УДК 510.647
DOI 10.17377/PAM.2017.17.3

Е. Е. Витяев, В. В. Мартынович

ВЕРОЯТНОСТНЫЕ ФОРМАЛЬНЫЕ ПОНЯТИЯ НА ЗАШУМЛЕННЫХ ДАННЫХ*

Предлагается качественно новый подход к решению проблемы обнаружения классов объектов на зашумленных контекстах. Для работы с шумом разработано логико-вероятностное обобщение формальных классов (понятий), а также показано, как решить возникающую проблему противоречивости логического вывода.

Ключевые слова: анализ формальных понятий, концептуальная решетка, Data Mining, индуктивное обучение, ассоциативные правила, задача кластеризации, вероятность, шум.

Введение

Сегодня, как никогда, вопрос автоматической обработки и хранения знаний, содержащихся в больших массивах данных, особенно актуален: осваивается Big Data, разрабатываются новые производственные и экспертные системы, интеллектуальные агенты, появляются новые методы Data Mining, умеющие получать знания из данных. Среди всех задач Data Mining мы особо выделим задачу кластеризации, которая заключается в разбиении данных на группы, исходя из соображений схожести (в некоторым смысле) образцов, без какой-либо предварительной информации о структуре данных.

Именно задача кластеризации активно обсуждается, и развиваются подходы к ее решению в теории анализа формальных понятий (АФП). В ней данные интерпретируются как формальный контекст (G, M, I), состоящий из множества объектов G, атрибутов M, и отношения принадлежности атрибутов объектам I. Основной задачей АФП является поиск формальных понятий, объединяющих схожие атрибуты и объекты, а также иерархии формальных понятий в виде концептуальной решетки, ее последующей обработки и визуализации.

Понятия, участвующие в построении концептуальной решетки, являются кластерами, группирующими объекты. В случае «хороших» данных, содержащих точную информацию, концептуальная решетка является идеальным решением задачи кластеризации. Однако в подавляющем большинстве практических задач данные являются зашумленными ввиду целого ряда причин:

1) ошибки измерения — погрешности приборов, шкал;
2) ошибки восприятия — несоответствие воспринимаемых фактов действительным;
3) помехи при передаче данных — механические искажения информации.

Существующие методы [1–3] построения решетки понятий на зашумленном контексте данных оказываются малопродуктивными, поскольку они рассчитаны на точные данные, и

* Работа выполненна при поддержке президентской программы по поддержке научных школ (грант НШ-6848.2016.1) и Российского научного фонда (грант РНФ 17-11-01176).

в условиях наличия шума (искажений данных) решетка понятий, полученная этими методами, оказывается заполненной зашумленными клонами исходных понятий.

Попытки решить эту проблему «в лоб» не выглядят многообещающими [1; 4]: основная их идея заключается в том, чтобы взять полную решетку понятий и избавиться от шумовых, побочных понятий путем фильтрации. Широко распространенной оценкой «побочности» понятия является Stability Index [4; 5]. Однако такой подход не справляется даже с простейшими задачами по устранению шума, а точность предсказания оставляет желать лучшего [6]. Кроме того, фильтрация решеток обладает еще одним существенным недостатком — исходная решетка понятий часто будет обладать экспоненциальными размерами [7], что неприемлемо для больших контекстов.

В данной работе мы рассматриваем проблему формирования понятий на зашумленных данных с точки зрения синтеза логики и вероятности. Мы опираемся на так называемые «естественные классы» и «естественную классификацию» [8; 9], изучаемые в когнитивных науках о функционировании человеческого сознания. Возможно, наиболее близким направлением из анализа формальных понятий являются работы, связанные с нечеткими структурами и формированием нечеткой решетки понятий [10].

Основная цель данной работы — разработка метода, который позволил бы восстанавливать исходную решетку понятий по зашумленному контексту и обладал рядом важных качеств:

- стабильность обнаруживаемых формальных понятий при незначительном шуме;
- потенциальная вычислимость, в обход генерации полной решетки понятий;
- решение проблемы противоречивости концептов, при включении отрицаний.

Содержательная часть статьи устроена следующим образом. В разделе 2 предлагается язык для описания вероятностных рассуждений на формальных контекстах. Для обнаружения нечетких знаний мы обобщаем понятие импликации, вводя на формальных контекстах вероятностную меру. На основе вероятностных импликаций построено обобщение оператора выводимости, расширяющего набор литер с помощью вероятностностных импликаций.

Поскольку совместность выводов играет важную роль в поиске формальных понятий, не все множества импликаций оказываются пригодны для логического вероятностного вывода. Необходимые и достаточные условия совместности изложены в виде теоретических результатов в разделе 3.

