

О количественной оценке неопределенности вывода решений с помощью гиперкубов многомерных данных

Аннотация. В работе исследуются задачи оценки неопределенности вывода решений с помощью гиперкубовых структур в условиях реальных возмущений информационного состава баз и хранилищ многомерных данных. Вводится понятие меры неопределенности данных. Получены критерии уменьшения (увеличения) меры неопределенности вывода решений при изменении основных параметров гиперкубов многомерных данных.

Ключевые слова: гиперкуб, многомерные данные гиперкуба, мера неопределенности, энтропия, мера энтропии, набор сечений.

Введение

В статье, являющейся продолжением цикла работ [1-4] изучается проблема оценки неопределенности вывода решений с помощью многомерной модели данных – гиперкуба, в условиях разреженности данных гиперкуба и неограниченности числа размерностей и градаций на шкалах размерностей. Неопределенность в сложной (большой) системе – это ситуация, когда полностью или частично отсутствует информация о возможных состояниях системы. Иначе говоря, когда в системе возможны те или иные непредсказуемые события, вероятностные характеристики которых неизвестны. Неопределенность является неизбежным спутником сложных систем; чем сложнее система, тем большее значение приобретает фактор неопределенности в ее поведении. В статье рассматриваются аналитические OLAP-системы, основное назначение которых связано с агрегированием числовых данных в иерархиях размерностей, поэтому обрабатываются структурированные данные. Нечеткие и вербальные данные, характерные для экспертных систем принятия решений [6-7], в статье не исследуются. Структуризация и приведение данных к числовому виду осуществляется в хранилищах с помощью специальных систем ETL, которые выявляют ошибочные данные и преобразуют их

форматы. То есть очистка и структуризация данных происходит на более низком уровне по сравнению с уровнем OLAP. Но структурирование данных не означает устранения всех возможных видов неопределенности. В таких системах неопределенность логического вывода возникает не из-за дефекта данных, а из-за закономерного отсутствия у аналитиков знаний о структурных связях между отдельными значениями куба данных при больших их решетках, что и приводит к необходимости рассматривать разреженные кубы. Аналитик, работая с большими разреженными кубами многомерных данных, часто попадает в ситуации, когда его запрос приводит к пустой ячейке куба. Это происходит тогда, когда запрос аналитика нарушает логику связей, существующую между данными.

В статье предложен подход, позволяющий устранять подобные ошибки за счет предварительной оценки различных покрытий гиперкуба, свободных от разреженных ячеек. В идеале можно предварительно определить все покрытия гиперкуба, не приводящие к пустым ячейкам. Тогда компьютерная система будет отвергать ошибочные запросы аналитика до начала вычисления сечений куба. Задача устранения такой неопределенности может быть решена с помощью теории энтропии. В этой теории мера неопределенности называется энтропией.

1. Основные свойства меры неопределенности вывода решений

Введем определение данного понятия [5, 8, 9]. Вместе со шкалами и их значениями зададим некоторое многомерное пространство (гиперкуб) HK_n исходов. Размерность HK_n вычисляется по формуле [5]:

$$\dim HK_n = \prod_{i=1}^n k_i,$$

где k_i – число значений на шкале i -ой размерности (критерия);

n – общее число размерностей гиперкуба данных HK_n ;

$\prod_{i=1}^n k_i$ – произведение величин $k_1, \dots, k_n$;

$\dim HK_n$ – общая размерность гиперкуба HK_n .

Определение 1. Логарифм размерности гиперкуба исходов HK_n называется энтропией критериального описания процесса генерации множества наборов сечений HK_n :

$$w_0 = \log_2 \prod_{i=1}^n k_i = \sum_{i=1}^n \log_2 k_i. \quad (1)$$

Определение 2. Энтропией конечного состояния гиперкуба HK_n называется величина

$$w_1 = \log_2 r,$$

где r – число различных наборов сечений $V_1, \dots, V_r$ гиперкуба многомерных данных, покрывающее HK_n .

