ОПТИМИЗАЦИЯ РАЗРАБОТКИ ДАННЫХ ПОСРЕДСТВОМ ПРОГНОЗИРУЮЩЕГО МОДЕЛИРОВАНИЯ

OPTIMIZATION OF DATA PROCESSING BASED ON PREDICTIVE MODELING

I. Kutsenko
N. Pavlova

Summary. The article considers multidimensional data analysis with using of OLAP technologies based on the example of the insurance company. The problem of predictive modeling using OLAP cubes and application of multidimensional data analysis in the insurance model with various insurance sums and partial losses are considered. As an analytical task is set up to calculate the net rate and gross rate and analyze the moving average of insurance company.

Keywords: Database, predicting in the insurance system, OPLAP technology, net rate, gross rate, loss-making company, moving average, standard deviation, insured sums, partial losses.

Куценко Ирина Львовна
К.ф.-м.н., доцент, Российский университет дружбы народов (РУДН)
i.kutsenko@mail.ru
Павлова Наталья Геннадиевна
К.ф.-м.н., доцент, Российский университет дружбы народов (РУДН)
natasharussia@mail.ru

Аннотация. В статье рассматривается многомерный анализ данных с привлечением OLAP технологий на примере работы страховой компании. Рассматривается проблема прогнозирующего моделирования с использованием OLAP-кубов и применение многомерного анализа данных в модели страхования с различными страховыми суммами и частичными убытками. Ставится аналитическая задача по вычислению нетто-ставки и брутто-ставки и производится анализ по скользящему среднему страховой компании.

Ключевые слова: База данных, прогнозирование в системе страхования, OPLAP технологии, нетто-ставка, брутто-ставка, убыточность компании, скользящее среднее, среднее квадратическое отклонение, страховые суммы, частичные убытки.

На сегодняшний день, многие компании и организации приходят к мнению, что без существования объективных, своевременных данных о положении рынка, анализа взаимоотношений с конкурентами и партнёрами, прогнозировании его перспектив, постоянной оценки эффективности функционирования собственных структур, в дальнейшем их развитие станет фактически нереальным. Ввиду этого, на сегодняшний день проявляется внимание к средствам реализации и концепциям построения информационных систем, которые ориентированы на аналитическую обработку данных. И первым делом это касается систем управления базами данных, которые основаны на многомерном подходе.

1. Многомерный анализ данных

Согласно определениям Online Analytical Processing (OLAP), OLAP-куб, Реляционные данные, пользователь оперирует в многомерной модели данных со следующими понятиями:

- Измерение (Dimension)
- Ячейка (Cell) [1].

В многомерной модели данных Измерения играют роль индексов, используемых для определения конкретных значений (Показателей), находящихся в ячейках куба.

Показатель — это поле (обычно цифровое), у которого значения определяются фиксированным набором Измерений. Показатель может быть представлен переменной (Variable) или формулой (Formula).

Операции многомерного анализа данных: Кросс-детализация(Drill-across), проекция(Projection), смена измерений, кубический срез(Dice), укрупнение(Roll-up).

Кубический срез (Dice). Эта операция позволяет выбрать подмножество необходимых точек из всего n-мерного пространства области определения куба, применения к ним критерия P. Эта операция равнозначна подобной операции реляционной алгебры. Операция кубического среза предоставляет возможность ограничить представление по нескольким измерениям. Отсюда следует, что получаем подкуб исходных данных [2].

$$\text{Срез}(x) = \boldsymbol{\sigma p}(\text{Сисх}(x)) = \begin{cases} \text{Сисх}(x), \text{ если } P(x) \\ \text{Undef, если } P(x) \end{cases}$$

2. Модель страхования с различными страховыми суммами и частичными убытками

Далее рассмотрим применение многомерного анализа данных в модели страхования с различными страховыми суммами и частичными убытками.

Введём следующие обозначения:

V_j — страховая сумма для j-го клиента, j = 1,…, n;

S_j — частичный убыток j-го клиента, $0 \le S_j \le V_j$;

$N_j = \begin{cases} 1, & \text{q — число исков по j-му договору;} \\ 0, & \text{1-q} \end{cases}$

$X = \sum_{j=1}^{n} X_j$ — общее возмещение по портфелю.