В разделе 5 вводится вероятностное обобщение формальных понятий и предлагается алгоритм прямого поиска обобщенных формальных понятий.

1. Основы анализа формальных понятий

Начнем с определения основных элементов АФП. Большинство определений здесь берет свое начало еще с классических трудов по АФП [11; 12], другие взяты из [13].

Определение 1. Формальный контекст — это тройка (G, M, I), где G и M — произвольные множества объектов и атрибутов, а $I \subseteq G \times M$ — бинарное отношение, выражающее принадлежность атрибута объекту.

На формальном контексте ключевую роль играют операторы взятия производной.

Определение 2. $A \subseteq G$, $B \subseteq M$. Тогда

1) $A^{\uparrow} = \{m \in M \mid \forall g \in A, (g, m) \in I\}$;
2) $B^{\downarrow} = \{g \in G \mid \forall m \in B, (g, m) \in I\}$.

Операторы производной связывают между собой подмножества объектов и атрибутов и по сути являются соответствием Галуа между множеством объектов и атрибутов контекста. В дальнейшем, если не возникает неоднозначностей, мы будем обозначать оба оператора производной единым символом «$'$».

Определение 3. Пара (A, B) — формальное понятие, если $A^{\uparrow} = B$ и $B^{\downarrow} = A$.

В силу этого определения понятие может быть полностью описано одним из своих компонентов — объемом A либо содержанием B, поскольку второй восстанавливается с помощью оператора производной. Если не возникает неоднозначностей, далее под формальным понятием $B \subseteq M$ мы имеем в виду пару (B', B).

Заметим, что обе суперпозиции операторов производной $^{\uparrow} \circ {}^{\downarrow}$ и $^{\downarrow} \circ {}^{\uparrow}$ являются операторами замыкания [11]. Следующее предложение связывает систему замкнутых множеств на контексте K, формальные понятия и операторы производных определены далее.

Предложение 1. *Пара $(B^{\downarrow}, B)$ является формальным понятием тогда и только тогда, когда $B^{\downarrow\uparrow} = B$, т. е. B замкнуто.*

Множество формальных понятий образует решетку, называемую **решеткой формальных понятий**.

Атрибуты могут состоять в многочисленных и довольно сложных зависимостях. Такие зависимости в направлениях Data Mining и Machine Learning принято описывать в терминах импликаций [11; 13].

Определение 4. Импликация — это пара атрибутных множеств $(B, C) \in M \times M$, записанная в форме $B \rightarrow C$. Импликация истинна на контексте K, если $\forall g \in G (B \not\subseteq g'$ или $C \subseteq g')$. Множество всех истинных импликаций на K будем обозначать $\mathrm{Imp}(K)$.

Определение 5. Для любого множества импликаций L построим оператор непосредственного вывода f_L, который добавляет все возможные заключения импликаций к исходному множеству атрибутов X:

$$f_L(X) = X \cup \{C \mid B \subseteq X,\ B \rightarrow C \in L\}.$$

Следующая теорема ниже лежит в основе определения вероятностных формальных понятий как неподвижных точек вероятностного аналога оператора вывода.

Теорема 1 [12]. *Для любого множества $B \subseteq M$, $f_{\mathrm{Imp}(K)}(B) = B \Leftrightarrow B'' = B$.*

2. Вероятностная логика на формальном контексте

Каждое небольшое изменение, внесенное в контекст, кардинально влияет на концептуальную решетку. Количество фиктивных понятий будет увеличиваться с уровнем шума, причем

качественная оценка скорости имеет экспоненциальный характер [7]. Устойчивость понятий относительно возможного шума можно обеспечить за счет оценки устойчивости импликаций, описывающих эти понятия. Естественным способом оценки будет привлечение вероятности для оценки правдоподобия атрибутных импликаций.

Определение 6. Для конечного формального контекста $K = (G, M, I)$ определим сигнатуру σ_K, содержащую лишь множество предикатных символов, совпадающее с M.

Определение 7. Классические логические конструкции переносятся тривиально:

1) Term_K — множество термов включает все символы переменных и ничего больше;
2) At_K — атомами будут все возможные выражения $m(t)$, где $m \in \sigma_K$ и $t \in \mathrm{Term}_K$;
3) L_K — литеры включают все атомы $m(t)$ плюс их отрицания $\neg m(t)$;
4) Φ_K — определяется индуктивно: всякий атом — формула, и для любых $\Phi, \Psi \in \Phi_K$ синтаксические конструкции $\Phi \wedge \Psi, \Phi \vee \Psi, \Phi \to \Psi, \neg\Phi$ — тоже формулы.

