ПРИМЕНЕНИЕ ПРИНЦИПОВ ЭКСПЕРТНОЙ КЛАССИФИКАЦИИ ДЛЯ АНАЛИЗА ФОРМАЛЬНЫХ ПОНЯТИЙ

С.Ю. Соловьев,
доктор физико-математических наук, профессор кафедры алгоритмических языков, факультет вычислительной математики и кибернетики Московского государственного университета им. М.В. Ломоносова

Д.Е. Стельмашенко,
инженер факультета вычислительной математики и кибернетики Московского государственного университета им. М.В. Ломоносова

E-mail: soloviev@glossary.ru, dashanikolaeva@gmail.com
Адрес: г. Москва, ГСП-1, Ленинские горы, МГУ имени М.В. Ломоносова, 1, стр. 52

В статье рассматривается вариант задачи конструирования решетки формальных понятий и обсуждаются особенности алгоритма ее решения. Для повышения эффективности алгоритма предлагается использовать принципы экспертной классификации.

Ключевые слова: анализ формальных понятий, метод экспертной классификации, принцип доминирования по характерности, шкала характерности.

1. Введение

В последнее время анализ формальных понятий (англ. Formal Concept Analysis; FCA) выступает своеобразным собирателем различных методов обработки неколичественных данных. В работе рассматривается вопрос о применении основных принципов экспертной классификации для одной задачи анализа формальных понятий. Систематическому изложению этих родственных направлений искусственного интеллекта посвящены монографии [1-4], здесь же приводится минимально необходимый набор сведений из FCA (п. 2), а также адаптированные под конкретную задачу (п. 3) основы экспертной классификации (п. 4). Результатом такого «симбиоза» стали новые подходы к обработке формальных понятий (п. 5) и формулировки новых задач.

2. Некоторые определения FCA

Формальный контекст есть тройка множеств (G, M, I), в которой элементы G называются объектами, элементы M – признаками, а I есть подмножество из $G \times M$. Если $(g, m) \in I$, то говорят «объект g обладает признаком m» либо, что то же самое, «признак m встречается в описании объекта g». При переходе от традиционных неколичественных моделей проблемных областей (первичных моделей) к формальным контекстам (вторичным моделям) каждый атрибут-свойство X с конечным числом значений $X_1, \ldots, X_n$ кодируется набором (0,1)-признаков $m_{X1}, \ldots, m_{Xn}$ по принципу «одно значение атрибута – один признак контекста».

Формальный контекст $K = (G, M, I)$ порождает так называемые формальные понятия. Пара непустых подмножеств (A, B), где $A \subseteq G$ и $B \subseteq M$, называется формальным понятием, если (1) каждый объект из A обладает всеми признаками из B, и (2) никакая другая пара $(A \cup A', B \cup B')$ не удовлетворяет требованию (1). Пары множеств $(\varnothing, M)$ и $(G, \varnothing)$ являются формальными понятиями по определению. Если (A, B) – формальное понятие контекста K, то A называется объемом понятия, B – его содержанием. В дальнейшем изложении будем полагать, что множества G и M конечны, а все термины «контекст», «понятие» и «содержание» автоматически подразумевают спецификацию «формальный».

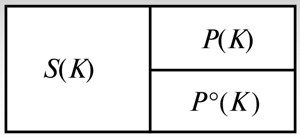
Множество $C(K)$ всех понятий контекста K образует полную решетку, с отношением частичного порядка

$$(A_1, B_1) \prec (A_2, B_2) \Leftrightarrow B_1 \supseteq B_2.$$

Соответственно, решеткой также является множество $S(K) = \{B \mid (A, B) \in C(K)\}$ с отношением частичного порядка $\supseteq$. Решетки $C(K)$ и $S(K)$ изоморфны; располагая одной из них, можно построить другую решетку. Рассмотрим подробнее те наборы признаков, которые не относятся к $S(K)$ и образуют множество $S^\circ(K) = 2^M \setminus S(K)$.