Введём стандартную нормальную случайную величину

$$Z = \frac{X - m_X}{\sigma_X}$$

С математическим ожиданием

$$E\{Z\} = E\left\{\frac{X - m_X}{\sigma_X}\right\} = \frac{E\{X\} - m_X}{\sigma_X} = 0$$

И дисперсией

$$Var\{Z\} = Var\left\{\frac{X - m_X}{\sigma_X}\right\} = \frac{Var\{X\}}{\sigma^2_X} = 1$$

Определим аргумент функции распределения $x_{кр}$

$F(x) = P\{Z \le x_{кр}\} = \gamma$ [3]

Следовательно, среднее значение случайной величины Z не превысит критического значения с вероятностью γ

$$Z = \frac{X - m_X}{\sigma_X} \le x_{кр}(\gamma) \qquad \textbf{(2.1)}$$

Запишем неравенство (1) для величины общего возмещения по портфелю

$$X \le m_X + x_{кр}(\gamma) \times \sigma_X \qquad \textbf{(2.2)}.$$

Нужно выразить числовые характеристики m_X, σ_X через статистические оценки, полученные по данным за прошлый период, для того чтобы мы могли пользоваться формулой (2.2).

Допустим, что нам известны данные за прошлый период:

n^* — число заключённых договоров,

v_k^* — страховые суммы договоров, k = 1,…, n*,

N^* — общее число исков,

S^* — средний ущерб, который возместила компания,

S_k^* — ущерб, возмещённый компанией k-му клиенту, предъявившему иск.

Будем искать несмещённые оценки математического ожидания и дисперсии величины ущерба на одного клиента, используя данную статистическую информацию:

$S^* = \frac{1}{N^*} \sum_{k=1}^{N^*} S_k^*$ — несмещённая оценка среднего ущерба на клиента,

$$(R^*)^2 = \frac{1}{N^* - 1} \sum_{k=1}^{N^*} (S_k^* - S^*)^2$$

— несмещённая оценка дисперсии ущерба.

Также запишем вспомогательные приближённые неравенства для выражения основных числовых характеристик возмещения по портфелю через статистические оценки.

1) $\sum_{j=1}^{n} E\{S_j\} \approx n \times S^*$ — ущерб по j-му договору

2) $\sum_{j=1}^{n} Var\{S_j\} \approx n \times (R^*)^2$, где $(R^*)^2$ — несмещённая оценка дисперсии ущерба каждого договора по данным прошлого года

3) $\sum_{j=1}^{n} E^2\{S_j\} \approx n(S^*)^2$

Выразим числовые характеристики общего возмещения по портфелю через статистические данные, используя полученные приближённые равенства:

$$E\{X\} = E\sum_{j=1}^{n} X_j = \sum_{j=1}^{n} E\{X_j\} =$$

$$= \sum_{j=1}^{n} E\{X_j \times N_j\} = \sum_{j=1}^{n} E\{X_j\} \times E\{N_j\} =$$

$$= \sum_{j=1}^{n} E\{S_j\} \times q = n \times S^* \times q \qquad \textbf{(2.3)}$$

Формулу для дисперсии получим с учётом предпосылки независимости случайных величин $X_j, j = 1, \ldots, n$:

$$Var\{X\} = Var\sum_{j=1}^{n} X_j = \sum_{j=1}^{n} Var\{X_j\} =$$

$$= \sum_{j=1}^{n} [E\{X^2{}_j\} - E^2\{X_j\}]$$

Выразим члены, которые входят под знак суммы, через выборочные данные:

$$E\{X^2_j\} = E\{S^2_j \times N^2_j\} = E\{S^2_j\} \times E\{N^2_j\} = E\{S^2_j\} \times q = (Var\{S_j\} + E^2\{S_j\}) \times q,$$