Определение 8. Для множества литер $L \subseteq \mathrm{L}_K$ мы определяем их конъюнкцию: $\wedge L = \bigwedge\limits_{P \in L} P$. Аналогично, $\neg L = \{\neg P \mid P \in L\}$.

Для сигнатуры σ_K и контекста-модели K определим интерпретацию предикатных символов следующим образом: $K \vDash m(x) \Leftrightarrow (x, m) \in I$. Факт истинности формулы φ на модели K, суженной до объекта g, записываем как $g \vDash \varphi \Leftrightarrow K_g \vDash \varphi$.

Определение 9. Рассмотрим произвольную вероятностную меру μ на множестве G, определенную в Колмогоровском смысле. Тогда контекстная вероятностная мера на множестве формул

$$\nu : \Phi_K \to [0, 1],\ \nu(\varphi) = \mu(\{g \mid g \vDash \varphi\}).$$

Определение 10. Набор литер M является ν-совместным, если $\nu(\wedge M) > 0$.

В дальнейшем понятие совместности набора литер рассматривается в рамках вероятностной меры контекста, и в отсутствие неоднозначности символ меры будет опускаться.

Рассмотрим некоторое подмножество атомов $L = \{m_i \in \sigma_K\} \subseteq \mathrm{At}(K)$ и $m \in \sigma_K$. Формула $m_1 \wedge m_2 \ldots \wedge m_k \to m = \wedge\{m_i\} \to m$ определяет импликацию на контексте, т. е. пару $(\{m_i\}_{i=1,..k}, \{m\})$. Наличие вероятностной меры на контексте позволяет определить достоверность импликаций на основе условной вероятности.

Определим правила на контексте и их составные части.

Определение 11. Пусть $C, H_i \in \mathrm{L}_K$, $C \notin \{H_1, H_2, \ldots H_k\}, k \geq 0$. Тогда

1) **правило** $R = (H_1, H_2, \ldots, H_k \to C)$ есть импликация $(H_1 \wedge H_2 \ldots \wedge H_k \to C)$;
2) **посылкой** $R^{\leftarrow}$ правила R называется набор литер $\{H_1, H_2, \ldots, H_k\}$;
3) **заключением** правила является $R^{\rightarrow} = C$;
4) **длиной правила** мы называем мощность его посылки $|R^{\leftarrow}|$;
5) если $R_1^{\leftarrow} = R_2^{\leftarrow}$ и $R_1^{\rightarrow} = R_2^{\rightarrow}$, тогда $R_1 = R_2$.

Определение 12. Вероятностью правила R является значение

$$\eta(R) = \nu(R^{\rightarrow} | \wedge R^{\leftarrow}) = \frac{\nu(\wedge R^{\leftarrow} \wedge R^{\rightarrow})}{\nu(\wedge R^{\leftarrow})}.$$

Если знаменатель $\nu(\wedge R^{\leftarrow})$ нулевой, вероятность правила остается неопределенной.

Определение 13. R_1 — **подправило** R_2 ($R_1 \sqsubset R_2$), если $R_1^{\rightarrow} = R_2^{\rightarrow}$, $R_1^{\leftarrow} \subset R_2^{\leftarrow}$.

Определение 14. R_1 **уточняет** R_2 ($R_1 > R_2$), если $R_2 \sqsubset R_1$ и $\eta(R_1) > \eta(R_2)$.

Основным техническим инструментом в доказательстве теоретических результатов будет служить понятие псевдоправила, с помощью которого удается уточнять уже существующее правило, а также связанная с ним теорема об уточнении.

Определение 15. Псевдоправило — это формула вида $R = ((P_1 \wedge \ldots \wedge P_k) \wedge \neg(N_1 \wedge \ldots \wedge N_s) \Rightarrow T)$; для псевдоправила R также определены посылка $R^{\leftarrow} = (P_1 \wedge \ldots \wedge P_k) \wedge \neg(N_1 \wedge \ldots \wedge N_s)$ и заключение $R^{\rightarrow} = T$; литеры P_i называются позитивной посылкой псевдоправила, а литеры N_j — негативной посылкой; вероятностью псевдоправила R является значение

$$\eta(R) = \nu(R^{\rightarrow} | R^{\leftarrow}) = \frac{\nu(\wedge R^{\leftarrow} \wedge R^{\rightarrow})}{\nu(\wedge R^{\leftarrow})}.$$

Теорема 2 (об уточнении). *Пусть* $S = ((\wedge A) \wedge \neg(\wedge B)) \Rightarrow G)$ *есть псевдоправило, а* $R = ((\wedge A) \Rightarrow G)$ *— соответствующее ему правило, полученное удалением негативной посылки, и, более того,* $\eta(S) > \eta(R)$. *Тогда для* R *существует уточнение* $R' > R$ *сформированное с помощью добавления литер из негативной посылки псевдоправила* S.

