УДК 681.332

ПОМЕХОУСТОЙЧИВОСТЬ КОДИРОВАНИЯ ЭВРИСТИК В ИНТЕЛЛЕКТУАЛЬНОЙ БАЗЕ ДАННЫХ

к.т.н. Т.Н. Новожилова
(представил д.т.н., проф. Г.Г. Асеев)

Рассмотрены модели экспертного опроса, обеспечивающие помехоустойчивость интегрированного логического вывода при автоматизации кодирования экспертных знаний по анализу финансовой деятельности в среде приобретения знаний.

Выдвижение гипотез является неотъемлемой частью эвристического поиска в интеллектуальной базе данных (ИБД): эксперт выдвигает гипотезы относительно частичных решений, чтобы когнитолог их проверил на общих. Главная трудность кодирования экспертных знаний в ИБД – определение неверных гипотез (помех) и эффективный возврат от них в исходное состояние. Кроме того, следует понимать, что “нулевая гипотеза”, часто используемая как удобный способ формулирования гипотезы эксперимента, никогда не может быть “принята” в свете полученных данных. Она может быть только или “отвергнута”, или “не отвергнута”. Эта точка зрения согласуется со всеми положениями юмовской философии науки, которые подчеркивают *невозможность дедуктивного доказательства индуктивных законов*. Чтобы установить закономерности вывода эвристических решений, необходимо свести данные опроса к моделям экспертов: выдвигается гипотеза **X** и делается попытка определить модель вывода [1]. Автоматизация приобретения знаний предполагает распространение *эвристик* и функциональных зависимостей ($\mathbf{B_i}$) не только на другие $\mathbf{DS_i}$ (ситуации) или периоды времени, но и на те воздействия, которые в теории, будто идентичны **X**, но в действительности отличаются от **X** теоретически несущественными элементами

$$\mathbf{X}=\{(\mathbf{B_i} <\mathbf{DS_1}>)\},\ \mathbf{DS_1} \Rightarrow \mathbf{X} \models \mathbf{DS_1} \vee \mathbf{DS_2} \Rightarrow \mathbf{X}. \qquad (1)$$

Так, если мы воспользуемся многочисленными независимыми реализациями **X**, то специфические иррелевантные детали не будут воспроизводиться каждый раз в том же виде, и *интерпретация* будет иметь больше шансов оказаться правильной [2]. Помехоустойчивость информационного обмена в среде обозначим как валидность (табл. 1). Применяется следующая система графических и символьных обозначений: **X** - событие, влияние которого подлежит опросу; **O** - обозначает как саму процедуру опроса, так и ее результаты; **X** и **O**, в одной строке, относятся к одним и тем же конкретным лицам; направление слева направо - временной порядок, а расположение **X** и **O** одно под другим - одновременность. **R** - случайное распределение экспертов по

разным режимам эксперимента (такая рандомизация - универсальный метод уравнивания групп перед введением воздействия в известных статистических границах); “ - ” - безусловная слабость плана; “ + ” - фактор под контролем; “ ? “ - возможность осложнений; пробел - иррелевантность фактора.

