МЕТОДЫ КОНТРОЛЯ ЧИСЛОВЫХ ПОКАЗАТЕЛЕЙ

ГУСЕВ Кирилл Вячеславович
старший преподаватель
ЛЕОНТЬЕВ Александр Савельевич
кандидат технических наук
ФГБОУ ВО «МИРЭА – Российский технологический университет»
г. Москва, Россия

В статье рассматриваются вопросы контроля числовых показателей в базах данных большого объема. Приводится характеристика различных классов числовых показателей. Описывается оригинальный методический подход грубого и уточненного контроля значений числовых показателей годичной периодичности 1-го и 2-го классов.

Ключевые слова: числовые показатели, метод контроля числовых показателей, базы данных большого объема.

В базах данных большого объема (хранилищах данных), содержащих миллионы значений различных показателей, принципиально невозможно обеспечить полностью достоверную информацию даже в том случае, когда первичная информация является безошибочной, так как всегда существуют ошибки операторов, осуществляющих ввод и контроль информации. Поэтому при разработке технологий контроля необходимо использование многоуровневых систем контроля и корректировки данных, включающих не только визуальный, но и программный контроль, на этапах как обновления, так и эксплуатации хранилищ данных. При этом должен быть обеспечен необходимый уровень достоверности информации в хранилищах.

При исследовании и реализации информационных технологий, использующих распределенные хранилища данных, возникают задачи оценки показателей достоверности (безошибочности) числовых показателей, разработки систем контроля информации, а также задачи, позволяющие на основе специальных технологических процедур выбирать для использования при решении определенных проблем согласованные корректные числовые показатели, достоверность которых более чем на порядок превышает достоверность числовых показателей в распределенных хранилищах в среднем.

Обобщенная характеристика различных классов числовых показателей. Как правило, числовые показатели федерального уровня, уровня федеральных округов и уровня субъектов РФ являются относительно устойчивыми и их изменение на интервале 1 год обычно не превышает 50%. Числовые показатели уровня города, промышленного предприятия и технического объекта могут изменяться на интервале 1 год в несколько раз.

Поэтому при разработке методики контроля достоверности целесообразно все числовые показатели разбить на два класса: 1 класс – показатели федерального, окружного и регионального уровней, 2 класс – показатели уровня города, промышленного предприятия и технического объекта. Диапазон изменения числовых коэффициентов, характеризующих показатели 1-го и 2-го классов, а также критерии оценки грубого и уточненного контроля достоверности показателей различных классов различаются для 1-го и 2-го классов.

Грубый и уточненный контроль числовых показателей годичной периодичности 1 класса. Каждому числовому показателю $A_i^{(1)}$ первого класса приписываются определяемые экспертно значения числовых коэффициентов $\{K_{ij}^{(1)}\}$, j=1,2,3, где $K_{i1}^{(1)} = n_i$ – количество лет (выборка), предшествовавших рассматриваемому году, за которые имеются числовые показатели $A_{i,m}^{(1)}$ ($m = 1, 2, \ldots n_i$).

Эта выборка будет использоваться при контроле достоверности числовых показателей;

$K_{i2}^{(1)}$ – числовой коэффициент, используемый для грубого контроля (первичного контроля) достоверности показателей $A_i^{(1)}$, если n_i>1;

$K_{i3}^{(1)}$ – числовой коэффициент, используемый для уточненного контроля достоверности $A_i^{(1)}$, если грубый контроль успешно завершен.

Грубый контроль числовых показателей. При грубом контроле оценивается порядок отношений $A_{i,m+1}^{(1)} / A_{i,m}^{(1)}$ ($m = 1, 2, \ldots, n_i$).

Если $|(A_{i,m+1}^{(1)} / A_{i,m}^{(1)}) - 1| \le K_{i2}^{(1)}$ ($m = 1, \ldots, n_i$ -1), то выборка корректна.

Если $|(A_{i,n_i+1}^{(1)} / A_{i,n_i}^{(1)}) - 1| \le K_{i2}^{(1)}$ при корректной выборке, то показатель $A_i^{(1)}$ прошел грубый контроль на достоверность.