ДОКАЗАТЕЛЬСТВО. Для краткости обозначим $\overline{A} = \wedge A$, $\overline{B} = \wedge B$. Перепишем вероятность псевдоправила S как

$$\eta(S) = \nu(G \mid \overline{A} \wedge \neg\overline{B}) = \nu(G \mid \overline{A} \wedge (\neg B_1 \vee \ldots \vee \neg B_m)). \tag{1}$$

Основной шаг состоит в представлении дизъюнкции как дизъюнкции взаимоисключающих конъюнкций:

$$\neg B_1 \vee \ldots \vee \neg B_m = \bigvee_{i=(0,\ldots,0)}^{i=(1,\ldots,1,0)} (B_1^{i_1} \wedge \ldots \wedge B_m^{i_m}),$$

где 0 в мультииндексе означает наличие отрицания, а 1 — его отсутствие. Все мультииндексы включаются в лексикографическом порядке, за исключением последнего $(1, \ldots, 1)$, который соответствовал бы конъюнкции $B_1 \wedge \ldots \wedge B_m$.

Теперь условная вероятность (1) может быть переписана как

$$\eta(S) = \nu(G \mid \bigvee_{i=(0,\ldots,0)}^{i=(1,\ldots,1,0)} (\overline{A} \wedge B_1^{i_1} \wedge \ldots \wedge B_m^{i_m})). \tag{2}$$

Предположим, что заключение теоремы ложно и всякое обобщение $R' \sqsupset R$, сформированное добавлением литер из $\{B_1, \ldots, B_m\}$ к посылке, не будет уточнением. Это означает, что выполняются все неравенства типа $\nu(G \mid \overline{A} \wedge B_1^{i_1} \wedge \ldots \wedge B_m^{i_m}) \leq \nu(G \mid \overline{A})$, если, конечно, соответствующие условные вероятности определены. Поскольку $\nu(\overline{A} \wedge \neg\overline{B}) \neq 0$, существует хотя бы один мультииндекс $(i_1, \ldots, i_m)$, для которого условная вероятность определена. Но тогда

$$\nu(G \wedge \overline{A} \wedge B_1^{i_1} \wedge \ldots \wedge B_m^{i_m}) \leq \nu(G \mid \overline{A})\nu(\overline{A} \wedge B_1^{i_1} \wedge \ldots \wedge B_m^{i_m});$$

$$\nu(G \mid \bigvee_{i=(0,\ldots,0)}^{i=(1,\ldots,1,0)} (\overline{A} \wedge B_1^{i_1} \wedge \ldots \wedge B_m^{i_m})) = \frac{\nu(\vee\, G \wedge \overline{A} \wedge B_1^{i_1} \wedge \ldots \wedge B_m^{i_m})}{\nu(\vee\, \overline{A} \wedge B_1^{i_1} \wedge \ldots \wedge B_m^{i_m})} =$$

$$= \frac{\sum \nu(G \wedge \overline{A} \wedge B_1^{i_1} \wedge \ldots \wedge B_m^{i_m})}{\sum \nu(\overline{A} \wedge B_1^{i_1} \wedge \ldots \wedge B_m^{i_m})} \leq \frac{\nu(G \mid \overline{A}) \sum \nu(\overline{A} \wedge B_1^{i_1} \wedge \ldots \wedge B_m^{i_m})}{\sum \nu(\overline{A} \wedge B_1^{i_1} \wedge \ldots \wedge B_m^{i_m})} = \nu(G \mid \overline{A}).$$

Последнее в силу (2) означает, что $\eta(S) \leq \eta(R)$, а это непосредственно противоречит условияем теоремы. Поэтому наше предположение неверно, и хотя бы для одного из правил мы имеем $\nu(G \mid \overline{A} \wedge B_1^{i_1} \wedge \ldots \wedge B_m^{i_m}) > \nu(G \mid \overline{A})$. □

3. Непротиворечивость предсказаний

Определение 16. Оператор предсказания, использующий множество правил $\mathcal{R}$ — это

$$\Pi_{\mathcal{R}}(L) = L \cup \{C \mid \exists R \in \mathcal{R} : R^{\leftarrow} \subseteq L, R^{\rightarrow} = C\}.$$

Иногда множество предсказывающих правил $\mathcal{R}$ оказывается избыточным и, возможно, несовместимым, что в результате приводит к противоречивым предсказаниям. Присутствие одновременно литеры и ее отрицания в множестве атрибутов ведет к известным логическим проблемам, поэтому свойство непротиворечивости набора литер является необходимым.

