С.Ю. Соловьев, Д.Е. Стельмашенко

# Метод экспертной классификации как инструмент анализа формальных понятий

**Аннотация:** в статье рассматривается вариант задачи конструирования решетки формальных понятий и обсуждаются особенности алгоритма ее решения. Для повышения эффективности алгоритма предлагается использовать метод экспертной классификации.

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

## Введение

Анализ формальных понятий (англ. Formal Concept Analysis; FCA) – отрасль интеллектуального анализа данных, в которой предполагается переход к понятийным структурам проблемных областей. Методология конструирования понятий из неколичественных данных составляет предмет и главную задачу FCA. В работе рассматривается разновидность главной задачи FCA, решение которой сводится к последовательности, вообще говоря, дорогостоящих итераций. На каждой итерации обрабатывается предварительно построенное множество кандидатов в понятия, и для некоторых из них принимаются положительные или отрицательные классификационные решения; процесс завершается, когда классификационные решения приняты для всех кандидатов. В интересах сокращения числа итераций в работе предлагается использовать метод экспертной классификации, позволяющий переносить классификационные решения, принятые для одних кандидатов на подмножества других кандидатов.

Систематическому изложению FCA и экспертной классификации посвящены монографии [1-4], здесь же приводится минимально необходимый набор сведений из FCA (п. 1), а также адаптированные под конкретную задачу (п. 2) основы экспертной классификации (п. 3). Результатом совмещения двух родственных методов стали новые подходы к обработке формальных понятий (п. 4) и постановки новых задач.

## 1. Некоторые определения FCA

FCA исходит из весьма общей модели данных, именуемой формальным контекстом. По определению, формальный контекст есть тройка множеств $(G, M, I)$, в которой элементы $G$ называются объектами, элементы $M$ – признаками, а $I$ есть подмножество пар из $G \times M$. Если $(g, m) \in I$, то говорят “объект $g$ обладает признаком $m$” либо, что тоже самое, “признак $m$ встречается в описании объекта $g$”.

Для конечных множеств $G$ и $M$ – а в дальнейшем изложении рассматривается именно такой случай – 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 пересечении строки $g$ и столбца $m$ стоит $\boxed{\times}$, если $(g,m)\in I$.

Формальный контекст $K=(G,M,I)$ порождает формальные понятия. Пара непустых подмножеств $(A,B)$, где $A\subseteq G$ и $B\subseteq M$, называется формальным понятием, если (a) каждый объект из $A$ обладает всеми признаками из $B$, и (b) никакая другая пара $(A\bigcup A',B\bigcup B')$ не удовлетворяет требованию (a). Кроме того, пары множеств $(\varnothing,M)$ и $(G,\varnothing)$ являются формальными понятиями по определению. Если $(A,B)$ – формальное понятие, то $A$ называется объемом понятия, $B$ – его содержанием. Далее используются только формальные понятия и формальные контексты, а потому указания на их формальные статусы опускаются.

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

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

Решеткой также является множество содержаний $S(K)=\{\,B\,|\,(A,B)\in C(K)\,\}$ с отношением частичного порядка $\supseteq$. Решетки $C(K)$ и $S(K)$ изоморфны, однако решетка $S(K)$ более удобна, так как охватывает все понятия контекста, но не зависит от наличия в контексте объектов-дубликатов, которым в $Tab(K)$ соответствуют строки, отличающиеся только кодами.

Контексты без объектов-дубликатов будем называть простыми контекстами. Не ограничивая общности, можно полагать, что в простом контексте $(G,M,I)$ каждый объект $g$ из $G$ однозначно описывается подмножеством признаков[1], а $I=\varphi(G)\equiv\{\,(g,m)\,|\,g\in G,m\in g\,\}$. При таком подходе для задания простого контекста достаточно перечислить описания его объектов.

Независимо от особенностей того или иного контекста $K=(G,M,I)$ для $S(K)$ всегда можно указать некоторое объемлющее его подмножество $U(K)$ элементов из $2^M$. В простейшем случае $U(K)=2^M$. $S(K)\subseteq U(K)\subseteq 2^M$. Ценность множества $U(K)$ состоит в том, что его элементы являются кандидатами в содержания формальных понятий. Кандидаты, не относящиеся к $S(K)$, образуют множество $S^{\circ}(K)=U(M)\setminus S(K)$. В свою очередь, элементы $S^{\circ}(K)$ подразделяются на запреты и все прочие.