Процедура регуляризации выборки. Если выборка некорректна, то прежде чем проводить грубый контроль необходимо осуществить регуляризацию выборки. Регуляризация осуществляется удалением из выборки тех элементов, для которых нарушено отношение порядка: то есть если $|(A_{i,m+1}^{(1)} / A_{i,m}^{(1)}) - 1| > K_{i2}^{(1)}$ для некоторого m ($m = 1, \ldots n_i$ -1), то удаляется элемент выборки $A_{i,m+1}^{(1)}$.

Если после регуляризации количество членов выборки будет не меньше 2, то переходим к процедуре грубого контроля $A_i^{(1)}$ (точнее $A_{i,n_i+1}^{(1)}$).

Если грубый контроль для $A_i^{(1)}$ успешно завершен, переходим к процедуре уточненного контроля.

Уточненный контроль числовых показателей. Если при грубом контроле осуществлялась регуляризация выборки, то при уточненном контроле в случае n_i=2 используется линейная интерполяционная модель [1], если n_i>2 будет использоваться линейная аппроксимационная модель - аппроксимация осуществляется по методу наименьших квадратов [2] по всей регуляризованной выборке.

Если грубый контроль был выполнен без регуляризации выборки, используется процедура выбора модели прогнозирования.

Процедура выбора модели прогнозирования (прогнозирование значений показателей $A_{i,n_i+1}^{(1)}$).

Если n_i=2, то для прогнозирования используется линейная интерполяционная модель [1].

Если n_i>2, то осуществляется исследование на выпуклость (вогнутость) функции, аппроксимирующей решетчатую функцию выборки.

Пусть $Y_{i,m}^{(1)} = (A_{i,m+1}^{(1)} / A_{i,m}^{(1)})$, ($m = 1, \ldots, n_i$ -1).

Если члены ряда $Y_{i,m}^{(1)}$, ($m = 1, \ldots n_i$ -1) строго убывают ($Y_{i,m+1}^{(1)} < Y_{i,m}^{(1)}$), то аппроксимирующая функция должна быть выпуклой и для аппроксимации используется интерполяционная параболическая модель [3], параметры которой выбираются по трем последним точкам выборки ($A_{i,n_i-2}^{(1)}$, $A_{i,n_i-1}^{(1)}$, $A_{i,n_i}^{(1)}$).

Если члены ряда $Y_{i,m}^{(1)}$, ($m = 1, \ldots n_i$ -1) строго возрастают ($Y_{i,m+1}^{(1)} > Y_{i,m}^{(1)}$), то аппроксимирующая функция должна быть вогнутой и для аппроксимации также используется интерполяционная параболическая модель [3], параметры которой выбираются по трем последним точкам выборки($A_{i,n_i-2}^{(1)}$, $A_{i,n_i-1}^{(1)}$, $A_{i,n_i}^{(1)}$).

Отметим, что если выполняются условия выпуклости или вогнутости, то в первом приближении может быть использована для прогнозирования линейная интерполяционная модель, параметры которой определяются по последним двум точкам выборки [1]. При этом ошибка прогнозирования будет меньше, чем при использовании линейной

аппроксимационной модели, параметры которой определяются по всей выборке [2].

Если не выполняются условия выпуклости или вогнутости, то в качестве модели прогнозирования выбирается линейная аппроксимирующая модель [2], параметры которой определяются по всей выборке по методу наименьших квадратов.

После выбора и настройки параметров аппроксимационной модели (модели прогнозирования) осуществляется прогнозирование показателя $A^{(1)}_{i,n_i+1(прог)}$ и вычисляется отношение

$$Y^{(1)}_{i,n_i+1} = A^{(1)}_{i,n_i+1} / A^{(1)}_{i,n_i+1(прог)}. \quad (1)$$

Затем определяется коэффициент

$$K^{(1)}_{i3(прог)} = |Y^{(1)}_{i,n_i+1} - 1|. \quad (2)$$

Если $K^{(1)}_{i3(прог)} \le K^{(1)}_{i3}$, то считается, что показатель $A^{(1)}_i$ ($A^{(1)}_{i,n_i+1}$) прошел уточненный контроль.

