СУБЪЕКТИВНЫЕ АСПЕКТЫ ФОРМИРОВАНИЯ И ОБРАБОТКИ ДАННЫХ В АНАЛИЗЕ ФОРМАЛЬНЫХ ПОНЯТИЙ

Д.Е. Самойлов [1], С.В. Смирнов [2]

[1] Самарский государственный аэрокосмический университет им. С.П. Королёва (научно-исследовательский университет), Самара, Россия,
[2] Институт проблем управления сложными системами РАН, Самара, Россия

В работе дается краткий обзор субъективных аспектов формирования и обработки данных в анализе формальных понятий. Показывается, что фундаментальная когнитивная процедура шкалирования, допуская различную интерпретацию, вводит в анализ новую информацию, без внимания к которой анализ в общем случае не будет корректным. Рассматриваются подлежащие учету зависимости между исследуемыми свойствами объектов, которые возникают при использовании шкал различного вида.

Ключевые слова: анализ формальных понятий, шкалирование, ограничение существования свойств.

Введение

Уже более трех десятилетий на стыке прикладной математики и компьютерных наук успешно развивается анализ формальных понятий (АФП) [1, 2]. Благодаря классическому (аристотелевскому) подходу к понятию как к фундаментальной ментальной сущности, определяемой объёмом и содержанием, и опоре на алгебраическую теорию решёток АФП внёс значительный вклад и продолжает стимулировать развитие анализа данных, представления знаний и других разделов информатики.
Когнитивный характер АФП проявляется в учете различных аксиологических установок исследователя. Очерк таких субъективных аспектов АФП и его приложений в анализе данных входит в задачу данной статьи, но главное внимание уделяется шкалированию первичных данных. Считаем, что генезис так называемых ограничений существования свойств [3], без учета которых решение задач АФП оказывается некорректным [4, 5], зачастую определяется именно процедурами шкалирования. В работе исследуется возникновение различных ограничений существования свойств в результате субъективного выбора типа шкал [6].

1. Субъективные аспекты классического АФП

1.1. Основные определения и модели

АФП имеет дело с массово встречающимися практическими приложениями, которые требуют анализа объектно-признаковых данных. Классический АФП ориентирован на обработку бинарного представления этих данных в виде совокупности оценок истинности семантических суждений вида b_{gm} = «объект g обладает свойством m» и использует следующие обозначения и модели:

- $\boldsymbol{K} = (G^*, M, I)$ – формальный контекст, где G^* - набор объектов изучаемой предметной области (ПрО), попавших в поле зрения исследователя (т.е. «обучающая выборка»

объектов ПрО), M - множество измеряемых у объектов свойств, I - бинарное соответствие «объекты-свойства» - совокупность оценок $\|b_{gm}\| \in$ {**Истина**, **Ложь**};

- операторы Галуа φ, ω (общая нотация «'») для контекста $\boldsymbol{K}$:
  $\varphi(X) = X' = \{m \mid m \in M, \forall g \in X\ ((g, m) \in I)\}$ - общие свойства объектов, составляющих $X \subseteq G^*$;
  $\omega(Y) = Y' = \{g \mid g \in G^*, \forall m \in Y\ ((g, m) \in I)\}$ - объекты, которые обладают всеми свойствами из $Y \subseteq M$;
  для множества объектов X множество их общих признаков X' служит описанием сходства объектов из множества X, а замкнутое множество X'' является кластером сходных объектов;
- (X, Y) – формальное понятие, у которого $X \subseteq G^*$ - объем, $Y \subseteq M$ - содержание, причем $X = Y'$, $Y = X'$;
- $\boldsymbol{B}(\boldsymbol{K})$ - множество формальных понятий контекста $\boldsymbol{K}$;
- $(\boldsymbol{B}(\boldsymbol{K}), \leq)$ – замкнутая решетка понятий, где $(X_1, Y_1) \leq (X_2, Y_2)$, если $X_1 \subseteq X_2$ (или $Y_1 \supseteq Y_2$).

Субъективный аспект формирования контекста $\boldsymbol{K}$ проявляется в когнитивной асимметрии «объектов» и «свойств»: формально объекты G^* независимы от исследователя ПрО, тогда как свойства M - результат продуцирования субъектом гипотез о ПрО на основе его текущих целевых установок, имеющейся у него совокупности априорных знаний и ресурсных возможностей.

1.2. Редуцирование решетки формальных понятий

Представление результатов АФП для последующего анализа может быть затруднено из-за большого количества выявляемых понятий. Для редуцирования решетки $(\boldsymbol{B}(\boldsymbol{K}), \leq)$ разработаны два основных способа отбора релевантных формальных понятий [7, 8].

Поддержкой множества свойств $Y \subseteq M$ для данного контекста $\boldsymbol{K}$ называется величина

$$\mathbf{supp}(Y) = |Y'| / |G^*|.$$

Множество $Y \subseteq M$ называется частым множеством свойств, если $\mathbf{supp}(Y) \geq minsupp \in [0, 1]$.

Если в решетке понятий сохранить лишь частые понятия, содержание которых - частые множества свойств, то она сократиться до так называемого «айсберга понятий» [7].