Для контекста $K=(G,M,I)$ собственное подмножество признаков $H$, $H\subset M$, будем называть запретом, если в описаниях объектов признаки из $H$ вместе никогда не встречаются:

[1] то есть $g\subseteq M$

$\forall g \in G \ \exists b \in H : (g,b) \notin I$. Множество всех запретов контекста $K$ обозначим $P(K)$. Кроме того, обозначим $P^{\circ}(K)$ множество $S^{\circ}(K) \setminus P(K)$. Таким образом, для произвольного кандидата $A$ из $U(K)$ возможны следующие классификационные решения:

- либо $A \in S(K)$,
- либо $A \in S^{\circ}(K)$ и дополнительно: либо $A \in P(K)$,

  либо $A \in P^{\circ}(K)$.

Наличие дополнительной классификации элементов $S^{\circ}(K)$ позволяет более тонко анализировать кандидатов в содержания.

**2. Задача совмещения формальных контекстов**

Любой нетривиальный контекст $K$ способен порождать некоторые производные контексты методом вычеркивания из $Tab(K)$ отдельных строк-объектов и/или столбцов-признаков. Прямая задача построения контекстов, производных от заданного, затруднений не вызывает. Иначе обстоит дело с обратной задачей, которая (в несколько ослабленной формулировке) состоит в построении решетки $S(K)$ неизвестного контекста $K$ по заданным контекстам $K_1$ и $K_2$, производным от $K$. При всех различиях в дополнительных предположениях и целевых установках многие известные подходы [1, 5, 6] к решению задачи совмещения не претендуют на точное построение решетки $S(K)$. Для случая простых контекстов строгое, хотя и трудоемкое решение задачи совмещения обеспечивает метод @FC [7,8]. Представим этот метод.

Назовем глобальным контекстом неизвестный простой контекст $K = (G, M, \varphi(G))$. Заданную проекцию множества $G$ на фиксированное подмножество признаков $M_T$ назовем локальным контекстом. Формально, локальный контекст $K_T$ есть производный контекст вида

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

На практике получение $K_T$ может быть связано, например, со снаряжением дорогостоящей экспедиции к местам обитания изучаемых объектов $G$, причем исследовательские возможности экспедиции позволяют установить для этих объектов признаки из относительно ограниченного множества $M_T$.

Для глобального контекста $K$ и пары локальных контекстов $K_1 = (G_1, M_1, I_1)$ и $K_2 = (G_2, M_2, I_2)$ в случае $M_1 \bigcup M_2 = M$ метод @FC предлагает сначала построить множество кандидатов[2]

$$V = \{ B_1 \bigcup B_2 \mid B_1 \in S(K_1), B_2 \in S(K_2) \}, \quad (1)$$

[2] кандидатов в содержания понятий глобального контекста $K$

а затем пошагово выделить из $V$ элементы $S(K)$. На каждом шаге построения принимается решение о запросе нового локального контекста[3], и по итогам его специальной обработки для некоторых кандидатов принимаются классификационные решения. Процесс формирования $S(K)$ заканчивается, когда классификационные решения приняты для всех кандидатов. Доказано, что $S(K) \subseteq V$, и метод @FC позволяет построить множество $S(K)$ за конечное число шагов. Запрос нового локального контекста, оформленный в виде множества его признаков $M_T$, формируется автоматически или в диалоге с пользователем. При этом возможные варианты запросов рассчитываются заранее, и задача сводится к выбору лучшего варианта.

В части обработки локальных контекстов $K_T$ метод @FC использует классификационные правила, которые позволяют относить некоторые элементы $v$ множества $V$ к классам $S(K)$ или $S^{\circ}(K)$. Типичные классификационные правила выглядят так:

(K1) Если $v \in P(K_T)$, и $v \cup v' \in V$ для некоторого $v'$, то $v \cup v' \in P(K)$.

(K2) Если $v \in P^{\circ}(K_T)$, то $v \in P^{\circ}(K)$.

(K3) Если $v \in P^{\circ}(K_T)$, и $v \cup v' \in V$ для некоторого $v'$ из $M \setminus M_T$, то $v \cup v' \in S^{\circ}(K)$.