Определение 3. Информацией о наборе сечений $V_1, \dots, V_r$ гиперкуба HK_n называется величина

$$w = w_0 - w_1 = \log_2 \prod_{i=1}^n k_i - \log_2 r. \quad (2)$$

Определение 4. Мерой неопределенности α задачи генерации описания наборов сечений $V_1, \dots, V_r$ гиперкуба многомерных данных HK_n называется отношение полезной информации w к энтропии исходного критериального описания HK_n :

$$\alpha \equiv w \cdot (w_0)^{-1} = 1 - \frac{\log_2 r}{\log_2 \prod_{i=1}^n k_i} = 1 - \frac{\log_2 r}{\sum_{i=1}^n \log_2 k_i}. \quad (3)$$

где величины w , w_0 определяются формулами (1), (2) соответственно. В работах [1, 6, 9] установлено следующее утверждение о свойствах меры неопределенности α .

Теорема 1. Величина α обладает следующими свойствами:

- 1) $0 \leq \alpha \leq 1$ при любом натуральном числе r из отрезка $[1, r_n]$, где $r_n = \prod_{i=1}^n k_i$;
- 2) $\alpha = 0$ тогда и только тогда, когда $r = r_n$;
- 3) $\alpha = 1$ тогда и только тогда, когда $r = 1$;

4) Функция $\alpha = \alpha(k, n, r)$ возрастает с увеличением аргументов n , k и убывает с ростом аргумента r .

Теорема 2. Пусть число оценок по шкалам гиперкуба одинаково и равно k . Число наборов сечений равно r . Предположим, что величина k увеличивается на $s \geq 1$, а величина r на $m \geq 1$. Тогда для того, чтобы мера неопределенности $\alpha(k_1, r_1) \equiv \alpha(k + s, r + m)$ была меньше меры неопределенности $\alpha(k, r)$ необходимо и достаточно выполнения неравенства:

$$\alpha(k_1, r_1) < \alpha(k, r) \leftrightarrow \log_2 r \cdot$$

$$\log_2(k + s) < \log_2(r + m) \cdot \log_2 k.$$

Доказательство теоремы 2. Для величин $\alpha(k_1, r_1)$, $\alpha(k, r)$ справедлива следующая цепочка соотношений:

$$\begin{aligned} \alpha(k_1, r_1) - \alpha(k, r) &= \frac{\log_2 r}{n \log_2 k} - \frac{\log_2(r + m)}{n \log_2(k + s)} = \\ &= \frac{1}{n} \cdot \left(\frac{\log_2 r \cdot \log_2(k + s) - \log_2 k \cdot \log_2(r + m)}{\log_2 k \cdot \log_2(k + s)} \right). \quad (4) \end{aligned}$$

Используя (4), находим

$$\alpha(k_1, r_1) < \alpha(k, r) \leftrightarrow \log_2 r \cdot$$

$$\log_2(k + s) < \log_2 k \cdot \log_2(r + m),$$

что и требовалось доказать.

Отметим, что теорема 2 дает критерий увеличения величины $\alpha(k, n, r)$ при увеличении параметров k и r , при этом $n = \text{const}$.

Теорема 3. Допустим теперь, что число критериев n и число наборов сечений r увеличивается соответственно на величины $q \geq 1$, $p \geq 1$: $n_1 = n + q$, $r_1 = r + p$. Тогда для того, чтобы мера неопределенности $\alpha(n_1, r_1) \equiv \alpha(n + q, r + p)$ была меньше меры неопределенности $\alpha(n, r)$ необходимо и достаточно, чтобы выполнялись неравенства:

$$\begin{aligned} \alpha(n + q, r + p) < \alpha(n, r) &\leftrightarrow r^{n+q} < (r + p)^n \leftrightarrow \\ &r^q < \left(1 + \frac{p}{r}\right)^n. \end{aligned}$$

Доказательство теоремы 3. Для разности $\alpha(n_1, r_1) - \alpha(n, r)$ будем иметь

$$\begin{aligned} \alpha(n_1, r_1) - \alpha(n, r) &= \\ &= \frac{1}{\log_2 k} \cdot \frac{(\log_2 r)(n+q) - n \cdot \log_2(r+p)}{n \cdot (n+q)}. \end{aligned}$$

Из последней формулы окончательно получаем, что

$$\begin{aligned} \alpha(n_1, r_1) < \alpha(n, r) &\leftrightarrow r^{n+q} < (r + p)^n \leftrightarrow r^q \\ &< \left(1 + \frac{p}{r}\right)^n. \end{aligned}$$

Теорема 3 полностью доказана. Она дает критерий увеличения величины $\alpha(k, n, r)$ при увеличении параметров n , r и $k = \text{const}$.