Так как $E\{N^2_j\} = 1^2 \times q + 0^2 \times (1-q) = q$,

$$E^2\{Xj\} = (E\{Xj\})^2 = (E(S_j \times N_j))^2 = (E\{S_j\} \times q)^2 =$$

$$= E^2\{S_j\} \times q^2$$

Итак, формула для дисперсии возмещения принимает вид

$$Var\{X\} =$$

$$= \sum_{j=1}^{n}[Var\{S_j\} \times q + E^2\{S_j\} \times q - E^2\{S_j\} \times q^2] =$$

$$= q\sum_{j=1}^{n} Var\{S_j\} + q(1\text{-}q)\sum_{j=1}^{n} E^2\{S_j\} = q \times n(R^*)^2 +$$

$$+ q(1\text{-}q) \times n(S^*)^2 \quad \textbf{(2.4)}$$

Тогда среднее квадратичное отклонение (ско) имеет вид

$$\sigma_X = \sqrt{qn[(R*)^2 + (1-q) \times (S*)^2]} \quad \textbf{(2.5)}$$

Подставим в (2.2) выражения числовых характеристик через выборочные данные (2.3), (2.5) и получим:

$$X \le n \times S^* \times q +$$

$$+ x_{кр} \times \sqrt{qn[(R*)^2 + (1-q) \times (S*)^2]} \quad \textbf{(2.6)}$$

В формулу (2.6) входят величины, оцененные по данным прошлого года.

Преобразуем эту формулу. Рассмотрим величину:

$S = \frac{1}{n}\sum_{j=1}^{n} v_j$ — среднюю страховую сумму для клиентов текущего года.

Теперь разделим неравенство (2.4) на nS (общую страховую сумму по рискам, принятым на страхование в данном году) и умножим на 100

$$\frac{100 \times X}{nS} \le \frac{S* \times 100}{S} \times q +$$

$$+ x_{кр} \times T_0 \times \sqrt{\frac{[\left(\frac{R*}{S*}\right)^2 + (1-q)]}{qn}} \quad \textbf{(2.7)}$$

$$Y = \frac{100 \times X}{nS} = \frac{X}{nS{:}100}$$

— нетто-ставка (фактическая убыточность страховой суммы со 100 рублей страховой суммы) — размер выплаты компании на каждые 100 рублей общей страховой суммы.

$$T_0 = \frac{S* \times 100}{S} \times q \quad \textbf{(2.8)}$$

— чистая нетто-ставка (сумма, которая берётся с каждых 100 рублей).

При решении задачи для её вычисления удобней пользоваться следующей формулой:

$$T_0 = \frac{S* \times 100}{S} \times q = \frac{q \times S* \times n}{n \times S{:}100} = \frac{E\{X\}}{n \times S{:}100}$$

$$T_p = x_{кр} \times T_0 \times \sqrt{\frac{[\left(\frac{R*}{S*}\right)^2 + (1-q)]}{qn}} \quad \textbf{(2.9)}$$

— рисковая надбавка.

При решении задачи для её вычисления удобней пользоваться следующей формулой:

$$T_0 = \frac{S* \times 100}{S} \times q = \frac{q \times S* \times n}{n \times S{:}100} = \frac{E\{X\}}{n \times S{:}100}$$

С учётом (2.8) и (2.9) выражение (2.7) принимает вид

$$Y \le T_0 + T_p$$

Обозначим нетто-ставку как T_H и запишем её структуру

$$T_H = T_0 + T_p \quad \textbf{(2.10)}$$

Запишем связь брутто-ставки с нетто-ставкой:

$$T_б = T_H + \frac{f}{100} \times T_б, \; т.е. \; T_б = \frac{100}{100-f} \times T_H, \quad \textbf{(2.11)}$$

где $T_б$ — брутто ставка;

f — нагрузка, идущая на выплаты сотрудникам.

Убыточность страховой суммы — это отношение суммы страховых выплат к страховой сумме застрахованных объектов (максимально возможная страховая выплата).

Введём обозначения:

У — убыточность;

СВ — сумма страховой выплаты:

СС — страховая сумма застрахованных объектов.

Рассчитаем убыточность страховой суммы по формуле:

$$У = \frac{СВ}{СС} * 100\% \quad \textbf{(2.12)}$$

3. Анализ по скользящему среднему

Скользящее среднее (Moving Average) — семейство функций, у которых значения в каждой точке опре-

База данных страховой компании А (Рис. 1).