Определение 17. Множество литер L будет непротиворечивым, если оно не содержит одновременно некоторого атома C и его отрицания $\neg C$.

Заметим также, что совместность и противоречивость тесно связаны друг с другом.

Предложение 2. *Если L — совместно, то L — непротиворечиво.*

ДОКАЗАТЕЛЬСТВО. Пусть $\exists G : G \in L$ и $\neg G \in L$, тогда $\nu(\wedge L) \leq \nu(G \wedge \neg G) = 0$. Это противоречит совместности L. □

Проблема противоречивости предсказаний широко известна как одна из фундаментальных проблем индуктивного вывода. Она берет начало еще с работ Гемпеля [14] и его последователей. Одно из возможных решений состоит в предъявлении к правилам требования максимальной специфичности [14]. Суть заключается в уточнении правил настолько, насколько это возможно, путем включения в них максимального количества доступной информации. Класс M_2, описанный ниже, выделяет правила, обладающие свойством максимальной специфичности.

Определение 18. $R \in M_1(C) \Leftrightarrow \eta(R) > \nu(R^{\rightarrow}), R^{\rightarrow} = C$.

Определение 19. $R \in M_2(C) \Leftrightarrow R \in M_1(C)$ и $[R \sqsubset \tilde{R} \Rightarrow \eta(\tilde{R}) \leq \eta(R)]$.

Класс Imp соответствует классическому множеству $\mathrm{Imp}\,(K)$ из определения 4.

Определение 20. $R \in \mathrm{Imp}\,(C) \Leftrightarrow R^{\rightarrow} = C$ и $\eta(R) = 1$.

Определение 21. Полные классы M_1, M_2 и Imp правил по всем литерам $C \in Lit(K)$:

1) $M_1 = \bigcup_C M_1(C)$;
2) $M_2 = \bigcup_C M_2(C)$;
3) $\mathrm{Imp} = \bigcup_C \mathrm{Imp}\,(C)$.

Пусть здесь и далее L — некоторое множество литер контекста K, т. е. $L \subseteq \mathrm{L}(K)$. Теперь мы докажем ключевую теорему совместности, которая гарантирующует корректность логических выводов по оператору предсказания. Она представляется основным результатом наших теоретических исследований.

Определение 22. Системой правил называется любое подмножество $\mathcal{R} \subseteq M_2$.

Теорема 3 (о совместности). *Пусть $\mathcal{R}$ — произвольная система правил. Тогда если L совместно, то его замыкание $\Pi_{\mathcal{R}}(L)$ также совместно.*

ДОКАЗАТЕЛЬСТВО. Рассмотрим все правила, которые добавляют литеры к образу предсказания на L: $Q = \{R \in \mathcal{R} \mid R^{\leftarrow} \subseteq L\}$. Занумеруем все элементы Q в произвольном порядке, $Q = \{Q_1, \dots Q_m\}$, и рассмотрим последовательность множеств $U_i = U_{i-1} \cup \{Q_i^{\rightarrow}\}, U_0 = L$. Мы покажем, что все U_i совместны.

$U_0 = L$ очевидно совместно согласно условиям теоремы.

Пусть U_i будет совместным. Обозначим $U = U_i, W = U_{i+1}, R = R_{i+1}$ и $G = R^{\rightarrow}$, $H = R^{\leftarrow}$, $N = U \setminus H$. Предположим, что W несовместно, т. е. $\nu(\wedge W) = 0$. Рассмотрим псевдоправило $F = (\wedge H \ \wedge \ \neg(\wedge N)) \Rightarrow G$. Возникает 2 случая.

Случай 1: $\nu(\wedge F^{\leftarrow}) \neq 0$. Тогда вероятность F определена и

$$\eta(F) = \frac{\nu(\wedge H \ \wedge \ \neg(\wedge N) \ \wedge \ G)}{\nu(\wedge H \ \wedge \ \neg(\wedge N))} = \frac{\nu(\wedge H \ \wedge \ G) - \nu(\wedge H \ \wedge \ (\wedge N) \ \wedge \ G)}{\nu(\wedge H) - \nu(\wedge H \ \wedge \ (\wedge N))} =$$
$$= \frac{\nu(\wedge H \ \wedge \ G) - \nu(\wedge W)}{\nu(\wedge H) - \nu(\wedge U)} = \frac{\nu(\wedge H \ \wedge \ G)}{\nu(\wedge H) - \nu(\wedge U)} > \frac{\nu(\wedge H \ \wedge \ G)}{\nu(\wedge H)} = \eta(R) > 0.$$

В соответствии с теоремой об уточнении найдется правило S, такое что $S > R$; однако это противоречит тому, что $R \in M_2$ ($R \in M_2$ декларирует неуточняемость правила R). Поэтому этот случай невозможен.