Теорема 4. Допустим, что задано конечное покрытие гиперкуба многомерных данных HK_n $V_1, \dots, V_r$. Предположим, что для дальнейшего их исследования мы добавляем еще один критерий с номером $n+1$. Число оценок на шкале его измерений равно k_{n+1} . Пусть число наборов сечений покрытия увеличилось на единицу:

$$V_1, \dots, V_r \Rightarrow V'_1, \dots, V'_r, V'_{r+1}.$$

Тогда для того, чтобы мера неопределенности α_{r+1} покрытия $V'_1, \dots, V'_{r+1}$, была меньше величины α_r , необходимо и достаточно, чтобы числа $k_1, \dots, k_n, k_{n+1}$ удовлетворяли соотношению:

$$\alpha_{r+1} < \alpha_r \Leftrightarrow T_n(\log_2(r+1) - \log_2 r) > \log_2 r \cdot \log_2 k_{n+1},$$

$$\text{где } T_n = \sum_{i=1}^n \log_2 k_i$$

Доказательство теоремы 4. Для величин α_r, α_{r+1} имеем:

$$\alpha_r = 1 - \frac{\log_2 r}{\log_2 \prod_{i=1}^n k_i} = 1 - \frac{\log_2 r}{\sum_{i=1}^n \log_2 k_i},$$

$$\begin{aligned} \alpha_{r+1} &= 1 - \frac{\log_2(r+1)}{\log_2 \prod_{i=1}^{n+1} k_i} = \\ &= 1 - \frac{\log_2(r+1)}{\sum_{i=1}^n \log_2 k_i + \log_2 k_{n+1}}. \end{aligned} \quad (5)$$

Используя (5), находим, что

$$\begin{aligned} \alpha_{r+1} - \alpha_r &= \\ &= \frac{-T_n \cdot \log_2(r+1) + \log_2 r \cdot \log_2 k_{n+1} + T_n \cdot \log_2 r}{T_n \cdot (T_n + \log_2 k_{n+1})}, \end{aligned} \quad (6)$$

где $T_n = \sum_{i=1}^n \log_2 k_i$.

Принимая во внимание (6), окончательно получаем, что $\alpha_{r+1} - \alpha_r < 0 \Leftrightarrow T_n(\log_2(r+1) - \log_2 r) > \log_2 r \cdot \log_2 k_{n+1}$, что и требовалось доказать.

Теорема 4 полностью доказана. Она устанавливает критерий уменьшения величины $\alpha(k, n, r)$ при следующих условиях $r \Rightarrow r+1$, $n \Rightarrow n+1$, $k_1, \dots, k_n \Rightarrow k_1, \dots, k_{n+1}$, где k_{n+1} – некоторое натуральное число.

Отметим, что теоремы 2-4 дают критерии уменьшения (увеличения) меры неопределенности $\alpha(k, n, r)$ покрытия $V_1, \dots, V_r$ при изменениях основных параметров n, k, r . Ниже будут рассмотрены примеры применения этих критериев исследования динамики изменения величины $\alpha(k, n, r)$ для некоторых частных случаев покрытий гиперкубов многомерных данных HK_n .

2. Примеры применения теорем 1-3 при исследовании меры неопределенности покрытий гиперкубов многомерных данных

В Табл. 1 и Табл. 2 число столбцов равно 3, число строк – 8. Значение элемента NULL предполагается одинаковым для выбранного столбца. Энтропия таблиц вычисляется по одной и той же формуле и имеет величину, равную 6:

$$w_0 = \log_2 8 \cdot 2 \cdot 4 = \log_2 64 = 6,$$

где 8-число строк (месяцев); 2-число различных элементов во втором столбце; 4-число различных элементов в третьем столбце.

Мера энтропии для таблиц вычисляется по формуле (определение 4):

$$\alpha = 1 - \frac{\log_2 r}{w_0} = 1 - \frac{\log_2 r}{6},$$

где r -число наборов сечений (строк), выбранных ЛПР (или покупателем продукции):

$$r(I)=2, \quad r(II)=21.$$

1) Список наборов сечений к Табл.1:

$$V_1 = \{\text{число сечений с точными данными}\} = 5,$$

$$V_2 = \left\{ \begin{array}{l} \text{число сечений} \\ \text{с неопределёнными данными} \end{array} \right\} = 3.$$