Контроль достоверности числовых показателей годичной периодичности 2 класса.

Отметим, что методика контроля достоверности числовых показателей годичной периодичности для 2-го класса $A^{(2)}_i$ практически совпадает с методикой контроля достоверности показателей годичной периодичности для 1-го класса за исключением процедуры сравнения отношений $A^{(2)}_{i,m+1} / A^{(2)}_{i,m}$ ($m = 1, \ldots, n_i - 1$) с числовым коэффициентом $K^{(2)}_{i2}$ и

$$Y^{(2)}_{i,n_i+1} = A^{(2)}_{i,n_i+1} / A^{(2)}_{i,n_i+1(прог)} \quad (3)$$

с числовым коэффициентом $K^{(2)}_{i3}$.

Каждому числовому показателю $A^{(2)}_i$ второго класса приписываются определяемые экспертно значения числовых коэффициентов $\{K^{(2)}_{ij}\}$ j=1, 2, 3, где $K^{(2)}_{i1} = n_i$ – количество лет (выборка), предшествовавших рассматриваемому году, за которые имеются числовые показатели $A^{(2)}_{i,m}$ ($m = 1, 2, \ldots n_i$). Эта выборка будет использоваться при контроле достоверности числовых показателей;

$K^{(2)}_{i2}$ – числовой коэффициент, используемый для грубого контроля (первичного контроля) достоверности $A^{(2)}_i$, если $n_i > 1$;

$K^{(2)}_{i3}$ – числовой коэффициент, используемый для уточненного контроля достоверности $A^{(2)}_i$, если грубый контроль успешно завершен.

Грубый контроль числовых показателей 2-го класса.

Если $(1/K^{(2)}_{i2}) \le |A^{(2)}_{i,m+1} / A^{(2)}_{i,m}| \le K^{(2)}_{i2}$ ($m = 1, \ldots n_i - 1$), то выборка корректна.

Если $(1/K^{(2)}_{i2}) \le |A^{(2)}_{i,n_i+1} / A^{(2)}_{i,n_i}| \le K^{(2)}_{i2}$, при корректной выборке, то считается, что показатель $A^{(2)}_i$ прошел грубый контроль на достоверность.

Если выборка некорректна, то прежде чем проводить грубый контроль необходимо осуществить регуляризацию выборки.

Регуляризация осуществляется удалением из выборки тех элементов, для которых нарушено отношение порядка. То есть если $|A^{(2)}_{i,m+1} / A^{(2)}_{i,m}| > K^{(2)}_{i2}$ или $|A^{(2)}_{i,m+1} / A^{(2)}_{i,m}| < (1/K^{(2)}_{i2})$, то удаляется элемент выборки $A^{(2)}_{i,m+1}$ ($m = 1, n_i - 1$). Однако, если при этом выполняется условие $(1/K^{(2)}_{i2}) \le |A^{(2)}_{i,m+2} / A^{(2)}_{i,m+1}| \le K^{(2)}_{i2}$, то удаляется элемент выборки $A^{(2)}_{i,m}$.

Если после регуляризации количество членов выборки будет не меньше 2, то переходим к процедуре грубого контроля показателя $A^{(2)}_i$ ($A^{(2)}_{i,n_i+1}$).

Если грубый контроль для $A^{(2)}_i$ успешно

завершен, переходим к процедуре уточненного контроля.

Уточненный контроль для числовых показателей 2-го класса.

Процедура выбора модели прогнозирования числовых показателей 2-го класса полностью соответствует процедуре выбора модели прогнозирования числовых показателей 1-го класса.

После выбора и настройки параметров аппроксимационной модели (модели прогнозирования) осуществляется прогнозирование показателя $A^{(2)}_{i,n_i+1(прог)}$ и вычисляется значение

$$Y^{(2)}_{i,n_i+1} = A^{(2)}_{i,n_i+1} / A^{(2)}_{i,n_i+1(прог)}. \quad (4)$$

Если $(1/K^{(2)}_{i3}) \le |Y^{(2)}_{i,n_i+1}| \le K^{(2)}_{i3}$, то считается, что показатель $A^{(2)}_{i,n_i+1}$ прошел уточненный контроль.