Случай 2: $\nu(\wedge F^{\leftarrow}) = 0$. Тогда

$$\nu(\wedge F^{\leftarrow}) = \nu(\wedge H \ \wedge \ \neg(\wedge N)) = 0 \Rightarrow \nu(\wedge H \ \wedge \ \neg(\wedge N) \ \wedge \ G) = 0;$$
$$0 = \nu(\wedge H \ \wedge \ (\wedge N) \ \wedge \ G) = \nu(\wedge H \ \wedge \ G) - \nu(\wedge H \ \wedge \ \neg(\wedge N) \ \wedge \ G) = \nu(\wedge H \ \wedge \ G).$$

Последнее влечет $\eta(R) = 0$, но $R \in M_1$ ($0 = \eta(R) > \eta(\varnothing \Rightarrow G) \geq 0$). □

Следствие 1. *Если L совместно, то $\Pi_{\mathcal{R}}(L)$ непротиворечиво.*

4. О несовместных наборах

Немного более сложным, но по-прежнему разрешимым оказывается вопрос о судьбе противоречивых множеств литер. Для их характеризации предлагается конструкция ν-максимальных по совместности подмножеств исходного множества литер L.

Определение 23. Мы говорим, что M ν-максимально в L или $M \underset{\nu}{\subseteq} L$, если M — максимальное по включению подмножество L, такое что M совместно.

Определение 24. Множество правил $\mathcal{R}$ называется точным, если $\mathrm{Imp} \subseteq \mathcal{R}$.

Теорема 4 (о замыкании ν-максимального). *Пусть $M \underset{\nu}{\subseteq} L$, а $\mathcal{R}$ — точная система правил. Тогда $M \cup \neg(L \setminus M) \subseteq \Pi_{\mathcal{R}}(M)$.*

ДОКАЗАТЕЛЬСТВО. Предположим, x принадлежит выражению в левой части формулы. Случай $x \in M$ очевиден: $x \in \Pi_{\mathcal{R}}(M)$ по определению оператора предсказания.

Если же $x \in L \setminus M$, то по определению ν-максимального подмножества множество $M \cup \{x\}$ несовместно (иначе получим большее по включению совместное множество). Поэтому

$$\nu(\wedge M \ \wedge \ x) = 0;$$
$$\nu(\wedge M \ \wedge \ \neg x) = \nu(\wedge M) - \nu(\wedge M \ \wedge \ x) = \nu(\wedge M);$$

Положим $R = (\wedge M \Rightarrow \neg x)$. Из соотношений выше нетрудно вычислить вероятность правила R:

$$\eta(R) = \frac{\nu(\wedge M \ \wedge \ \neg x)}{\nu(\wedge M)} = 1.$$

Значит, R — импликация, и $R \in \mathrm{Imp}(K) \subseteq \mathcal{R}$. Поэтому R неизбежно добавит $\neg x$ в образ предсказания $\Pi_{\mathcal{R}}(M)$. □

Теорема 5. *Рассмотрим $M \underset{\nu}{\subseteq} L$, $N \underset{\nu}{\subseteq} L$ и $M \neq N$. Тогда*

1) $\exists x \ : \ x \in \Pi_{\mathcal{R}}(M)$ *и* $\neg x \in \Pi_{\mathcal{R}}(N)$;

2) $\Pi_{\mathcal{R}}(M) \supseteq \Pi_{\mathcal{R}}(M \cap N) \subseteq \Pi_{\mathcal{R}}(N)$ *и* $\Pi_{\mathcal{R}}(M) \neq \Pi_{\mathcal{R}}(N)$.

ДОКАЗАТЕЛЬСТВО. 1. $M \neq N$ означает, что найдется $x \in M \setminus N \subseteq L \setminus N$. Поскольку $N \cup \neg(L \setminus N) \subseteq \Pi_{\mathcal{R}}(N)$, то заключаем, что $\neg x \in \Pi_{\mathcal{R}}(N)$.

2. $\Pi_{\mathcal{R}}(N)$ — совместно и непротиворечиво, и $\neg x \in \Pi_{\mathcal{R}}(N)$; это значит, что $x \notin \Pi_{\mathcal{R}}(N)$ и $x \in \Pi_{\mathcal{R}}(M) \setminus \Pi_{\mathcal{R}}(N)$. Далее, $M \cap N \subset M$, и поэтому $\Pi_{\mathcal{R}}(M \cap N) \subseteq \Pi_{\mathcal{R}}(M)$. □

Последние две теоремы позволяют заключить, что существует вложение из множества ν-максимальных подмножеств L в множество неподвижных точек, притом каждая из неподвижных точек покрывает все атомы из L (в виде атомов непосредственно или их отрицаний).