2) Энтропия конечного состояния:

$$w_1 = \log_2 2 = 1.$$

2) Информация о наборе сечений:

$$w = w_0 - w_1 = 6 - 1 = 5.$$

3) Мера неопределенности:

$$\alpha = 1 - \frac{w_1}{w_0} = 1 - \frac{\log_2 2}{6} = 1 - \frac{1}{6} = \frac{5}{6} \approx 0,83.$$

1) Список наборов сечений к Табл.2:

$$V_1 = \{\text{число сечений с точными данными}\} = 5,$$

$$V_2 = \{\text{сечения с максимальной ценой}\} = 1,$$

$$V_3 = \{\text{число сечений с минимальной ценой}\} = 1,$$

$$V_4 = \{\text{число сечений со средней ценой}\} = 3,$$

$$V_5 = \left\{ \begin{array}{l} \text{число сечений с неопределёнными} \\ \text{данными} \end{array} \right\} = 3,$$

$$V_6 = \left\{ \begin{array}{l} \text{число сечений с определёнными} \\ \text{данными в столбце 2} \end{array} \right\} = 6,$$

$$V_7 = \left\{ \begin{array}{l} \text{число сечений с определёнными} \\ \text{данными в столбце 3} \end{array} \right\} = 5,$$

$$V_8 = \left\{ \begin{array}{l} \text{число сечений с неопределёнными} \\ \text{данными в столбцах 2 и 3} \end{array} \right\} = 2,$$

$$\begin{aligned}
 V_9 &= \left\{ \begin{array}{l} \text{число сечений с одинаковыми} \\ \text{точными данными в столбце 2} \end{array} \right\} = 6, \\
 V_{10} &= \left\{ \begin{array}{l} \text{число сечений с неопределенными} \\ \text{данными в столбце 2} \end{array} \right\} = 2, \\
 V_{11} &= \left\{ \begin{array}{l} \text{число сечений с одинаковыми} \\ \text{точными данными в столбцах} \\ \text{2 и 3 одновременно} \end{array} \right\} = 3, \\
 V_{12} &= \left\{ \begin{array}{l} \text{число сечений с неопределенными} \\ \text{данными в столбце 3} \end{array} \right\} = 3, \\
 V_{13} &= \left\{ \begin{array}{l} \text{число сечений с минимальной ценой} \\ \text{в 1}^{\text{ом}} \text{ полугодии} \end{array} \right\} = 1, \\
 V_{14} &= \left\{ \begin{array}{l} \text{число сечений со средней ценой} \\ \text{в 1}^{\text{ом}} \text{ полугодии} \end{array} \right\} = 3, \\
 V_{15} &= \left\{ \begin{array}{l} \text{число сечений с максимальной ценой} \\ \text{в 1}^{\text{ом}} \text{ полугодии} \end{array} \right\} = 1, \\
 V_{16} &= \left\{ \begin{array}{l} \text{число сечений с неопределенными} \\ \text{данными в 1}^{\text{ом}} \text{ полугодии} \end{array} \right\} = 1, \\
 V_{17} &= \left\{ \begin{array}{l} \text{число сечений, имеющих как} \\ \text{определенные, так и неопределенные} \\ \text{данные} \end{array} \right\} = 1, \\
 V_{18} &= \left\{ \begin{array}{l} \text{месяц 1}^{\text{го}} \text{ полугодия} \\ \text{с минимальной ценой продукта} \end{array} \right\} = \text{январь}, \\
 V_{19} &= \left\{ \begin{array}{l} \text{месяц 1}^{\text{го}} \text{ полугодия с максимальной} \\ \text{ценой продукта} \end{array} \right\} = \text{май}, \\
 V_{20} &= \left\{ \begin{array}{l} \text{месяцы 1}^{\text{го}} \text{ полугодия} \\ \text{со средней ценой продукта} \end{array} \right\} = \text{февраль -} \\
 &\quad \text{апрель}, \\
 V_{21} &= \left\{ \begin{array}{l} \text{месяц года с определенными} \\ \text{и неопределенными данными} \end{array} \right\} = \text{июнь}.
 \end{aligned}$$

2) Энтропия конечного состояния:

$$w_1 = \log_2 21 = \log_2 3 + \log_2 7.$$

3) Информация о наборе сечений:

$$w = w_0 - w_1 = 6 - \log_2 3 - \log_2 7 \approx 1,62.$$

4) Мера неопределенности:

$$\begin{aligned}
 \alpha &= 1 - \frac{w_1}{w_0} = 1 - \frac{\log_2 3 + \log_2 7}{6} = \\
 &= \frac{6 - 1,58 - 2,80}{6} = 1,62 : 6 = 0,27.
 \end{aligned}$$

Отметим также, что примеры вычисления меры α показывают уменьшение её величины с ростом количества наборов сечений r ($r = 2$ для Табл. 1, $r = 21$ для Табл. 2), что полностью согласуется с утверждением 4 теоремы 1.