Более тонкий подход связан с выявлением в $(\boldsymbol{B}(\boldsymbol{K}), \leq)$ понятий, устойчивых к изменению объёма поддержки в обучающей выборке объектов [8].

Индекс устойчивости формального понятия (X, Y) определяется выражением

$$\sigma(X, Y) = |\{Z \subseteq X \mid Z' = Y\}| / |2^X|.$$

Понятие (X, Y) считается устойчивым, когда $\sigma(X, Y) \geq \sigma_{min} \in [0, 1]$, и редуцирование решетки означает сохранение в ней лишь наиболее устойчивых формальных понятий.

Очевидно, что субъективный характер выбора порогов и для поддержки множества свойств, и для индекса устойчивости понятий не связан с вовлечением в анализ данных дополнительной информации (знаний) о ПрО.

1.3. Импликации на подмножествах свойств

Импликация на подмножествах свойств формального контекста ***K*** есть зависимость вида $A \to B$, $A, B \subseteq M$, при условии, что все объекты, обладающие свойствами A, также обладают всеми свойствами из B, т.е. $A' \subseteq B'$. Частичная импликация в контексте ***K*** отличается тем, что не располагает полной поддержкой в обучающей выборке объектов [1, 2].
Вводимый в АФП показатель достоверности (*confidence*) частичных импликаций позволяет, субъективно выбрав порог достоверности, расширить множество релевантных эмпирических закономерностей, но, как и прежде, не сопровождается использованием дополнительных данных о ПрО.

2. Субъективные аспекты шкалирования свойств

Базовой формой эмпирической информации о ПрО служит таблица «объекты-свойства» [6, 9], которая трактуется в АФП как многозначный контекст (G^*, M, V, $\boldsymbol{I}$), где G^* и M имеют уже указывавшийся смысл, V - множество значений свойств, а $\boldsymbol{I}$ - тернарное отношение между G^*, M и V ($\boldsymbol{I} \subseteq G^* \times M \times V$), определенное для всех пар из $G^* \times M$.
Для приведения многозначного контекста к бинарному виду применяют фундаментальную когнитивную процедуру - концептуальное шкалирование [1], которая неформально означает субъективное конструирование «покрытия» домена значений каждого свойства многозначного контекста, т.е. образование новых отличительных свойств объектов ПрО, измеряемых в субъективно формируемых шкалах.
Шкалой свойства $m \in M$ является бинарный контекст $\boldsymbol{S}_m = (G_m, M_m, I_m)$, где G_m – значения шкалы, M_m – вводимые шкалой новые свойства объектов ПрО, I_m – соответствие, выражающее особенности субъективного восприятия ПрО её исследователем.
Покажем, что, осуществляя концептуальное шкалирование, субъект вводит в анализ качественно новую информацию о ПрО, без учёта которой практическое применение АФП становится проблематичным (решению возникающих при этом задач посвящены работы [4, 5]).

2.1. Использование номинальной шкалы

Наиболее распространённым приёмом шкалирования служит использование номинальной шкалы, или бинарной шкалы наименований [6]. Пример такой шкалы дает таблица 1.

Табл. 1. Шкала роста мужчин

Рост мужчин, см	Низкий	Средний	Высокий
< 168	×		
168-175		×	
> 175			×

Покрытие исходного домена значений шкалируемого свойства в этом случае строго дизъюнктивно; образцом более сложного подхода может быть использование нечеткой шкалы наименований.
В любом варианте очевидно, что вводимым номинальной шкалой свойствам объектов ПрО присуща парная несовместимость E [3, 4] (например, E(Низкий, Высокий)). И это -

новая существенная информация о ПрО, которую добавляет исследователь к уже имеющимся данным в исходном многозначном контексте.

2.2. Другие типы шкал

Эффекты применение других типов шкал опишем на примерах из [10], конечно, не исчерпывающих способы выражения исследователем своего субъективного восприятия ПрО.

Порядковую шкалу целесообразно использовать для сохранения упорядоченности значений в домене многозначного свойства.

Так домен многозначного свойства «Материальное положение (МП)» может быть описан следующими выражениями (от тяжелого до благополучного) [10]:

1) денег не хватает даже на питание;
2) на питание денег хватает, но не хватает на покупку одежды и обуви;
3) на одежду и обувь денег хватает, но приобретение бытовой техники позволить не можем;
4) денег вполне хватает на приобретение бытовой техники, но не можем купить новый автомобиль;
5) денег хватает на всё, кроме таких дорогих приобретений как квартира, дом;
6) материальных затруднений не испытываем, при необходимости могли бы приобрести квартиру, дом.

Для исследователя естественной шкалой для такого многозначного свойства будет таблица 2.

Табл. 2. Шкала материального положения

≤	$МП_1$	$МП_2$	$МП_3$	$МП_4$	$МП_5$	$МП_6$
1	×					
2	×	×				
3	×	×	×			
4	×	×	×	×		
5	×	×	×	×	×	
6	×	×	×	×	×	×

Это шкалирование устанавливает между вновь введенными свойствами бинарное отношение обусловленности C [3, 5]: $i < k \leftrightarrow C(МП_i, МП_k)$.