Таблица1

Помехоустойчивость кодирования эвристик в ИБД

Схемы опроса в ИБД	Источники невалидности (помех)											
	Внутренней								Внешней			
	1	2	3	4	5	6	7	8	9	10	11	12
1. XO	-	-				-	-			-		
2. O_1XO_2	-	-	-	-	?	+	+	-	-	-	?	
3. $\underline{XO_1}$ O_2	+	?	+	+	+	-	-	-		-		
4. RO_1XO_2 $RO_3\ O_4$	+	+	+	+	+	+	+	+	-	?	?	
5. RO_1XO_2 $RO_3\ O_4$ $R\ XO_5$ $R\ O_6$	+	+	+	+	+	+	+	+	+	?	?	
6. RXO_1 $R\ O_2$	+	+	+	+	+	+	+	+	+	?	?	
7.$O_1O_2O_3\ O_4XO_5\ O_6O_7O_8$	-	+	+	?	+	+	+	+	-	?	?	
8. $X_1O_1\ X_0O_2\ X_1O_3\ X_0O_4$	+	+	+	+	+	+	+	+	-	?	-	-
9. $DS_a\ X_1O\ DS_b\ X_0O$	+	+	+	+	+	+	+	+	-	?	?	-
10. $\underline{O_1XO_2}$ $O_3\ O_4$	+	+	+	+	?	+	+	-	-	?	?	
11. Сбалансированные $\underline{X_1O\ X_2O\ X_3O\ X_4O}$ $\underline{X_2O\ X_4O\ X_1O\ X_3O}$ $\underline{X_3O\ X_1O\ X_4O\ X_2O}$ $X_4O\ X_3O\ X_2O\ X_1O$	+	+	+	+	+	+	+	?	?	?	?	-
12. $R\ O_1\ (X)$ $R\ \ X\ O_2$	-	-	+	?	+	+	-	-	+	+	+	
13. $R\ \ O_1\ (X)$ $\underline{R\ \ X\ O_2}$ $R\ O_3$ $R\ \ O_4$	+	+	+	+	+	+	+	-	+	+	+	

Параллельные строки, не разделенные пунктиром, - уравненные посредством рандомизации группы экспертов; пунктирной линией отделяются друг от друга группы, не уравненные рандомизацией.

В ИБД рассматриваются следующие классы внешних переменных (неконтролируемых конкурентных гипотез: если эти переменные не контролируются в плане опроса, то они могут дать эффекты, которые смешиваются с воздействием **X**) [2]: 1**)** фон; 2) естественное развитие; 3) эффект тестирования; 4) инструментальная погрешность; 5) статистическая регрессия; 6) со-

став групп экспертов; 7) выбывание; 8) взаимодействие фактора отбора с естественным развитием. Факторы, ставящие под угрозу репрезентативность эксперимента: 9) реактивный эффект; 10) взаимодействие состава групп и **X**; 11) реакция экспертов на эксперимент, обусловленная особенностью мышления эксперта; 12) взаимодействие между разными **X**.

Схемы опроса в ИБД - средство проверки каузальных отношений, если:

1) соблюдается временная последовательность между двумя переменными **X** и **O**: причина должна предшествовать по времени следствию;

2) воздействия статистически связаны с эффектом, так как если возможная причина и эффект не связаны друг с другом, одно не может быть причиной другого. Разработаны критерии для принятия решения о существовании "действительной" ковариации в результатах (статистические показатели действуют как фильтры). Но, к сожалению, они подвержены ошибкам даже в тех случаях, когда используются в ИБД должным образом, с их помощью не всегда удается выявить как действительно существующую, так и ложную композицию ковариаций. Поэтому логично было бы показать когнитологу причины, которые порождают неверные выводы о ковариации.

3) нет правдоподобного альтернативного объяснения появления **O** помимо **X**. Это условие соблюсти труднее, что, главным образом, связано с валидностью альтернативных интерпретаций.

Для предупреждения ошибок интерпретации при использовании планов 1,2,3 полезно провести реверсивный контрольный анализ. Отсутствие полной корреляции данных может быть вызвано как «ошибками», так и систематическими источниками дисперсии, характерными для того или иного количественного показателя. Поэтому привлекательна идея о исследовании в среде приобретения знаний типов мышления и восприятия информации [1]. Результаты экспертного опроса после их типизации переносятся на модели экспертов. Поскольку содержание, вкладываемое в понятия, может меняться, в ИБД используются следующие отношения:

$$\mathbf{X_1 \alpha X_2 \Leftrightarrow "L(X_1) \subseteq L(X_2)"},\ \mathbf{XV\beta \Leftrightarrow "V^+ \subseteq L(X)},\ \mathbf{V^- \cap L(X) = \varnothing"} \quad (2)$$