Рассмотрим два набора сечений гиперкубов данных $X(r)$, $X(r')$ со следующими основными параметрами:

$$X(r): r = 4; n = 10; k_1 = k_2 = \dots = k_{10} = 5; \quad (7)$$

$$\begin{aligned}
 X(r'): r' = r + 1 = 5; n' = n + 1 = 11; \\
 k'_1 = k'_2 = \dots = k'_{10} = 5; k'_{11} = 12. \quad (8)
 \end{aligned}$$

Табл.1

Месяц	Количество (кг)	Стоимость (руб.)
январь	100	800
февраль	100	1000
март	100	1000
апрель	100	1000
май	100	1200
июнь	NULL	NULL
июль	NULL	NULL
август	NULL	NULL

Табл.2

Месяц	Количество (кг)	Стоимость (руб.)
январь	100	800
февраль	100	1000
март	100	1000
апрель	100	1000
май	100	1200
июнь	100	NULL
июль	NULL	NULL
август	NULL	NULL

Заметим, что набор сечений $X(r')$ получается из $X(r)$ добавлением одного нового критерия k'_{11} и одного нового сечения. Оценим меры неопределенности α_r , $\alpha_{r'} = \alpha_{r+1}$ наборов сечений $X(r)$, $X(r')$. Используя соотношения (1)- (3) и результат теоремы 3, находим, что при $r = 4$:

$$\alpha_{r+1} < \alpha_r \Leftrightarrow 10 \log_2 5 (\log_2 5 - 2) > 2 \log_2 k'_{11}. \quad (9)$$

Принимая во внимание (8), получаем, что

$$\alpha_5 < \alpha_4 \Leftrightarrow 5x(x - 2) > 2 + \log_2 3, \quad (10)$$

где $x \equiv \log_2 5$.

Оценим величины x , $y \equiv 2 + \log_2 3$. Для величины $x = \log_2 5$ имеем:

$$2,312 < x < 2,4. \quad (11)$$

Действительно, так как $[10, 11]$ $\lg 5 = \log_{10} 5 > 0,6985$; $\lg 2 = \log_{10} 2 < 0,302$, то

$$\log_2 5 = \frac{\lg 5}{\lg 2} > \frac{0,6985}{0,302} > 2,312. \quad (12)$$

Далее находим, что $2^{12} = 4096 > 3125 = 5^5$, следовательно, $2^{12/5} = 2^{2,4} > 5$, откуда получаем, что

$$2,4 > \log_2 5. \quad (13)$$

Используя (12), (13), окончательно получаем, что $2,312 < \log_2 5 < 2,4$. Это и доказывает

(11). Для величины $y = 2 + \log_2 3 = \log_2 12$ имеет место следующее неравенство:

$$y < 3,6. \quad (14)$$

В самом деле, для доказательства неравенства (14) нам достаточно установить, что $\log_2 3 < 1,6$. Имеем: $2^8 = 256 > 243 = 3^5$, откуда находим, что $\log_2 3 < \frac{8}{5} = 1,6$.

Это и требовалось доказать.

Принимая во внимание (11), (14), получаем, что для выполнения соотношения $\alpha_5 < \alpha_4$ достаточно, чтобы величина $x = \log_2 5$ удовлетворяла неравенству:

$$5x(x-2) > 3,6. \quad (15)$$

Имеем: $5x(x-2) = 5\log_2 5 \cdot (\log_2 5 - 2) \geq 11,56 \cdot 0,312 = 3,60672 > 3,6$, что и доказывает справедливость соотношения (15). Используя (15), окончательно получаем, что $\alpha_5 < \alpha_4$ при всех k'_{11} таких, что $1 \leq k'_{11} \leq 12$.

Таким образом, нами установлен следующий достаточный признак уменьшения меры неопределенности наборов сечений $X(r)$, $X(r')$, где $r = 4$, $r' = 5$.