(K4) Если $v \in P^{\circ}(K_T)$, $\{v' \mid v \cup v' \in V\} = \{\varnothing, v'\}$, и $v' \neq \varnothing$, то $v \cup v' \in S(K)$.

(K5) Если $v \in S(K_T)$ и $\{v' \mid v \cup v' \in V\} = \{\varnothing\}$, то $v \in S(K)$.

С одной стороны, классификационные правила представляют собой строго доказанные достаточные условия [8], а с другой стороны – они вполне пригодны для использования в программах. Обработка очередного локального контекста заканчивается, когда все классификационные правила оказываются неприменимыми.

При высокой стоимости локальных контекстов возникает потребность в технологиях, позволяющих при их обработке классифицировать как можно больше кандидатов в содержания из $V$. 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 контролируемых предположениях существенно увеличить количество классификационных решений.

### 3. Основы экспертной классификации в терминах FCA

Метод экспертной классификации позволяет в диалоге с экспертом построить неизвестное множество объектов $G^E$ некоторого целевого простого контекста $(G^E, M, \varphi(G^E))$. При этом множество признаков $M$, задействованных в описаниях объектов, считается (a) известным, (b) состоящим из $k$ непересекающихся подмножеств $M_1^E$, ..., $M_k^E$ и (c) порождающим универсум

[3] то есть решение о снаряжении новой экспедиции

$$U_k^E = M_1^E \times ... \times M_k^E \quad (2)$$

такой, что $G^E \subseteq U_k^E$. Собственно говоря, на эксперта возлагается обязанность последовательно относить предъявляемые ему элементы универсума либо к $G^E$, либо к $U_k^E \setminus G^E$.

В интересах скорейшего построения множества объектов метод экспертной классификации постулирует гипотезу, согласно которой элементы каждого множества $M_i^E$ различаются по степени характерности для объектов $G^E$. Принимая эту гипотезу, эксперт способен заранее построить на элементах множества $M_i^E$ асимметричное, антирефлексивное и транзитивное отношение характерности $<_i$; запись "$m_{i1} <_i m_{i2}$" означает, что признак $m_{i1}$ менее характерен для объектов $G^E$, чем признак $m_{i2}$. Экспериментально установлено, что фраза "быть характерным для объектов $G^E$" понятна эксперту без разъяснений, и он вполне способен построить $k$ штук отношений характерности – по одному для каждого множества $M_i^E$.

Заметим, что в оригинальном описании экспертной классификации предполагается линейность всех отношений характерности. В настоящей работе вместо линейного порядка предполагается наличие частичного порядка, что сказывается на методе хранения данных и на организации диалога с экспертом при извлечении знаний об отношениях характерности.

Совокупность построенных экспертом отношений характерности $<_1$, ..., $<_k$ порождает рефлексивное, антисимметричное и транзитивное отношение доминирования $\leq$ на всевозможных элементах универсума $U_k^E$. По определению, пара подмножеств $u_1$ и $u_2$ связана отношением $u_1 \leq u_2$, если для любого $i = 1, ..., k$ выполняется одно из двух: либо $m_{i1} = m_{i2}$, либо $m_{i1} <_i m_{i2}$, где $u_1 \bigcap M_i^E = \{ m_{i1} \}$, $u_2 \bigcap M_i^E = \{ m_{i2} \}$.

Если отношение доминирования удается построить, то механизм переноса классификационных решений, полученных для некоторого объекта $g$ из $U_k^E$, описывается двумя вполне разумными правилами:

(П1) Если $g \in G^E$, и $g \leq u$, то $u \in G^E$;

(П2) Если $g \in U_k^E \setminus G^E$, и $u \leq g$, то $u \in U_k^E \setminus G^E$.

**4. Экспертная классификация и задача совмещения контекстов**

Метод @FC решения задачи совмещения простых контекстов оперирует локальными контекстами $K_1 = (G_1, M_1, I_1)$ и $K_2 = (G_2, M_2, I_2)$ в интересах построения решетки $S(K)$ неизвестного глобального контекста $K$. Дополнительно потребуем

$$M_1 \cap M_2 = \varnothing \quad (3)$$

и положим $M_1^E = S(K_1)$ и $M_2^E = S(K_2)$.

