

## ОПРЕДЕЛЕНИЕ И ОЦЕНКА ХАРАКТЕРИСТИК СВОЙСТВ УНИВЕРСАЛЬНЫХ СУЩНОСТЕЙ ПРЕДМЕТНЫХ ОБЛАСТЕЙ

*Розглянуто проблеми створення моделей предметних областей та їхнього втілення в базах та сховищах даних. Наведені методи розв'язання проблем, що виникають при оцінюванні адекватності сутностей предметних областей та моделей цих сутностей. Зокрема, запропоновано математичне підґрунтя розв'язання проблеми визначення характеристик властивостей таких сутностей.*

*Рассмотрена проблема создания моделей предметных областей и их реализации в базах и хранилищах данных. Приведены методы решения проблем, которые возникают при оценке адекватности сущностей предметных областей и моделей этих сущностей. В частности, предложены математические основы решения проблемы определения характеристик свойств таких сущностей.*

*The problems of the subject domains models creation and their embodiment in databases and data warehouses are considered. Methods of the decision of problems which arise at estimation of adequacy of the subject domains entities and these models entities are shown. In particular, mathematical basics of the decision of a problem of definition of properties characteristics such entities are offered.*

Разработка систем информационной поддержки принятия решений в значительной степени зависит от степени соответствия баз и хранилищ, данных предметным областям (ПрО). Базы данных (БД) должны отражать процессы информационного обмена между предметными областями и внешней средой, а также между их объектами. Соответственно, необходима разработка математических методов проверки адекватности БД предметным областям как их информационным моделям, корректировки структур описания БД и инструментальных средств их наполнения выборочными данными [6].

Основными элементами ПрО являются объекты этих ПрО или проекции Универсальных Сущностей на эти ПрО [7]. Как известно из теории баз данных, объект ПрО, а следовательно, и вся универсальная сущность определяются своими свойствами, формализуемыми в соответствующих моделях в виде атрибутов.

Поэтому оценка адекватности модели ПрО должна начинаться с оценки адекватности моделЕЙ универсальной сущности и её свойств.

Для каждого свойства универсальной сущности, приблиЖЕННО описываемой

конечным множеством свойств, строится метод измерения степени его проявлений в одной из шкал: интервальной, балльной, порядковой или номинальной. Таким образом, для  $r$ -й универсальной сущности возможны следующие варианты принадлежности свойств к этим шкалам измерения в зависимости от степени их проявления:

а) все свойства принадлежат к интервальной, балльной или порядковой шкале измерения;

б) свойства распределяются на два подмножества, в одном из которых все свойства принадлежат к номинальной шкале, а во втором – к шкале, отличной от номинальной;

в) все наиболее информативные свойства, описывающие универсальную сущность с заданной степенью полноты и определенности, измеряются в номинальной шкале.

Для корректного анализа характеристик свойств сущностей предлагается осуществлять прогнозирование значений этих свойств, зависящее от их принадлежности к соответствующим шкалам.

Рассмотрим особенности прогнозирования  $(K + 1), \dots, (K + S)$  экземпляров  $r$ -й сущности на основе множества кортежей, размещенных в ее реляционном отношении как результат  $K$  статистических испытаний.

При этом предполагается, что данная совокупность информации является представительной выборкой, описывающей выделенную универсальную сущность с необходимыми полнотой и мерой определенности.

Свойства могут представляться детерминированной неизвестной функцией времени  $t$  и/или некоторым подмножеством других свойств, случайной величиной или случайным процессом. Кроме того, свойство может измеряться в балльной шкале, но закон представления относится к классу случайных процессов.

Понятие случайного процесса используется, в общем случае, для восстановления закона распределения свойства, измеренного в виде действительных чисел.

Множество всех элементарных случайных событий  $U(t)$  потенциально бесконечно и может изменяться во времени. В силу того, что количество испытаний всегда конечно, будем считать, что  $\bar{U}_K$  содержит конечное количество элементарных случайных событий, выбранных с помощью метода последовательного анализа.