Утверждение 1. Пусть наборы сечений гиперкубов $X(r)$, $X(r')$ многомерных данных удовлетворяют условиям (7), (8) соответственно. Тогда для того, чтобы для таких сечений выполнялось неравенство $\alpha_{r'} < \alpha_r$ достаточно, чтобы величина $k'_{11} = k'_{n+1}$ удовлетворяла соотношениям: $1 \leq k'_{11} \leq 12$.

Рассмотрим три следующих набора сечений гиперкубов многомерных данных.

- 1) $X(r)$: $r = 4$, $n = 10$; $k_1 = k_2 = \dots = k_{10} = 5$;
- 2) $X(r'')$: $r'' = r + 12 = 16$; $n'' = n = 10$; $k'_1 = \dots = k'_{10} = 25$;
- 3) $X(r''')$: $r''' = r + 12 = 16$; $n''' = n = 10$; $k'''_1 = \dots = k'''_{10} = 26$.

Напомним (см. теорему 2), что мера неопределенности $\alpha(k_1, r_1)$ меньше величины меры неопределенности $\alpha(k, r)$ тогда и только тогда, когда выполняется неравенство:

$$\log_2 r \cdot \log_2(k + s) < \log_2 k \cdot \log_2(r + m). \quad (16)$$

В нашем случае справедливо следующее утверждение.

Утверждение 2:

2.1. Пусть для наборов сечений гиперкубов выполнены условия (1) и (2), тогда

$$\alpha(X(r)) = \alpha(X(r'')),$$

где $\alpha(X(r))$, $\alpha(X(r''))$ – меры неопределенности, соответственно гиперкубов $X(r)$, $X(r'')$.

2.2. Наборы $X(r)$, $X(r''')$, удовлетворяющие условиям (1) и (3), связаны соотношением

$$\alpha(X(r)) < \alpha(X(r''')),$$

где $\alpha(X(r))$, $\alpha(X(r'''))$ – меры неопределенности, соответственно гиперкубов $X(r)$, $X(r''')$.

Доказательство:

2.1. Для гиперкубов $X(r)$, $X(r'')$ имеем:

$$k+s=25;$$

$$\log_2 r \cdot \log_2(k + s) = \log_2 4 \cdot \log_2 25 = \log_2 625; \quad (17)$$

$$\log_2 k \cdot \log_2(r + m) \log_2 5 \cdot \log_2 16 = 4\log_2 5 = \log_2 625.$$

Утверждение 2.1. доказано.

2.2. Для $X(r)$, $X(r''')$, учитывая условия (1), (3), находим

$$\log_2 r \cdot \log_2(k + s) = \log_2 676 > \log_2 625. \quad (18)$$

Принимая во внимание (18), получаем, что в случае выполнения условий (1), (3)

$$\alpha(X(r''')) > \alpha(X(r)),$$

что и требовалось доказать.

Утверждение 2 полностью доказано.

Утверждение 3. Для наборов сечений гиперкубов $X(r)$, $X(r'')$, $X(r''')$, удовлетворяющих условиям (1) - (3), выполнено:

$$\text{при } 5 \leq k''_i \leq 25 \quad (i=1, \dots, n)$$

$$\alpha(X(r)) \geq \alpha(X(r''));$$

$$\text{при } k'''_i \geq 26 \quad (i=1, \dots, n)$$

$$\alpha(X(r)) < \alpha(X(r''')).$$

Отметим, что все утверждения дают критерии уменьшения, равенства, увеличения меры неопределенности гиперкуба $X(r)$ при варьировании его основных параметров r, n, k , удовлетворяющих условиям (1)-(3).

Предложенный подход особенно важен при работе OLAP-кубов в условиях сверхбольших данных (BigData) как, например, при решении задач мобильными операторами, имеющими миллионы абонентов и терабайтные объемы биллинговых данных, или при анализе крупных розничных торговых сетей, оперирующих миллионами покупателей и огромной номенклатурой продукции. Часто имеющуюся в базах или хранилищах данных статистику необходимо представить в структурированном виде по сечениям гиперкуба и получить знания о логических связях многомерных данных. После этого можно применить методы классификации структурированных данных с помощью экспертов [12] либо с использованием компьютерных аналитических процедур.