год	средняя страховая сумма на каждого клиента	сумма страховой выплаты	число предъявленных исков	число договоров	нагрузка, идущая на выплату з/п сотрудникам
2006	20000	15000	10	505	32
2007	30250	13500	11	281	32
2008	40000	11000	40	400	32
2009	42350	8000	25	350	32
2010	31000	17000	8	1000	32
2011	24650	16500	17	625	32
2012	52300	19320	22	777	32
2013	46555	14400	20	1500	32
2014	35250	18500	13	512	32
2015	60000	12000	15	300	32
2016	25000	20000	19	4500	32

деления равны среднему значению исходной функции за предыдущий период.

$$\text{Moving average} = \frac{\sum_{i=1}^{n} P_i}{n},$$

где P_i — цена,

N — период скользящей средней. Это основной параметр при построении, его еще называют длина сглаживания.

Для обнаружения основных тенденций, как правило, используют скользящее среднее.

Взвешенное скользящее среднее (Weighted Moving Average) — скользящее среднее, при вычислении которого значение каждого члена исходной функции, начиная с меньшего, равно соответствующему члену арифметической прогрессии. То есть, при вычислении взвешенного скользящего среднего, мы считаем последнее значение исходной функции более значимым, чем предыдущие. Взвешенное скользящее среднее вычисляется по следующей формуле:

$$\text{ВВС}_t = \sum_{i=0}^{n-1} w_{t-i} * p_{t-i},$$

где ВВС_t — значение взвешенного скользящего среднего в точке t;

n — количество значений исходной функции для расчёта скользящего среднего;

w_{t-i} — нормированный вес (весовой коэффициент) t-i-го значения исходной функции;

p_{t-i} — значение исходной функции в момент времени, отдалённый от текущего на i интервалов.

Простое скользящее среднее численно равно среднему арифметическому значений исходной функции за определённый период времени. Вычисляется по следующей формуле:

$$\text{СС}_t = \frac{1}{n}\sum_{i=0}^{n-1} p_{t-i} = \\ = \frac{p_t + p_{t-1} + \cdots + p_{t-i} + \cdots + p_{t-n+2} + p_{t-n+1}}{n},$$

где СС_t — значение простого скользящего среднего в точке *t*;

t — количество значений исходной функции для расчёта скользящего среднего;

p_{t-i} — значение исходной функции в точке $t–i$.

Кумулятивное скользящее среднее (Cumulative Moving Average) численно равно среднему арифметическому значений исходной функции за весь период наблюдений:

$$\text{ССС}_t = \frac{1}{t}\sum_{i=1}^{t} p_i = \frac{p_t + p_{t-1} + \cdots + p_2 + p_1}{t},$$

где ССС_t кумулятивное скользящее среднее в момент t, t — количество доступных для вычисления интервалов, p_i — значение исходной функции в точке i.

4. Прогноз разорения страховой компании

Представим базу данных в виде OLAP-куба со следующими осями: время, количество договоров, средняя страховая сумма на каждого клиента, число предъявленных исков, сумма страховой выплаты.

Проведём кубический срез за период времени: 2016–2017 года и рассчитаем брутто-ставку, нетто-ставку.

Таблица 1. Данные для подсчета скользящего среднего

n*	300		q ≈ (N*)/(n*)	0,05
N*	15		T0=((S*×100)/S)×q	1,2
S*	60000		$T_p = x_{кр} \times T_0 \times \sqrt{\frac{\left[\left(\frac{R*}{S*}\right)^2 + (1-q)\right]}{qn}}$	0,161847
R*	45000			
N	4500			
S	250000		$T_H = T_0 + T_p$	1,361847
F	32		$T_б = \frac{100}{100 - f} \times T_H$	2,002716
γ	0,95			
γ(0,95)=	1,645			
			$y = \frac{CB}{S} * 100\%$	8
CB	20000			

И проведём кубический срез за период времени — 2017 год и рассчитаем убыточность компании за 2017 год.