На основе закона больших чисел имеет место соотношение

$$P_{K \rightarrow \infty} \left( (U - \bar{U}_K) \subseteq U_\varepsilon \right) \rightarrow 1, \quad (1)$$

где  $U$  – полное множество элементарных случайных событий,  $\bar{U}_K$  – множество элементарных случайных событий, полученное в результате  $K$  статистических испытаний,  $U_\varepsilon$  – множество всех элементарных событий, вероятность наступления которых в совокупности сколь угодно мала. Следовательно, они определяют случайное событие, вероятность наступления которого не превосходит сколь угодно малую величину  $\varepsilon > 0$ . Это означает, что  $U_K$  содержит все элементарные случайные события, которые определяют случайные события, вероятность наступления которых намного больше нуля.

Поскольку  $K$  не может выбираться сколь угодно большим и, более того, на практике оно всегда ограничено, возникает проблема

выбора меры неопределенности оценок различных характеристик свойств универсальных сущностей. Для получения таких оценок целесообразно использовать усиленный закон больших чисел. Пусть мы оцениваем степень неопределенности представления  $U$  с помощью  $U_K$ . Тогда, согласно усиленному закону больших чисел, будет иметь место **Утверждение 1.**

Для любых  $\varepsilon > 0$  и  $\eta > 0$  существует такое  $K_0$ , что для любого  $S$  и  $K \in [K_0, K_0 + S]$  выполняется

неравенство

$$P((U - \bar{U}_K) \subseteq U_\varepsilon) > 1 - \eta:$$

$$\forall \varepsilon > 0, \forall \eta > 0$$

$$\exists K_0 \left[ \forall S, K \in [K_0, K_0 + S] \Rightarrow \Rightarrow P((U - \bar{U}_K) \subseteq U_\varepsilon) > 1 - \eta. \quad (2) \right.$$

В виде вероятностных оценок с учетом того, что  $P(U) = 1$ , получаем соотношение

$$P((1 - P(\bar{U}_K)) < \varepsilon) > 1 - \eta. \quad (3)$$

Очевидно, что при заданных  $\varepsilon > 0$  и  $K$ , из приведенного неравенства может быть найдено значение  $\eta$ . При небольшом количестве испытаний  $K$  показатель  $(1 - \eta)$  утверждает, что степень неопределенности представления  $U$  с помощью  $U_K$  может быть значительной. Если ожидается, что  $\eta$  принимает небольшие значения, то из приведенного неравенства следует, что вероятность адекватного представления полного множества всех элементарных случайных событий с помощью множества элементарных случайных событий  $U_K$  будет на уровне  $(1 - \eta)$ . При значениях  $\eta \rightarrow 1$  гарантированная вероятность будет близкой к 0, что будет подтверждать невысокий уровень представительности данных, полученных в результате  $K$  статистических испытаний. Отсюда автоматически следует значительная степень неопределенности оценок любых характеристик рассмотренной совокупности свойств универсальной сущности  $E_r(t)$ .

К важным характеристикам выделенного

множества наиболее информативных свойств относятся оценки их математических ожиданий  $\{\bar{M}(A_{1r}(t)), \bar{M}(A_{2r}(t)), \dots, \bar{M}(A_{mr}(t))\}$  или сегментные оценки математических ожиданий  $\{\bar{M}^{sp}(A_{1r}(t)), \bar{M}^{sp}(A_{2r}(t)), \dots, \bar{M}^{sp}(A_{mr}(t))\}$ .

Статистические оценки математических ожиданий на сегментах фактора времени  $t$  используются тогда, когда оценки математического ожидания  $A_{jr}(t)$  на множестве сегментов (интервалов)  $\{[T_{j0}, T_{j1}]; [T_{j1}, T_{j2}], \dots, [T_{j_{f-1}}, T_{jf}]\}$  не являются постоянными и, следовательно, свойства  $A_{jr}(t)$  как случайные процессы не являются стационарными в узком, широком или квазистационарном смысле.

Такое же замечание характерно относительно любых других статистических характеристик, функций распределения вероятностей и, самое главное, прогнозных значений соответствующих свойств рассмотренной  $r$ -й универсальной сущности  $E_r(t)$ . Отсюда, в силу усиленного закона больших чисел, справедливо вытекает **Утверждение 2**.