Непротиворечивость и совместность неподвижных точек была доказана ранее. Для случая несовместных множеств ответ дает следующая теорема.

Теорема 6 (о несовместности). *Если L несовместно, то $\Pi_{\mathcal{R}}(L)$ противоречиво.*

ДОКАЗАТЕЛЬСТВО. Найдем ν-максимальное подмножество в L и обозначим его как M. $M \neq L$, иначе L было бы совместным. Поэтому найдется $x \in L \setminus M$. Множество $\{x\}$ расширяется до максимального совместного $N \underset{\nu}{\subseteq} L$. По построению $x \in N \setminus M \Rightarrow M \neq N$. По теореме о замыкании ν-максимального найдется y, такое что $y \in \Pi_{\mathcal{R}}(M)$ и $\neg y \in \Pi_{\mathcal{R}}(N)$:

$$\left.\begin{array}{l} M \subseteq L \\ N \subseteq L \end{array}\right\} \Rightarrow \left.\begin{array}{c} y \in \Pi_{\mathcal{R}}(M) \subseteq \Pi_{\mathcal{R}}(L) \\ \neg y \in \Pi_{\mathcal{R}}(N) \subseteq \Pi_{\mathcal{R}}(L) \end{array}\right\} \Rightarrow \Pi_{\mathcal{R}}(L) \text{ — противоречиво.}$$

□

5. Вероятностные формальные понятия

По аналогии с теоремой 1 неподвижные точки оператора предсказания — очевидные кандидаты на роль содержаний вероятностных понятий. Для объемов понятий идея состоит в том [4; 5], чтобы использовать все возможные прообразы оператора замыкания, т. е. все множества $M : \Pi_{\mathcal{R}}(M) = B$, для композиции их производных в единый объем понятия A. Это позволит восстановить исходные соответствия объекты-понятия, а также включить все объекты одного и того же класса в единый объем понятия.

Определение 25. Вероятностным формальным понятием на K называется пара множеств объектов и атрибутов (A, B), удовлетворяющая соотношениям

$$\Pi_{\mathcal{R}}(B) = B, \quad A = \bigcup_{\Pi_{\mathcal{R}}(C)=B} C'.$$

Чтобы отличать вероятностные понятия от классических понятий на контексте K, последние мы будем называть четкими формальными понятиями. Выбор такого множества A в качестве объема понятия также основывается на справедливости следующей теоремы, связывающей вероятностные и четкие формальные понятия.

Теорема 7 [3]. *Пусть K будет формальным контекстом.*

1. *Если (A, B) — это четкое понятие на K, тогда существует вероятностное понятие на том же самом контексте (N, M) такое, что $A \subseteq N$, и $B \subseteq M$.*
2. *Если (N, M) — это вероятностное понятие на K, тогда найдется семейство четких понятий на том же контексте C, таких что*

$$\forall (A, B) \in C \; (\Pi_{\mathcal{R}}(B) = M), \quad N = \bigcup_{(A,B)\in C} A.$$

Для построения иерархической структуры классов заметим, что получившийся оператор $\Pi_{\mathcal{R}}$ в действительности является оператором замыкания. Поэтому достаточно найти вероятностные понятия, содержащие максимальное количество признаков, получая оставшуюся часть концептуальной решетки из попарных пересечений понятий.

Наиболее важной, с практической точки зрения, представляется возможность использовать оператор предсказания в качестве инструмента прогнозирования в реальных задачах, связанных с Data Mining. Алгоритмический процесс начинается с генерации множества прогнозирующих (предсказывающих) правил. Теорема совместности требует, чтобы эти правила были подмножеством M_2. Поэтому предположим, что на контексте K уже была обнаружена система правил $\mathcal{R} \subseteq M_2$. Определение 25 указывает **процедуру поиска замыканий** наборов литер.

Вход: формальный контекст $K = (G, M, I)$, система правил $\mathcal{R} \subseteq M_2$.

Выход: множество всех вероятностных формальных понятий (относительно $\mathcal{R}$).

1. Установим $k = 1$ и сгенерируем $C^{(1)} = \{\Pi_{\mathcal{R}}(R^{\leftarrow}) \mid R \in \mathcal{R}\}$.

   Семейство $C^{(1)}$ может быть любым стартовым набором гипотез.

2. На шаге $k \geq 1$, в случае если $C^{(k)} = \varnothing$, алгоритм заканчивает свое выполнение и печатает на выход список всех обнаруженных неподвижных точек.