В настоящее время весьма популярны шкалы с разделением и порядком, описанные в [10] на примере закрытого вопроса «Чувствуете ли Вы себя в безопасности? (Б)» с вариантами ответа:

1) безусловно да;
2) скорее да;
3) скорее нет;
4) безусловно нет.

Субъективное понимание значений этого домена значений может быть выражено двупорядковой шкалой, доставляемой таблицей 3.

Табл. 3. Шкала безопасности

	$Б_1$	$Б_2$	$Б_3$	$Б_4$
1	×	×		
2		×		
3			×	
4			×	×

В этом примере исследователь субъективно расширяет имеющиеся эмпирические данные о ПрО, вводя следующие бинарные отношения между вновь вводимыми свойствами:

- $E = \{(Б_1, Б_3), (Б_1, Б_4), (Б_2, Б_3), (Б_2, Б_4)\}$;
- $C = \{(Б_1, Б_2), (Б_4, Б_3)\}$.

Заключение

Анализ субъективных аспектов формирования и обработки данных в АФП необходимо начинается указанием на аксиологические основы формирования исходных данных о ПрО.

Субъективно устанавливаемые пороговые значения различных показателей как правило используются для формирования классов эквивалентности получаемых результатов и непосредственно интерпретируются в рамках АФП.

Фундаментальная когнитивная процедура шкалирования, напротив, связана с привнесением субъектом дополнительной информации об исследуемой ПрО, которая должна быть учтена ещё на этапе формирования бинарного формального контекста АФП [4, 5] и существенно влияет на выводимую структуру формальных понятий.

Разумеется, генезис ограничений существования свойств не исчерпывается субъективными действиями исследователя при конструировании шкал для значений свойств, наблюдаемых у объектов обучающей выборки. В общем плане источником этих ограничений служат априорные знания субъекта, релевантные исследуемой ПрО.

Благодарности

Работа выполнена при проведении исследований по теме «Модели и методы формирования согласованной системы понятий о предметной области управления в процессах коллективного принятия решений» в рамках государственного задания Институту проблем управления сложными системами РАН на 2016 год, а также при государственной поддержке Программы повышения конкурентоспособности Самарского государственного аэрокосмического университета среди ведущих мировых научно-образовательных центров на 2013-2020 годы.

Литература

1. Ganter, B. Formal Concept Analysis. Mathematical foundations / B. Ganter, R. Wille. - Berlin-Heidelberg: Springer-Verlag, 1999. - 290 p.
2. Formal Concept Analysis Homepage [Электронный ресурс]. – URL: - http://www.upriss.org.uk/fca/fca.html (дата обращения 30.04.2016).
3. Lammari, N. Building and maintaining ontologies: a set of algorithms / N. Lammari, E. Metais // Data & Knowledge Engineering. - 2004. - Vol. 48(2). - P. 155-176.

4. Офицеров, В.П. Метод альфа-сечения нестрогих формальных контекстов в анализе формальных понятий / В.П. Офицеров, В.С. Смирнов, С.В. Смирнов // Проблемы управления и моделирования в сложных системах: Труды XVI междунар. конф. (30 июня - 03 июля 2014 г., Самара, Россия). – Самара: СамНЦ РАН, 2014. - С. 228-244.
5. Семенова, В.А. Модели и методы интеллектуального анализа неполных данных для построения формальных онтологий / В.А. Семенова, С.В. Смирнов // Информационные технологии и нанотехнологии (ИТНТ-2015) [Электронный ресурс]: Материалы Международной конф. и молодежной школы (29 июня-1 июля 2015 г., Самара, Россия). - Самара: Изд-во СамНЦ РАН, 2015. - С. 194-198.
6. Загоруйко, Н.Г. Прикладные методы анализа данных и знаний / Н.Г. Загоруйко. – Новосибирск: Изд-во Института математики СО РАН, 1999. - 270 с.
7. Stumme, G. Computing Iceberg Concept Lattices with Titanic / G. Stumme, R. Taouil, Y. Bastide, N. Pasqier, L. Lakhal // Journal on Knowledge and Data Engineering. – 2002. – Vol. 42(2). – P. 189-222.
8. Кузнецов, С.О. Устойчивость как оценка обоснованности гипотез, получаемых на основе операционального сходства / С.О. Кузнецов // НТИ. Сер.2. – 1990. – №12. – С. 21-29.
9. Барсегян, А.А. Анализ данных и процессов / А.А. Барсегян, М.С. Куприянов, И.И. Холод, М.Д. Тесс, С.И. Елизаров. - СПб.: БХВ-Петербург, 2009. – 512 с.
10. Игнатов, Д.И. Решетки формальных понятий для анализа данных социологических опросов / Д.И. Игнатов, О.Н. Кононыхина // Интегрированные модели и мягкие вычисления в искусственном интеллекте: Труды V-й Международной н.-т. конф. (20-30 мая 2009 г., Коломна, Россия). Т 1. – М.: Физматлит, 2009. – С. 230-240.