Подмножество признаков H контекста $K = (G, M, I)$ будем называть запретом, если $H \neq M$ и в описаниях объектов признаки из H вместе никогда не встречаются: $\forall g \in G \;\; \exists b \in H : \; (g, b) \notin I$. Требование $H \neq M$ объясняется лишь тем, что в содержаниях понятий – элементах множества $S(K)$ – все признаки запретов встречаются исключительно в понятии $(\varnothing, M)$. Множество всех запретов контекста K обозначим $P(K)$. Кроме того, обозначим $P^\circ(K)$ множество $S^\circ(K) \setminus P(K)$. Описанная классификация всевозможных наборов признаков представлена на *рис. 1.*

Эквивалентным представлением контекста $K = (G, M, I)$ является объектно-признаковая таблица $Tab(K)$, в которой именами столбцов служат признаки из M, кодами строк – объекты из G, а на



$S^\circ(K) = P(K) \cup P^\circ(K)$

Рис. 1. Строение множества 2^M для контекста K

пересечении строки g и столбца m стоит $\boxed{\times}$, если $(g, m) \in I$.

Общее определение контекста допускает наличие в нем объектов-дубликатов, которым в $Tab(K)$ соответствуют строки, отличающихся только кодами. Объекты-дубликаты не влияют на структуру решетки понятий, поэтому для многих приложений интерес представляют контексты без объектов-дубликатов, именуемые далее простыми контекстами. Не ограничивая общности, можно полагать, что в простом контексте (G, M, I) каждый объект g из G представляет собой подмножество признаков, а $I = \varphi(G) \equiv \{(g, m) \mid g \in G, m \in g\}$. При таком подходе для задания простого контекста достаточно перечислить описания его объектов.

3. Задача совмещения формальных контекстов

Очевидно, что любой нетривиальный контекст K способен порождать некоторые производные контексты посредством вычеркивания из $Tab(K)$ отдельных строк-объектов и/или столбцов-признаков. На содержательном уровне задача совмещения контекстов состоит в конструировании решетки $C(K)$ неизвестного контекста K по производным от него контекстам $K_1 = (G_1, M_1, I_1)$ и $K_2 = (G_2, M_2, I_2)$. Известные подходы к решению задачи совмещения различаются дополнительными предположениями и целевыми установками.

Для контекстов общего вида в случае $G_1 = G_2 = G$, $M_1 \cup M_2 = M$ и $M_1 \cap M_2 = \varnothing$ задача совмещения контекстов часто решается посредством так называемой «Nested Line» диаграммы [1]. Построение «Nested Line» диаграммы предполагает подстановку диаграммы Хассе решетки $C(K_1)$ в каждый узел соответствующей диаграммы решетки $C(K_2)$. Считается, что построенная таким образом диаграмма более наглядна, чем диаграмма Хассе решетки $C(K)$.

Для контекстов общего вида в случае $G_1 \cup G_2 = G$ и $M_1 \cup M_2 = M$ искомую решетку $C(K)$ предлагается [5] строить по контексту $K = (G_1 \cup G_2, M_1 \cup M_2, I_1 \cup I_2)$. Метод используется для пополнения баз знаний. Модификация метода [6] позволяет выявить в объединенном множестве признаков дубликаты и модифицировать отношение $I_1 \bigcup I_2$.

Упомянутые подходы к решению задачи совмещения не претендуют на точное построение решетки $C(K)$. Иначе устроен метод @FC [7,8], о котором речь пойдет далее. Название метода происходит от первых букв ключевых слов его описания: Alternative design, Testing, Formal context, Concept. Метод @FC использует специальную терминологию простых контекстов.

Если $K = (G, M, \varphi(G))$ – простой контекст, заданный множеством своих объектов G (глобальный контекст), то проекцию G на фиксированное подмножество признаков M_T будем называть локальным контекстом. Формально, локальный контекст K_T есть

$$(G_T, M_T, \varphi(G_T)), \text{ где } G_T = \{ g \bigcap M_T \mid g \in G \}.$$

На практике получение K_T может быть связано, например, со снаряжением целой экспедиции к местам обитания изучаемых объектов G, причем исследовательские возможности экспедиции ограничены набором признаков M_T.