где $\boldsymbol{\alpha}$ - отношение предпочтения (общего **ранга** словоупотребления понятий, из которых складывается **X)**: наименование; понятие; дисциплинарное; междисциплинарное; $\boldsymbol{\beta}$ - отношение совместимости **форм** словоупотребления понятий **X**: простое, составное, сложное, комплексное, системное, метасистемное; $\mathbf{V^+}$, $\mathbf{V^-}$ - множества позитивных и негативных примеров из **подпространства знаний** $\mathbf{V \in H_0(X)}$: обыденное; экспертное; теоретическое; категориальное; методологическое; эзотерическое; дидактическое; общее; **L (X)**- логика, определяемая **X** [3].

При кодировании эвристик решается вопрос об уровне детализации представления **X**: в отношении всех атрибутов устанавливается терминальный уровень ($\mathbf{H_3}$), семантика которого учитывается схемой ИБД. Основные символьные структуры $\mathbf{H_3}$ - векторы атрибутов и объекты со связанными с ними парами атрибут – значение (**а - з**) [4]:

$$\langle элемент \rangle ::= \langle элемент-вектор \rangle \mid \langle аз-элемент \rangle$$

$$\langle элемент-вектор \rangle ::= \left(\{ \langle значение \rangle \}^{\div} \right)$$

$$\langle аз-элемент \rangle ::= \left(\langle объект \rangle \{ \langle атрибут \rangle \langle значение \rangle \}^{\div} \right).$$

Семантика атрибутов уточняется с помощью любого необходимого числа добавочных атрибутов в объектной составляющей, отражающей предикатную структуру $\mathbf{H_{0i}}$ (описания $\mathbf{DS_j}$ серией бухгалтерских проводок). Объект без свойств (как самостоятельная сущность) в ИБД не существует.

Под экстенсионалом модели эксперта $\mathbf{EXT(H_{0i},\ \Delta IS(X),\ t)}$ будем понимать множество всех фактов, принимаемых экспертом во внимание (где $\mathbf{H_{0i}}$ - множество аксиом модели эксперта, характеризующих изменения функциональных зависимостей $\mathbf{\Delta B/\Delta t}$; $\mathbf{\Delta IS}$- прирост индикатора устойчивости). Под интенсионалом модели эксперта $\mathbf{H_{0i}}$ будем понимать функцию $\mathbf{INT(X_i)}$, которая вырабатывает множество $\mathbf{X_i(\Delta B/\Delta t) \in C}$ (где $\mathbf{C}$ - открытое множество фактов, характеризующих с различных сторон $\mathbf{DS_j}$), являющихся конкретизацией подпространства $\mathbf{H_0}$ при $\sum_{j=1}^{72} \mathbf{DS_j} \xrightarrow{\mathbf{B}} \mathbf{H_{0i}}$ [5]. Рассмотрим некоторые элементарные операции ИДБ, приводящие $\mathbf{C}$ к разделению на классы $\mathbf{H_i=\{H_{0i},\ H_{1i},\ H_{2i}\}}$. Разбиение $\mathbf{C}$ определено не полностью, задана лишь некоторая информация $\mathbf{i_0 \in DS_j}$ о классах $\mathbf{H_i}$. Задача состоит в том, чтобы по $\mathbf{i_0 \in DS_j}$ и $\mathbf{EXT(H_{0i},\ \Delta IS(X),\ t)}$ определить значения $\mathbf{INT(X_i) : X_i \in H_{0i}}$.