Заклyчение

В статье вводятся характеризующие неопределенность гиперкубов многомерных данных понятия энтропии и меры энтропии. Исследуется влияние основных параметров: общего числа размерностей n , числа значений на шкале i -ой размерности k_i ($i=1, \dots, n$), числа наборов сечений g , покрывающих гиперкуб, при их варьировании на поведение меры энтропии. Получены достаточные условия уменьшения, равенства, увеличения меры энтропии гиперкубовых структур многомерных данных при возмущении (или изменении) перечисленных параметров гиперкуба: размерности и значений на шкалах размерности ведут к увеличению меры энтропии (неопределенности) гиперкуба данных, а увеличение наборов сечений с заранее определенным выводом уменьшает меру энтропии, уменьшая тем самым неопределенность логического вывода в условиях разреженных кубов.

Литература

- Макаров И.М., Рахманкулов В.З., Ахрем А.А., Ровкин И.О. Построение СППР на основе OLAP-технологии // Информационные технологии и вычислительные системы. 2005. №1. С. 19-30.
 - Макаров И.М., Рахманкулов В.З., Ахрем А.А., Ровкин И.О. Исследование свойств гиперкубовых структур в OLAP-системах // Информационные технологии и вычислительные системы. 2005. №2. С. 4-9.
 - Макаров И.М., Ахрем А.А., Рахманкулов В.З., Ахрем А.А., Вашеvник Т.Л., Филюков Р.Ю. Энтропийные методы анализа информации // Труды ИСА РАН. 2011. Вып. 1. Т. 1. С. 36-40.
 - Ахрем А.А., Макаров И.М., Рахманкулов В.З. Математическая теория виртуализации процессов проектирования и трансфера технологий. М.: Физматлит. 2013. 316 с.
 - Gray Robert M. Entropy and Information Theory, 2nd edition – Springer, 2011. – 409 p.
 - Ларичев О.И. Вербальный анализ решений. М.: Наука. 2006. 181 с.
 - Вагин В.Н., Головина Е.Ю., Загорянская А.А., Фомина М.В. Достоверный и правдоподобный вывод в интеллектуальных системах. 2-е изд. М.: Физматлит. 2008. 712 с.
 - Осипов Г.С. Методы искусственного интеллекта. М.: Физматлит. 2011. 296 с.
 - Попков Ю.С. Математическая демоэкономика. макро-системный подход. М.: ЛЕНАНД. 2013. 560 с.
 - Бронштейн И.Н., Семендяев К.А. Справочник по математике для инженеров и учащихся втузов. М.: Наука. 1981. 387 с.
 - Выгодский М.Я. Справочник по элементарной математике. М.: АСТ: Астрель. 2014. 509 с.
 - Петровский А.Б., Ройзензон Г.В. Интерактивная процедура снижения размерности признакового пространства в задачах многокритериальной классификации. Поддержка принятия решений/Под ред. А.Б. Петровского Труды ИСА РАН. М.: Изд. ЛКИ/URSS. 2008. Т. 35. С. 43-53.
- Рахманкулов Виль Закирович.** Заведующий лабораторией ИСА ФИЦ ИУ РАН. Окончил Московский авиационный институт в 1960 году. Доктор технических наук, профессор. Количество печатных работ: более 175. Область научных интересов: системный анализ, виртуальное моделирование, интеллектуальный анализ данных, системы бизнес-аналитики, теория управления, компьютерная автоматизация производства. E-mail: vilrakh@mail.ru
- Ахрем Андрей Афанасьевич.** Старший научный сотрудник ИСА ФИЦ ИУ РАН. Окончил Московский Государственный Университет в 1977 году. Кандидат физико-математических наук. Количество печатных работ: более 110. Область научных интересов: математическая теория систем, математическое и виртуальное моделирование сложных технических систем. E-mail: vilrakh@mail.ru
- Южанин Кирилл Викторович.** Инженер-исследователь ИСА ФИЦ ИУ РАН. Окончил Московский инженерно-физический институт в 2012 году. Количество печатных работ: 1. Область научных интересов: теория распознавания образов, системный анализ. E-mail: vilrakh@mail.ru
- Вашевник Татьяна Леонидовна.** Научный сотрудник ИСА ФИЦ ИУ РАН. Окончила Московский экономико-статистический институт в 1983 году. Автор 20 печатных работ. Область научных интересов: системный анализ, математическое моделирование. E-mail: vilrakh@mail.ru

Estimation of uncertainty for logic output produced by multi-dimensional data hypercubes

A.A. Akhrem, V.Z. Rakhmankulov, T.L. Vashevnik, K.V. Yuzhanin

This paper investigates the problem of uncertainty estimation for logical output produced by multidimensional hypercube structures under the conditions of real disturbances of data in information databases and multidimensional data warehouses. The meaning of measure of data uncertainty is introduced. The criteria of decrease (increase) of uncertainty measures for logic output have been received taking into account the different changes of hypercube basic parameters.