Последовательность свойств (атрибутов)  $A_{1r}(t), A_{2r}(t), \dots, A_{mr}(t)$  как последовательность случайных процессов подчиняется усиленному закону больших чисел, если

$$\begin{aligned} & \forall \varepsilon > 0, \forall \eta > 0 \\ & \exists n_0 \left[ \forall s, n \in [n_0, n_0 + s] \Rightarrow \right. \\ & \Rightarrow P \left( \max_{n_0 \leq n \leq n_0 + s} \left| \frac{1}{n} \sum_{j=1}^n A_{jr}(t) - \right. \right. \\ & \left. \left. - \frac{1}{n} \sum_{j=1}^n M(A_{jr}(t)) \right| < \varepsilon \right) > 1 - \eta. \quad (4) \end{aligned}$$

Данное утверждение легко переносится на любой интервал оценивания или систему интервалов. Таким же образом может быть обобщена на случайные процессы теорема Гливленко [2]. Сущность этого обобщения: статистические оценки одномерных  $\bar{F}_K(A_{jr}(t_g))$  и всех многомерных функций распределения по усиленному закону боль-

ших чисел сходятся к  $F(A_{jr}(t_g))$  и, соответственно, к многомерным функциям распределения вероятности, где  $F(A_{jr}(t_g))$  — функция распределения вероятности свойства  $A_{jr}(t)$ ;  $\bar{F}_K(A_{jr}(t_g))$  — статистическая оценка этой функции на основе ее  $K$  выборочных значений в виде эмпирической функции распределения вероятностей.

При небольших значениях  $K$ , в силу усиленного закона больших чисел, выражение  $\bar{F}_K(A_{jr}(t_g))$  будет отличаться по абсолютной величине от выражения  $F(A_{jr}(t_g))$  не более, чем на  $\varepsilon > 0$  с вероятностью  $(\eta - 1)$ , где  $\eta > 0$ . Таким образом, при недостаточной представительности совокупности выборочных данных будет сохраняться неопределенность в представлениях о законе распределения каждого атрибута универсальной сущности.

В том случае, когда представительность выборочной совокупности не может быть улучшена, тогда сохраняемая мера неопределенности может быть оценена по выражению  $(\eta - 1)$ . Оценка величины  $(\eta - 1)$  может быть получена с помощью эмпирической функции распределения вероятностей  $\bar{F}_K(A_{jr}(t))$ .

Рассмотрим случай, когда все атрибуты принадлежат к номинальной шкале. Пусть для каждого атрибута на основе реляционного отношения определены его активные домены. Последовательность  $(ADom(A_{1r}(t)), ADom(A_{2r}(t)), \dots, ADom(A_{mr}(t)))$  представляет множество активных доменов всех наиболее информативных свойств  $r$ -й универсальной сущности, упорядоченных по убыванию их информационной ценности. В качестве меры информационной ценности использованы энтропийные меры Шеннона [8] и Кульбака [5].

Найдем декартово произведение приведенной последовательности активных доменов свойств (атрибутов)  $r$ -й универсальной сущности. Декартово произведение является множеством всех упорядоченных  $m$ -ок, т.е.

наборов (кортежей) из  $m$  элементов, составленных из значений активных доменов.

Построим классификацию множества всех упорядоченных  $m$ -ок для построения меры сходства экземпляров  $r$ -й универсальной сущности. При этом будем исходить из информационной ценности атрибутов. Пусть  $H(A_{1r}(t), A_{2r}(t), \dots, A_{mr}(t))$  — энтропия всего множества атрибутов  $H(A_{1r}(t), A_{2r}(t), \dots, A_{nr}(t))$ , где  $n < m$ , — энтропия множества наиболее важных атрибутов, выбранная так, что разность этих двух энтропий по абсолютной величине меньше заданного  $\alpha$ . Порог  $\alpha$  выбирается на основе прогностических возможностей первых  $n$  наиболее информативных свойств.