Переход от логики бухучёта $\mathbf{L_0(H_0,\ DS_j\, B_0, \alpha_0, \mu_0, \beta_0)}$, где $\mathbf{H_0^{\pm} = \overline{B_0^{\pm} DS_j}}$, к логике финансистов $\mathbf{L_1(H_0 \dot{+} H_1,\ DS_j, B_0 \dot{+} B_1, \alpha_0 \dot{+} \alpha_1, \mu_0 \dot{+} \mu_1, \beta_0 + \beta_1)}$, где $\mathbf{H_1^{\pm} = \overline{B_1^{\pm} DS_j}}$ – *тривиальное расширение*, от $\mathbf{L_1}$ к $\mathbf{L_0}$ – тривиальное сокращение, $\mathbf{\mu_0(\{X_0, X_1\}, \{\alpha_0, \alpha_1\}, \{\beta_0, \beta_1\}) = \mu_0(X_0, \alpha_0, \beta_0) + \mu_1(X_1, \alpha_1, \beta_1)}$, где $\mathbf{X_0, \alpha_0, \beta_0 \in H_0, X_1, \alpha_1, \beta_1 \in H_1}$, $\mathbf{\mu_0}$ - метрика $\mathbf{H_0 \dot{+} H_1}$. Переход от $\mathbf{L_1}$ к языку статистической теории $\mathbf{L_2(H_0 \dot{+} H_2,\ DS_j, B_0 \dot{+} B_2, \alpha_0 \dot{+} \alpha_2, \mu_0 \dot{+} \mu_2, \beta_0 \dot{+} \beta_2)}$, где $\mathbf{H_2^{\pm} = \overline{B_2^{\pm} DS_j}}$ назовём *нейтральным расширением*, от $\mathbf{L_2}$ к $\mathbf{L_0}$ – нейтральным сокращением.

Переход от $\mathbf{L_0(H_0,\ DSj,\ B_0,\ \alpha_0,\ \mu_0,\ \beta_0)}$, к логике интегрированного логического вывода ИДБ (язык когнитолога) $\mathbf{L_3(H_3,\ DSj,\ B_3,\ \alpha_3,\ \mu_3,\ \beta_3)}$, где $\mathbf{B_3 = UB_0 \oplus UB_1 \oplus UB_2}$; $\mathbf{H_3 = UH_0}$; $\mathbf{\mu_3(X_3, \alpha_3, \beta_3) = \mu(U^{-1}X_3, U^{-1}\alpha_3, U^{-1}\beta_3)}$, а $\mathbf{U}$ – оператор ИБД, отображающий $\mathbf{H_0}$ на $\mathbf{H_3}$, назовем *операцией изометрии.*

Для $\sum_{j=1}^{72} \mathbf{DS_j} = \overline{\overline{B_3^{\pm}} \overline{H_3^{\pm}}}$, где $\mathbf{H_3^{\pm} = \sum H_{0i}^{\pm} \oplus \sum H_{1i}^{\pm} \oplus \sum H_{3i}^{\pm}}$ перепишем $\mathbf{L_3}$:

$$L_3 = (H_3^+ \dot{+} H_3^-,\ DS_j, B_3^+ \dot{+} B_3^-, \alpha_3^+ \dot{+} \alpha_3^-, \mu_3^+ \dot{+} \mu_3^-, \beta_3^+ \dot{+} \beta_3^-) \qquad (3)$$

Таким образом, на основе (2) и (3) ИБД предоставляет когнитологу возможность перехода от представления $\mathbf{DS_j}$ на языке бухгалтерского учета, который понятен только бухгалтерам, на другие языки представления экономической информации: язык финансистов и описание $\mathbf{DS_j}$ через понятия о среднем, дисперсии, корреляции, типе распределения. При анализе многозначных зависимостей базис многозначной зависимости рассматривается в том же смысле, что и замыкание атрибута для функциональных зависимостей.

Применение плана 4 позволяет контролировать семь первых конкурентных гипотез (таблица 1). Но если экспериментальная группа подвергается однократному экспериментальному воздействию и контрольная группа также исследуется однократно, то посторонние частные события могут быть конкурентными объяснениями отличия $\mathbf{O_1}$-$\mathbf{O_2}$ от $\mathbf{O_3}$-$\mathbf{O_4}$. Возможный выход - *рандомизация* отдельных сеансов эксперимента с учетом тех ограничений, которые связаны с необходимостью уравнивания источников смещения. Естественное развитие и эффект тестирования контролируются постольку, поскольку они одинаковым образом проявляется в экспериментальной и контрольной группах. Одно из средств предупреждения ошибок интерпретации из-за регрессионных артефактов - проведение *параллельного анализа* экстремальных показателей предварительного тестирования в контрольной группе и использование этих данных при интерпретации изменения значений показателей. Действие фактора отбора исключается в той степени, в какой рандомизация обеспечивает эквивалентность групп в момент **R**.