3. Иначе на шаге $k \geq 1$ вычисляем множество $A = \{g \in G \mid \Pi_{\mathcal{R}}(g' \cap B) = B\}$ для каждого $B \in C^{(k)}$. Если $A \neq \varnothing$, пара (A, B) добавляется к списку обнаруженных неподвижных точек.

4. Генерируем новое множество гипотез $C^{(k+1)} = \{\Pi_{\mathcal{R}}(B \cup C) \mid B, C \in C^{(k)}\} \setminus C^{(k)}$, которое соответствует объединению двух понятий на решетке понятий FCA.

5. Полагаем $k := k + 1$ и идем на шаг 2.

Заключение

Вероятностные формальные понятия являются удобной формализацией: они описывают зашумленные контексты в вероятностных, нечетких терминах и могут быть использованы для обнаружения точных формальных понятий даже на зашумленных контекстах, не прибегая к построению полной решетки понятий. При этом решена проблема противоречивости, обычно возникающая при построении нечетких операторов логического вывода, а также дана характеризация противоречивых наборов.

Более детального внимания заслуживает изучение влияния различных видов шумов на устойчивость алгоритма. Интересна зависимость максимально допустимого уровня шума от специфики данных и его влияние на время выполнения алгоритма.

Для обоснования актуальности метода планируется ряд сравнений характеристик скорости и точности с известными алгоритмами обучения без «учителя».

Список литературы

1. *Klimushkin M., Obiedkov S., Roth C.* Approaches to the Selection of Relevant Concepts in the Case of Noisy Data // ICFCA 2010 Proceedings. LNAI, 2010. Vol. 5986. P. 255–266.

2. *Bayardo Jr., R., Goethals B., Zaki M. (eds.)* Proc. of the IEEE ICDM Workshop on Frequent Itemset Mining Implementations. CEUR-WS.org, 2004.

3. *Витяев Е. Е., Демин А. В., Пономарев Д. К.* Вероятностное обобщение формальных понятий // Программирование. 2012. Т. 38, № 5. С. 18–34.

4. *Kuznetsov S. O.* On Stability of a Formal Concept // Annals of Mathematics and Artificial Intelligence. 2007. Vol. 49. P. 101–115.

5. *Kuznetsov S. O.* Concept Stability as a Tool for Pattern Selection // CEUR Workshop proceedings, ECAI. 2014. Vol. 1257. P. 51–58.

6. *Prokasheva O., Onishchenko A., Gurov S.* Classification Based on Formal Concept Analysis and Biclustering: Possibilities of the Approach // Computational Mathematics And Modeling. 2014. Vol. 23. No. 3. P. 329–336.

7. *Emilion R., Levy G.* Size of Random Galois Lattices // Discrete Applied Math. 2009. Vol. 157. P. 2945–2957.

8. *Rosch E., Lloyd B. B.* Principles of Categorization // Cognition and Categorization. Lawrence Elbaum Associates, 1978.

9. *Rehder B.* Categorization as Causal Reasoning // Cognitive Sci. 2003. Vol. 27. No. 5. P. 709–748.

10. *Quan T. T., Hui S. C., Cao T. H.* A Fuzzy FCA-Based Approach to Conceptual Clustering for Automatic Generation of Concept Hierarchy on Uncertainty Data // CLA 2004 CEUR Workshop proceedings, Belohlavek R., Snasel V. (Eds.). 2004. Vol. 110. P. 507–522.

11. *Ganter B.* Formal Concept Analysis: Methods, and Applications in Computer Science. TU Dresden, 2003.

12. *Ganter B., Wille R.* Formal Concept Analysis — Mathematical Foundations. Berlin; Heidelberg; N. Y.: Springer, 1999.

13. *Ganter B., Obiedkov S.* Implications in Triadic Formal Contexts. TU Dresden: Springer, 2004.

14. *Hempel C.* Inductive Inconsistencies // Synthese. 1960. Vol. 12. P. 439–469.

15. *Vityaev E. E., Martynovich V. V.* Probabilistic Formal Concepts with Negation // Perspectives of System Informatics, A. Voronkov, I. Virbitskaite (Eds.). LNCS, 2015. Vol. 8974. P. 385–399.

Материал поступил в редколлегию 05.04.2016

Адреса авторов

ВИТЯЕВ Евгений Евгеньевич
Институт математики им. С. Л. Соболева СО РАН
пр. Акад. Коптюга, 4, Новосибирск, 630090, Россия
Новосибирский государственный университет
ул. Пирогова, 1, Новосибирск, 630090, Россия
vityaev@math.nsc.ru

МАРТЫНОВИЧ Виталий Валерьевич
Институт математики им. С. Л. Соболева СО РАН
пр. Акад. Коптюга, 4, Новосибирск, 630090, Россия
vilco@ya.ru