Между множеством кандидатов в содержания $V$ и формально построенным универсумом $U_2^E = M_1^E \times M_2^E$ (см. формулы (1) и (2) при $k=2$) существует взаимно однозначное соответствие

$$\Psi(v) = (v \bigcap M_1, v \bigcap M_2).$$

С учетом требования (3) отображение $\Psi$ представляет собой утверждение о возможности однозначной раскраски всех кандидатов из $V$ двумя цветами: признаки из $M_1$ окрашиваются, скажем, красным цветом, а признаки $M_2$ – зеленым.

При таком подходе отношение доминирования $\leq$, построенное для целевого простого контекста $(S^E, M_1^E \cup M_2^E, \varphi(S^E))$, где $S^E = \Psi(S(K))$, оказывается весьма полезным для решения задачи совмещения. Сконструированное экспертом/экспертами на универсуме $U_2^E$ отношение $\leq$ естественным образом порождает отношение доминирования $\leq$ на множестве $V$, а также модификации правил П1 и П2:

$$v_1 \leq v_2 \quad \Leftrightarrow \quad \Psi(v_1) \leq \Psi(v_2),$$

(П1м) Если $v' \in S(K)$, и $v' \leq v$, то $v \in S(K)$;

(П2м) Если $v \in S^\circ(K)$, и $v' \leq v$, то $v' \in S^\circ(K)$.

Вновь полученные классификационные правила П1м и П2м расширяют области действий правил типа К1 – К5. Так, в случае исполнения правила К3 все кандидаты из $\{x \mid x \leq v \cup v'\}$ зачисляются в $S^\circ(K)$. Аналогично, в случае срабатывания правила К5 все кандидаты из $\{x \mid v \leq x\}$ зачисляются в $S(K)$.

*Замечание 1.* Успех описанного подхода к усовершенствованию метода @FC за счет привлечения правил П1м и П2м находится в прямой зависимости от количества экземпляров отношений характерности $<_1$ и $<_2$, выявленных экспертом на множествах $S(K_1)$ и $S(K_2)$.

*Замечание 2.* Задача построения отношения характерности $<_1$, определенного на $S(K_1)$, осложняется большим[4] количеством элементов в $S(K_1)$, а также сложно устроенными описаниями этих элементов. В связи с этим наиболее перспективным “лобовым” подходом к построению отношений характерности является метод балльных оценок, выставляемых элементам $S(K_1)$.

*Замечание 3.* Если множество признаков $M_1$ допускает разбиение $M_1^1 \cup ... \cup M_1^m$ на подклассы значений некоторых однозначных атрибутов, то отношение характерности $<_1$, определенное на $S(K_1)$, можно сконструировать из отношения доминирования $\leq_1$,

[4] нетипично большим для традиционных задач экспертной классификации

определенного на $M_1^1 \times ... \times M_1^m$, путем удаления из $\leq_1$ всех пар вида $u \leq_1 u$. Замечания 2 и 3 в равной степени относятся к отношению $<_2$, определенному на $S(K_2)$.

*Замечание 4.* В ходе решения задачи совмещения контекстов помимо первоначально построенного отношения доминирования $\leq$, полученного при посредничестве универсума $U_2^E$, могут появиться и другие пригодные для использования отношения доминирования. Организация непротиворечивого сосуществования двух и более отношений доминирования составляет предмет самостоятельной задачи.

## Заключение

По большому счету, отношение доминирования, построенное методом экспертной классификации, является эвристикой, сокращающей перебор кандидатов в содержания понятий. В связи с этим возникает необходимость хранить для каждого принятого классификационного решения историю его вывода с тем, чтобы иметь возможность отказаться от неподтвердившихся решений. По завершении итерационного процесса истории выводов позволяют вывести итоговую оценку достоверности построенного множества содержаний.

**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. Guan-yu L., Shu-peng L., Yan Z. Formal Concept Analysis based Ontology Merging Method // Proceeding of 3rd IEEE International Conference on Computer Science and Information Technology. Vol. 8, 2010. 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., Vol. 183, Formal Ontology in Information Systems, 2008. P. 311-323.
7. Стельмашенко Д.Е. Алгоритм восстановления глобального контекста // Сборник статей молодых ученых факультета ВМК МГУ. Вып. 9, 2012. С.191-209.
8. Соловьев С.Ю., Стельмашенко Д.Е. Подходы к исследованию формальных контекстов // Информационные процессы. Том 11, № 2, 2011. С. 277-290.