Наиболее применяемая процедура обработки для плана 4 состоит в определении для каждой группы приращений показателей от предварительного тестирования и вычислении **t**-критерия для приращений, наблюдаемых в экспериментальной и контрольной группах. В большинстве случаев когнитологу следует предпочесть *ковариационный анализ*, в котором показатели предварительного тестирования берутся в качестве сопутствующих переменных и которому предшествует процедура формирования рандомизированных блоков или «распределения по уровням» показателей предварительного тестирования. Использование этого более точного анализа предпочтительно.

Большинство исследований по приобретению знаний перегружены «ошибками первого рода»: *констатация эвристик, не подтвержденных перекрестной валидизацией*. План 5 - первая попытка эксплицитно учитывать факторы внешней валидности (таблица1). Путем параллельного использования элементов плана 4, когда предварительное тестирование не проводится ни в экспериментальной, ни в контрольных группах, когнитологом может быть определен не только главный эффект тестирования, но и эффект его взаимодействия с **X**: расширяются возможности обобщения результатов (1); эффект **X** воспроизводится четырьмя различными способами: $\mathbf{O_2}>\mathbf{O_1}$; $\mathbf{O_2}>\mathbf{O_4}$; $\mathbf{O_5} > \mathbf{O_6}$ и $\mathbf{O_5} > \mathbf{O_3}$. Если все эти четыре соотношения выполняются, то правомерность вывода эвристики значительно возрастает. Обобщению резуль-

татов опроса косвенно способствует и то, что применение плана 5 позволяет выяснить общую вероятность взаимодействия тестирования и **X**, благодаря чему облегчается интерпретация данных прошлых и будущих опросов по плану 4. Точно так же сравнение $\mathbf{O_6}$ с $\mathbf{O_1}$ и $\mathbf{O_3}$ позволяет выявить комбинированный эффект естественного развития и фона. Некоторые возможные результаты для серии периодических опросов представлены на рис.1.

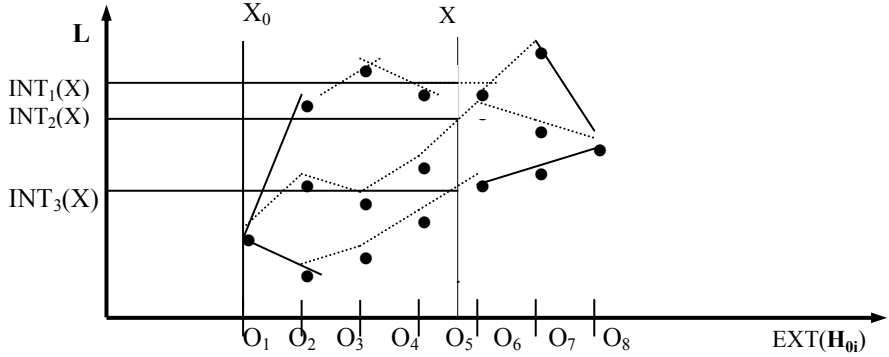

Рис.1. Результаты опроса мнений экспертов в точке X

Единой статистической процедуры, в которой одновременно использовались результаты всех наблюдений, не существует. Асимметрия этого плана исключает дисперсионный анализ приращений. Можно обработать данные итогового тестирования согласно простой схеме 2x2 дисперсионного анализа (не обращая внимания на предварительные тестирования, отведя им роль лишь дополнительного параметра воздействия **X**). Если главный эффект предварительного тестирования и эффект взаимодействия настолько малы, что ими можно пренебречь, желательно провести ковариационный анализ $\mathbf{O_4}$ и $\mathbf{O_2}$, используя результаты предварительного тестирования в качестве сопутствующей переменной.