Данные за 2016 год:

1) Число договоров $n* = 300$

2) Число предъявленных исков N* = 15

3) Средняя страховая сумма на каждого клиента $S* = 60000$ рублей

4) Оценка скол возмещения R* = 45000 рублей

(R* = $\sqrt{\frac{1}{N*-1} \sum_{k=1}^{N*} (S_k * - S*)^2}$

рассчитывается по этой формуле)

Данные за 2016 год:

1) Число договоров $n = 4500$

2) Средняя страховая сумма на каждого клиента $S = 250000$ рублей

$f = 32\%$ — нагрузка, идущая на выплату зарплаты сотрудникам

$\gamma = 0,95$ — уровень надёжности (с вероятностью 0,95 все иски будут удовлетворены).

СВ=20000

Оценим вероятность страхового события:

$$q \approx \frac{N*}{n*} \approx 0,05$$

По формуле (4) рассчитаем чистую нетто-ставку

$$T_0 = \frac{S* \times 100}{S} \times q = 1,2$$

(с каждых 100 рублей берётся страховая сумма в размере 1,2 рубля)

Возьмём критическое значение из таблиц нормального распределения:

$$\gamma(0,95) = 1,645$$

$$T_p = x_{кр} \times T_0 \times \sqrt{\frac{\left[\left(\frac{R*}{S*}\right)^2 + (1-q)\right]}{qn}} = 0,161847$$

Рассчитаем нетто-ставку по формуле (10):

$$T_H = T_0 + T_p = 1,361847$$

Таблица 2

Год	средняя страховая сумма на каждого клиента	скользящее среднее по 2 годам	скользящее среднее по 3 годам	скользящее среднее по 4 годам
2006	20000			
2007	30250			
2008	40000	25125		
2009	42350	35125	30083,33333	
2010	31000	41175	37533,33333	33150
2011	24650	36675	37783,33333	35900
2012	52300	27825	32666,66667	34500
2013	46555	38475	35983,33333	37575
2014	35250	49427,5	41168,33333	38626,25
2015	60000	40902,5	44701,66667	39688,75
2016	25000	47625	47268,33333	48526,25
2017		42500	40083,33333	41701,25

Год	сумма страховой выплаты	скользящее среднее по 2 годам	скользящее среднее по 3 годам	скользящее среднее по 4 годам
2006	15000			
2007	13500			
2008	11000	14250		
2009	8000	12250	13166,66667	
2010	17000	9500	10833,33333	11875
2011	16500	12500	12000	12375
2012	19320	16750	13833,33333	13125
2013	14400	17910	17606,66667	15205
2014	18500	16860	16740	16805
2015	12000	16450	17406,66667	17180
2016	20000	15250	14966,66667	16055
2017		16000	16833,33333	16225

Рассчитаем брутто-ставку по формуле (11):

$$T_б = \frac{100}{100-f} \times T_Н = 2,002716$$

Рассчитаем убыточность компании по формуле (12):

$$У = 8\%$$

Далее проведём анализ по скользящему среднему для страховой компании А (таблица 2).

В результате была получена нетто-ставка за 2016 год, равную 1,361847, брутто-ставку, равную 2,002716 и убыточность компании за 2016 год, равную 8% (или 8 рублей с каждых 100 рублей) и провели анализ по скользящему среднему для страховой компании А. Также дан прогноз на 2017 год по сумме страховой выплаты и по средней страховой суммы на каждого клиента. Для того чтобы уменьшить убыточность компании, мы должны уменьшить количество страховых выплат. Расчёт брутто-ставки неразрывно связан со всей деятельностью страховой компании. Также она влияет на уровень развития компании, прибыль и затраты.

ЛИТЕРАТУРА

1. Коннолли Т., Бегг К., Страчан А. Базы данных: Проектирование, Реализация и сопровождение. Теория и практика — М.: Вильямс, 2000. С. 981–991
2. Висков А. В. Модель многомерного представления данных и методы ее анализа: Автореф. дис. на соиск. учен. ст. канд. физ-мат. наук/ А. В. Висков — М., 2010
3. Бабешко Л. О. Математическое моделирование финансовой деятельности // Учебное пособие. Москва: Изд-во КноРус 2011

Журнал «Современная наука: актуальные проблемы теории и практики»