Две  $m$ -ки будем считать подобными, если первые  $n$  их значений совпадают, а другие могут принимать любые значения. При таком подходе всё декартово произведение будет разбито на  $\bar{n}$  классов. Пусть в декартовом произведении существует  $M$   $m$ -ок, а в каждом классе, соответственно,  $(M_1, M_2, \dots, M_{\bar{n}})$  экземпляров. На начальном этапе, т.е. при формировании декартового произведения активных доменов, приведенные величины определяются мощностями соответствующих декартовых произведений, причем  $M_1 = M_2 = \dots = M_{\bar{n}} = M_0$ :

$$\begin{aligned} M &= |ADom(A_{1r}(t)) \times ADom(A_{2r}(t)) \times \dots \\ &\quad \dots \times ADom(A_{mr}(t))| = \\ &= |ADom(A_{1r}(t))| \times |ADom(A_{2r}(t))| \times \dots \\ &\quad \dots \times |ADom(A_{mr}(t))|, \end{aligned} \quad (5)$$

относительно всех  $m$  атрибутов,

$$\bar{n} = |ADom(A_{1r}(t)) \times ADom(A_{2r}(t)) \times \dots \times ADom(A_{nr}(t))|, \quad (6)$$

относительно  $n$  наиболее информативных атрибутов,

$$\begin{aligned} M_0 &= |ADom(A_{(n+1)r}(t)) \times ADom(A_{(n+2)r}(t)) \times \\ &\quad \dots \times ADom(A_{mr}(t))|, \end{aligned} \quad (7)$$

относительно других  $m - n$  атрибутов.

Необходимо отметить, что декартово произведение не будет пересчитываться при следующих испытаниях, а новые экземпляры сущности для проведения оценок будут или

добавляться к существующим классам, или формировать новый. При этом, как любое  $M_i$  так и  $M$  в целом будут инкрементироваться на 1 или на  $L$  — количество новых экземпляров сущности, полученных в результате статистических испытаний за пределами  $K$  начальных испытаний.

Обозначим полученные классы множеством  $\{B_1, B_2, \dots, B_{\bar{n}}\}$ . Очевидно, что каждый класс определяет некоторое случайное событие. Оценки вероятностей наступления событий определяются на основе общеизвестного метода оценивания [4]. При этом получаем

$$\bar{P}(B_i) = \frac{M_i}{M}, \quad (8)$$

где  $i = \overline{1, \bar{n}}$ ,  $\sum_{i=1}^{\bar{n}} M_i = M$  или, на начальном этапе,  $M = M_0 \bar{n}$ .

Пусть действительные (точные) значения вероятностей будут соответственно  $P(B_1), P(B_2), \dots, P(B_{\bar{n}})$ ...

На основе законов больших чисел или усиленного закона больших чисел получаем соотношение

$$P \left( \left| P(B_i) - \frac{M_i}{M + L} \right| < \varepsilon \right) \rightarrow 1, \quad (9)$$

для всех  $i = \overline{1, \bar{n}}$

**Утверждение 3.** В случае, когда сложно обеспечить представительность выборочной совокупности, целесообразно использовать усиленный закон больших чисел, который, относительно этого случая, принимает вид:

$$\begin{aligned} &\forall \varepsilon > 0, \forall \eta > 0 \\ &\exists M_0 \left[ \forall S, M \in [M_0, M_0 + S] \Rightarrow \right. \\ &\Rightarrow P \left( \left| P(B_i) - \frac{M_i}{M + L} \right| < \varepsilon \right) > 1 - \eta_i, \end{aligned} \quad (10)$$

для всех  $i = \overline{1, \bar{n}}$ .

Допуская, что  $M$  задано, из этой системы неравенств можно найти  $\eta$  как максимальное значение среди  $\{\eta_1, \eta_2, \dots, \eta_{\bar{n}}\}$ .

Прогнозирование наступления после  $K$  испытаний одного из событий  $\{B_1, B_2, \dots,$

$B_{\bar{n}}$  } осуществляется с привлечением Байесовской процедуры оценивания условных вероятностей [1, 3, 4].

Предположим, что после  $K$ -го экземпляра реляционного отношения  $r$ -й универсальной сущности наступило событие  $B_g$ .

Вычислим условные вероятности:

$$P(B_j | B_g) = \frac{P(B_g | B_j)P(B_j)}{\sum_{i=1}^{\bar{n}} P(B_g | B_i)P(B_i)}, \quad (11)$$

для всех  $j = \overline{1, \bar{n}}$ ,  $j \neq g$ .