В рамках доверительных пределов, устанавливаемых статистической моделью ИБД, рандомизации достаточно и без предварительного тестирования, чтобы удостовериться в «равенстве» экспериментальной и контрольной групп к введению дифференциального режима эксперимента. Так, при экспертных исследованиях приходится предлагать экспертам для оценки совершенно новые способы вывода решений и в этой обстановке предварительное тестирование в обычном смысле слова невозможно. Схема 6 годится для этих случаев и представляет собой как бы половину плана 5, при этом в нем контролируется как главный эффект тестирования, так и его взаимодействие с **X**, но в отличие от плана 5 они не измеряются. Если доступны данные, полученные до введения **X**, ими когнитологу нужно воспользоваться для формирования блоков, определения факторных уровней или как сопутствующими переменными. Это следует сделать так как, *во-первых*, статистические модели ИБД, соответствующие плану 4 обладают большей мощностью, чем те, кото-

рые применяются для плана 6; *во-вторых*, доступность таких показателей позволяет проверить наличие взаимодействия **X** с начальным уровнем и тем самым составить более полное представление о возможности обобщения результатов. Применение t-критерия при обработке плана 6 - оптимально. Ковариационный анализ и формирование блоков по таким исходным показателям, как квалификация, оценки в тестах, профессиональные интересы, позволяют повысить мощность статистической процедуры примерно до того же уровня, что и при использовании предварительного тестирования. Причем, такой надежный комплексный показатель, как *коэффициент интеллекта*, может оказаться лучше короткого предварительного тестирования.

На концептуальной основе планов 4 и 6, добавляя к ним новые группы с новыми **X**, можно построить сложные факторные планы. В обычном однофакторном дисперсионном анализе когнитолог будет иметь несколько «уровней» воздействия ($\mathbf{X_1}$, $\mathbf{X_2}$, $\mathbf{X_3}$, и т.д., а также, возможно, **X** «без **X**»). Поскольку приобретаются эвристики высококвалифицированных специалистов в области анализа финансовой деятельности, которые (по понятным всем причинам) не спешат делиться своим опытом, в последующих схемах когнитолог рассматривает взаимодействие, перекрестные планы и планы с переключением, а так же конечные, фиксированные, рандомизированные и смешанные факторные модели экспертного опроса. Для анализа полученных данных применяются различные идеи кусочно-линейной аппроксимации интенсиональных отношений оболочки. Чтобы такое детализированное знание не стало источником ошибок, предусматривается интерактивный режим отладки моделей опроса экспертов и обработки результатов, что обеспечивает валидность интегрированного логического вывода среды приобретения знаний.

ЛИТЕРАТУРА

1. Новожилова Т.Н. Интегрированный логический вывод интеллектуальной базы данных // Системи обробки інформації. – Харків: НАНУ, ПАНМ, ХВУ. – 1999. – Вип. 2(6). – С. 9 - 22.

2. Campbell D.T., Stanley J. C. Experimental and Quasi- Experimental Designs for Research. Chicago, Rand McNally, 12th ed., by the American Educational Research Association, 1976.

3. Новожилова Т.Н. Анализ структуры знаний // Всеукраїнська науково-методична конференція “Гуманізація і гуманітаризація вищої технічної освіти”. – Харків: ХТУРЕ. – 2000. – С. 250 - 252.

4. Новожилова Т.Н. Эвристическая сила структурности экспертного знания // 4 - й Международный молодёжный форум «Радиоэлектроника и молодёжь в XXI веке». – Харьков: ХТУРЭ. – 2000. – С. 313 - 314.

5. Новожилова Т.Н. Интеллектуальная база данных как средство приобретения знаний // Системи обробки інформації. – Харків: НАНУ, ПАНМ, ХВУ. – 2000. – Вип.1(7).- С. 39 - 46.