Для глобального контекста K и пары локальных контекстов K_1, K_2 в случае $M_1 \cup M_2 = M$ метод @FC предлагает строить множество $S(K)$ в виде

$$V = \{ B_1 \bigcup B_2 \mid B_1 \in S(K_1), B_2 \in S(K_2) \}$$

с последующей пошаговой классификацией элементов V на элементы $S(K)$ и $S^\circ(K)$. Доказано, что метод @FC позволяет построить множество $S(K)$ за конечное число шагов. На каждом шаге построения принимается решение о запросе нового локального контекста, и по итогам его специальной обработки для некоторых элементов V принимаются классификационные решения. Процесс формирования $S(K)$ заканчивается, когда классификационные решения приняты для всех элементов множества V.

Запрос нового локального контекста в виде множества его признаков M_T формируется автоматически или в диалоге с пользователем. При этом варианты запросов известны заранее, и необходимо «лишь» выбрать лучший из них.

В части обработки локальных контекстов K_T метод @FC использует правила классификации, которые иногда позволяют относить элементы V к классам $S(K)$ или $S^\circ(K)$. Приведем в качестве примера два классификационных правила:

(K1) Если $v \in S(K_T)$ и $\{v' \mid v \bigcup v' \in V\} = \{\varnothing\}$, то $v \in S(K)$.

(K2) Если $v \in P(K_T)$ и $v \bigcup v' \in V$ для некоторого v', то $v \bigcup v' \in P(K)$.

С одной стороны, классификационные правила представляют собой строго доказанные утверждения [8], а с другой стороны – их можно использовать в качестве условных операторов алгоритма реализации метода @FC. Обработка очередного локального контекста заканчивается, когда все правила классификации оказываются неприменимыми.

При высокой стоимости локальных контекстов возникает потребность в технологиях их «глубокой переработки», позволяющих на каждом шаге классифицировать как можно больше элементов V. Аналогичная потребность исследовалась в 1980-х научным коллективом под руководством О.И. Ларичева. Результатом этой работы стал метод экспертной классификации, позволяющий при незначительных дополнительных, но проверяемых предположениях существенно увеличить количество классификационных решений в задаче совмещения формальных контекстов.

4. Принципы экспертной классификации

В терминах FCA главная идея метода экспертной классификации состоит в обоснованном переносе классификационных решений, полученных для одного объекта, на некоторые другие объекты. Базой знаний для обоснования и переноса служит отношение доминирования, которое, в свою очередь, строится из отношений характерности. Считается, что контекст $K = (G, M, I)$ представляет собой вторичную модель проблемной области, а в первичной модели используются атрибуты-свойства, принимающие конечное число значений. При этом контекст K наследует от первичной модели разбиение множества признаков $M = M_1 \bigcup M_2 \bigcup ... \bigcup M_k$, в котором каждое подмножество M_i соответствует ровно одному атрибуту.

Теоретическую основу метода экспертной классификации составляет гипотеза, согласно которой «признаки каждого множества M_i различаются по степени характерности для объектов G». Если для конкретной проблемной области гипотеза справедлива, то эксперт способен построить на элементах множества M_i асимметричное, антирефлексивное и транзитивное отношение характерности $<_i$; запись «$m_{i1} < m_{i2}$» означает, что признак m_{i1} менее характерен для объектов G, чем признак m_{i2}. Экспериментально установлено, что фраза «быть характерным для объектов *G*» и фразы, производные от нее, понятны эксперту без разъяснений, и он почти всегда способен построить *k* штук отношений характерности – по одному для каждого множества M_i.

Строго говоря, в оригинальном описании экспертной классификации предполагается линейность всех отношений характерности. Это обстоятельство существенно для представления данных и для организации диалога с экспертом при извлечения знаний об отношениях характерности. В настоящей работе вместо линейного порядка предполагается наличие частичного порядка. Соответственно, изменяется процедура опроса эксперта, в основу которой полагается сравнение на характерность в различных парах признаков из M_i.

Совокупность отношений характерности $<_1, ..., <_k$ порождает рефлексивное, антисимметричное и транзитивное отношение доминирования $\leq$ на всевозможных подмножествах признаков из *M*. По определению, пара подмножеств b_1 и b_2 связана отношением $b_1 \leq b_2$, если для любого $i = 1, ..., k$ выполняется одно из двух: либо $b_1 \cap M_i = b_2 \cap M_i$, либо $b_1 \cap M_i = \{m_{i1}\}$, $b_2 \cap M_i = \{m_{i2}\}$ и $m_{i1} < m_{i2}$.