Потом найдем  $\max_{j \in [1, \bar{n}]} P(B_j | B_g)$  относительно  $g$ . Пусть максимум достигнут для  $j = d$ . Тогда будем считать, что следующим событием будет событие  $B_d$ . Согласно определению множества событий  $\{B_1, B_2, \dots, B_{\bar{n}}\}$  значения первых  $n$  атрибутов однозначно определяются прогнозируемым случайным событием  $B_d$  и лежащей в основе его определения  $m$ -ки. Первые  $n$  значений номинальных атрибутов определяют все экземпляры универсальной сущности, включенные в класс  $B_d$ .

Значения других  $m - n$  атрибутов определяются таким элементом декартового произведения активных доменов, который принадлежит  $B_d$ , и собственной классификацией, к прогнозированию классов которой применима та же байесовская процедура классификации, когда в подклассе с максимальной условной вероятностью выбирается произвольный набор значений  $m - n$  атрибутов, которые вместе с первыми  $n$  прогнозными значениями определяют прогнозируемый экземпляр.

В том случае, когда все выделенные наиболее существенные атрибуты  $r$ -й универсальной сущности принадлежат к номинальной шкале измерения, прогнозирование экземпляров универсальной сущности, которые будут наблюдаться в следующих  $M$  испытаниях, осуществляется по следующей схеме. Предположим, что на основе экземпляров универсальной сущности, полученных

в результате  $K$  испытаний, которые являются, в общем случае, статистически зависимыми, сформировано реляционное отношение с  $K$  кортежами степени  $m$ .

Будем считать, что атрибуты реляционного отношения упорядочены по убыванию их информационной ценности. Предположим, что первые  $n$  атрибутов являются более существенными, чем следующие  $m - n$  атрибутов. Полученное множество из  $K$  экземпляров универсальной сущности на основе первых  $n$  атрибутов делится на  $L$  классов  $\{B_1, \dots, B_L\}$  согласно описанной выше схеме.

Множество априорных оценок вероятностей классов  $\{\bar{P}(B_1), \dots, \bar{P}(B_L)\}$  определяется соотношениями

$$\bar{P}(B_1) = \frac{M_1}{K}, \dots, \bar{P}(B_L) = \frac{M_L}{K}, \quad (12)$$

где  $M_i$  — количество экземпляров в каждом классе, полученных в результате  $q$  итераций построения приближенной математической

модели ПрО, причем  $\sum_{i=1}^L M_i = K$ . Характерной особенностью каждого класса является то обстоятельство, что в каждом принадлежащем ему экземпляре значения первых  $n$  атрибутов совпадают, а остальные  $m - n$  атрибутов могут принимать значение из соответствующих активных доменов.

Положим, что выполнено  $M$  последовательных шагов прогнозирования принадлежности к тому или иному классу очередного экземпляра универсальной сущности. На каждом шаге прогнозирования будем использовать байесовскую процедуру, основанную на оценивании условных вероятностей и выборе в качестве прогнозируемого того класса, для которого условная вероятность принимает наибольшее значение.

На каждом шаге прогнозирования возможны следующие случаи:

- а) прогноз является точным;
- б) прогноз является ошибочным, но структура классов не изменяется, т.е. не появляются новые классы;
- в) прогноз является ошибочным из-за появления нового класса, который связан с

появлением в очередном испытании экземпляра универсальной сущности, первые  $q$  атрибутов которого не совпадают ни с одним из существующих классов.

В каждом случае на каждой итерации прогнозирования будем получать две оценки вероятностей классов  $\{B_1, \dots, B_L\}$ . Одна осуществляется на основании прогнозных значений классов, а вторая – на основании реальной принадлежности очередного экземпляра к одному из существующих классов. Если наблюдается случай (в), количество классов увеличивается на единицу.

Положим, что при  $M$  шагах прогнозирования с целью оценки адекватности приближенной математической модели ПрО для заданной универсальной сущности получены два множества:

$$\left\{ \begin{aligned} \bar{P}_F(B_1) &= \frac{M_1^F}{K+M}, \dots \\ \dots, \bar{P}_F(B_{L+L_1}) &= \frac{M_{L+L_1}^F}{K+M} \end{aligned} \right\}; \quad (13)$$

$$\left\{ \begin{aligned} \bar{P}(B_1) &= \frac{M_1}{K+M}, \dots \\ \dots, \bar{P}(B_{L+L_1}) &= \frac{M_{L+L_1}}{K+M} \end{aligned} \right\}. \quad (14)$$

Первая – это прогнозные оценки вероятностей классов, а вторая — классические оценки вероятностей классов на основе статистического подхода.