Keywords: hypercube, hypercube multidimensional data, uncertainty measure, entropy, entropy measure, a set of cross-sections.

References

1. Makarov I.M., Rakhmankulov V.Z., Akhrem A.A., Rovkin I.O. Postroenie sistem podderzhki prinyatiya resheniy na osnove OLAP-tehnologii // Informatsionnye tekhnologii i vychislitel'nye sistemy. -2005.-№1.-p.19-30.
2. Makarov I.M., Rakhmankulov V.Z., Akhrem A.A., Rovkin I.O. Issledovanie svoystv giperkubovykh struktur v OLAP-sistemakh // Informatsionnye tekhnologii i vychislitel'nye sistemy. -2005.-№2.-p.4-9.
3. Makarov I.M., Akhrem A.A., Rakhmankulov V.Z., Akhrem A.A., Vashevnik T.L., Filyukov R.Yu. Entropiynye metody analiza informatsii // Trudy ISA RAN. -2011.-Vyp.1, t.1.-p.36-40.
4. Akhrem A.A., Makarov I.M., Rakhmankulov V.Z. Matematicheskaya teoriya virtualizatsii protsessov proektirovaniya i transfera tekhnologii. - M.: Fizmatlit, 2013.-316 p.
5. Gray Robert M. Entropy and Information Theory, 2nd edition – Springer, 2011. – 409 p.
6. Laritchev O.I. Verbalnyi analiz resheniy. –M.: Nauka, 2006. -181p.
7. Vagin V.N., Golovina E.U., Zagorianskaya A.A., Fomina M.V. Dostoverniy i pravdopodobnyi vyvod v intellektualnykh sistemakh. 2-e izd. - M.: Fizmatlit, 2008. 712 p.
8. Osipov G.S. Metody iskusstvennogo intellekta. -M.: Fizmatlit, 2011.-296 p.
9. Popkov J.S. Matematicheskaya demoeconomica. Makrosystemny podhod. – M.: LENAND, 2013, 560 p.
10. Bronshteyn I.N., Semendyaev K.A. Spravochnik po matematike dlya inzhenerov i uchashhikhsya vtuzov. - M.: Nauka, 1981.-387 p.
11. Vygodskiy M.Ya. Spravochnik po elementarnoy matematike. - M.: AST: Astrel', 2014.-509p.
12. Petrovsky A.B., Roizenzon G.V. Interaktivnaya procedura cnizhenia razmernosti priznakogo prostranstva v zadachah mnogokriteriakhnoy klassifikatsii// Podderzka priniatia reshenii: Trudy Instituta sistemnogo analiza RAS/ Pod red. A.B. Petrovskogo. M.: Izd. LKI/URSS, 2008. T.35. p. 43-53.

Rakhmankulov Vil Zakirovich. Head of Laboratory, ISA FRC CSC RAS. Graduated from Moscow Aviation Institute in 1960. Doctor of Sciences, professor. Author of more than 175 scientific papers and monographs. Research areas of interest: system analysis, virtual modeling, intelligent data analysis, business analytics, control and management theory, computer-based industrial automation. E-mail: vilrakh@mail.ru

Akhrem Andrey Afanasievich. Senior scientist, ISA FRC CSC RAS. Graduated from Moscow State University in 1977. Candidate of Sciences. Author of more than 110 scientific papers and monographs. Research areas of interest: mathematical theory of systems, mathematical and virtual modeling of complex technological systems. E-mail: vilrakh@mail.ru

Yuzhanin Kirill Viktorovich. Research engineer, ISA FRC CSC RAS. Graduated from Moscow Physics Engineering University in 2012. Author of 1 scientific paper. Research areas of interest: pattern recognition theory, system analysis. E-mail: vilrakh@mail.ru

Vashevnik Tatiana Leonidovna. Researcher, ISA FRC CSC RAS. Graduated from Moscow Economic Statistics Institute in 1983. Author of 20 scientific papers. Research areas of interest: system analysis, mathematical modeling. E-mail: vilrakh@mail.ru