Если отношение доминирования удается построить, то механизм переноса классификационных решений применительно к объектам контекста – элементам множества *G* – описывается двумя постулированными, но вполне разумными правилами:

(П1) Если $g_1 \in G$, $g \in 2^M$ и $g_1 \leq g$, то $g \in G$;

(П2) Если $g_1 \neq G$, $g \in 2^M$ и $g \leq g_1$, то $g \neq G$.

5. Метод доминирования для формальных контекстов

Метод @FC, в отличие от оригинальной версии экспертной классификации, оперирует не только описаниями объектов, но и другими подмножествами признаков. В связи с этим возникает необходимость корректного обобщения правил П1 и П2 на случай подмножеств из *V*, и главная роль в таком обобщении отводится формальной интерпретации отношения характерности.

Простейшая, хотя и не единственная теоретико-множественная интерпретация отношений $<_i$ имеет вид:

$$m_{i1} <_i m_{i2} \Leftrightarrow \{g \in G \mid m_{i1} \in g\} \subseteq \{g \in G \mid m_{i2} \in g\}.$$

Доказано, что при таком подходе к отношениям характерности правила П1 и П2 превращаются в формальные утверждения, а кроме того, справедливы также утверждения

(П3) Если $v_1 \in S(K)$, $v \in V$ и $v_1 \leq v$, то $v \in S(K)$;

(П4) Если $v_1 \in S^\circ(K)$, $v \in V$ и $v \leq v_1$, то $v \in S^\circ(K)$;

(П5) Если $v_1 \in P(K)$, $v \in V$ и $v \leq v_1$, то $v \in P(K)$;

(П6) Если $v_1 \in P^\circ(K)$, $v \in V$ и $v \leq v_1$, то $v \in P^\circ(K)$.

В доказательствах утверждений П3–П6 существенно используются свойства объектов глобального контекста, однако в окончательных формулировках объекты не фигурируют, что позволяет использовать утверждения П3–П6 в качестве дополнительных правил распространения классификационных решений метода @FC.

6. Заключение

Представленная модификация метода совмещения @FC применима для глобальных контекстов, удовлетворяющих гипотезе доминирования в ее простейшей интерпретации. Проверка гипотезы по решетке содержаний *S*(*K*) представляет собой самостоятельную теоретическую задачу, способную в перспективе породить новую версию алгоритма совмещения локальных контекстов. Кроме того, один из принципов экспертной классификации состоит в обязательном исследовании наборов признаков на границах классов. Адаптация этого принципа для задач FCA также представляется весьма перспективной.

Предложенный и намеченные алгоритмы могут использоваться при решении практических задач бизнес-информатики, сводимых к обработке формальных понятий. ■

Литература

1. Ganter B., Wille R. Formal Concept Analysis: Mathematical Foundations. – Springer, 1999.
2. Гуров С.И. Булевы алгебры, упорядоченные множества, решетки: Определения, свойства, примеры. – М.: ЛИБРОКОМ, 2013.
3. Ларичев О.И. Теория и методы принятия решений, а также Хроника событий в Волшебных Странах. – М.: Логос, 2000.

4. Ларичев О.И., Мечитов А.И., Мошкович Е.М., Фуремс Е.М. Выявление экспертных знаний. – М.: Наука, 1989.
5. Li G.Y., Liu S.P., Zhao Y. Formal Concept Analysis based Ontology Merging Method // Proceeding of 3rd IEEE International Conference on Computer Science and Information Technology. – 2010 – volume 8 – P. 279-282.
6. Bendaoud R., Napoli A., Toussaint Y. A proposal for an Interactive Ontology Design Process based on Formal Concept Analysis // Frontiers in Artificial Intelligence and Applications. – 2008 – volume 183: Formal Ontology in Information Systems – P. 311-323.
7. Стельмашенко Д.Е. Алгоритм восстановления глобального контекста // Сборник статей молодых ученых факультета ВМК МГУ. – 2012 – выпуск 9 – С.191-209.
8. Соловьев С.Ю., Стельмашенко Д.Е. Подходы к исследованию формальных контекстов // Информационные процессы. – 2011 – том 11 – № 2 – С. 277-290.