В случае (а) оба множества корректируются одинаково на каждом этапе прогнозирования. В случае (б) корректируются вероятности различных классов в первом и втором множествах. В случае (в) с первым появлением нового класса  $B_{L+S}$ , где  $1 \leq S \leq L_1$ , его прогнозная вероятность  $\bar{P}_F(B_{L+S}) = 0$ , а классическая —  $\bar{P}(B_{L+S}) = \frac{1}{K+S}$ . При дальнейшем прогнозировании этот класс уже используется или для случая (а), или для случая (б).

Очевидно, что если для каждого  $\varepsilon > 0$  имеет место соотношение

$$P\left(\left(\sum_{i=1}^{L+L_1} |\bar{P}_F(B_i) - P(B_i)|\right) < \varepsilon\right) \rightarrow 1, \quad (15)$$

при  $M \rightarrow \infty$ , то, в силу усиленного закона больших чисел, можно утверждать, что  $\bar{P}(B_i)$  и  $\bar{P}_F(B_i)$  по вероятности сходятся к  $P(B_i)$  для всех  $i = \overline{1, L+L_1}$ , и модель является адекватной для данной универсальной сущности.

Таким образом, предложенный подход и сформулированные для этого утверждения позволяют определить некоторые, в частности, вероятностные характеристики свойств сущностей ПрО и оценить степень их информативности. На основании полученных характеристик осуществляется итеративное прогнозирование значений этих свойств, эффективность которого проверяется сравнением значений атрибутов спрогнозированного экземпляра универсальной сущности со значениями атрибутов экземпляра, который будет наблюдаться при  $(K+1)$ -м испытании. Сравнение осуществляется по всем атрибутам  $r$ -й универсальной сущности с учетом их информационной ценности. Это, в свою очередь, даёт возможность оценить адекватность информационных моделей сущностей, что является предпосылкой оценки адекватности модели самой ПрО и, соответственно, повышения полноты, корректности и снижения степени неопределённости системы БД, в виде которой реализуется такая модель.

#### Список использованной литературы

1. Гмурман В.Е. Теория вероятностей и математическая статистика / В.Е. Гмурман. – М.: Высш. шк., 1999. – 432 с.
2. Гнеденко Б.В. Курс теории вероятностей / Б.В. Гнеденко. - 7-е изд., испр. – М: УРСС. - 2009. - 318 с.
3. Дубров А.М. Многомерные статистические методы / Дубров А.М., Мхитарян В.С., Трошин Л.И. – М.: Финансы и статистика, 2003. – 352 с.
4. Кобзарь А.И. Прикладная математическая статистика / А.И. Кобзарь. – М.: Физматлит, 2006. – 814 с.

5. Кульбак Р. Теория информации и статистика / Р. Кульбак: Пер. с англ. – М.: Наука, 1967. – 408 с.

6. Малахов Е.В. Оценка степени адекватности баз данных как информационных моделей предметных областей [текст] / Е.В. Малахов // Тр. Одес. политехн. ун-та. – 2004. – Вып. 1(21). – С. 82 – 86.

7. Малахов Е.В. Манипулирование мета-моделями предметных областей [текст] / Е.В. Малахов // Восточно-европейский журнал передовых технологий. – Харьков: 2007. – Вып. 5/3(29). – С. 6 – 10.

8. Шеннон К. Работы по теории информации и кибернетике / К. Шеннон. – М.: Изд-во иностранной литературы, 1963. – 832 с.

Получено 23.07.2010



Малахов Евгений Валериевич, канд. техн. наук, доцент, зав. каф. информационных систем в менеджменте Одес. нац. политехн. ун-та  
e-mail: mev@oru.ua



Востров Георгий Николаевич, канд. техн. наук, доцент, зав. каф. прикладной математики и информационных технологий в бизнесе Одес. нац. политехн. ун-та  
e-mail: mev@oru.ua