Аппроксимационные модели (модели прогнозирования) значений числовых показателей.

Линейная интерполяционная модель.

В том случае, когда $K^{(1)}_{i1} = n_i = 2$ или $K^{(2)}_{i1} = n_i = 2$, то есть числовой показатель (в том числе после регуляризации) n_i, характеризующий выборку элементов $A^{(1)}_i$ или $A^{(2)}_i$ за предшествующие годы, равен 2 используется линейная интерполяционная модель.

Пусть $\mathbf{x}_1$ и $\mathbf{x}_2$ значение года, предшествующего рассматриваемому году $\mathbf{x}=\mathbf{x}_3$, для которого рассчитывается прогнозные значения числовых показателей.

a_{ix1} – значение показателя $A^{(1)}_i$ или $A^{(2)}_i$ в x_1 году.

a_{ix2} – значение показателя $A^{(1)}_i$ или $A^{(2)}_i$ в x_2 году.

Тогда интерполяционная линейная функция равна:

$$\mathbf{a}_{ix3прог}(x)=\mathbf{a}_{ix1}*\mathbf{L}_0(x)+\mathbf{a}_{ix2}*\mathbf{L}_1(x)(\mathbf{x}=\mathbf{x}_3). \quad (5)$$

$$\mathbf{L}_0(x)=(\mathbf{x}-\mathbf{x}_2)/(\mathbf{x}_1-\mathbf{x}_2);\ \mathbf{L}_1(x)=(\mathbf{x}-\mathbf{x}_1)/(\mathbf{x}_2-\mathbf{x}_1). \quad (6)(7)$$

Если $\mathbf{x}_2-\mathbf{x}_1=1$, то

$$\mathbf{L}_0(x) = -(\mathbf{x}-\mathbf{x}_2);\ \mathbf{L}_1(x)=(\mathbf{x}-\mathbf{x}_1). \quad (7)(8)$$

Если $\mathbf{x}=\mathbf{x}_3$ и $\mathbf{x}_2-\mathbf{x}_1 = 1$, $\mathbf{x}_3-\mathbf{x}_1=2$,

то $$\mathbf{L}_0(\mathbf{x}) =\mathbf{L}_0(\mathbf{x}_3)=-(\mathbf{x}_3-\mathbf{x}_2)= -1; \quad (9)$$

$$\mathbf{L}_1(\mathbf{x}) =\mathbf{L}_1(\mathbf{x}_3)= (\mathbf{x}_3-\mathbf{x}_1)=2; \quad (10)$$

Тогда имеет место следующая расчетная формула, в соответствии с которой осуществляется прогнозирование числовых показателей:

$$\mathbf{a}_{ix3прог}(\mathbf{x})=\mathbf{a}_{ix3прог}(\mathbf{x}_3)=-\mathbf{a}_{ix1} + 2\mathbf{a}_{ix2}. \quad (11)$$

Линейная аппроксимационная модель (метод наименьших квадратов).

Пусть объем выборки n_i числового показателя A_i больше 2:

$$K^{(1)}_{i1} = n_i > 2 \text{ или } K^{(2)}_{i1} = n_i > 2. \quad (12)$$

Пусть имеется множество годов $\mathbf{x}_1$, $\mathbf{x}_2$, …$\mathbf{x}_n$, предшествующих $\mathbf{x}_{n+1}$ году ($n=n_i$), для которого осуществляется прогнозирование числового показателя A_{i,n_i+1}.

Пусть A_{ij} значение показателя $A^{(1)}_{ij}$ или $A^{(2)}_{ij}$ в j-ом году (x_j), j=1,…n.

$A_{i,(прог)}(x_{n+1})$ – прогнозное значение показателя $A^{(1)}_{i,n+1}$ или $A^{(2)}_{i,n+1}$ в x_{n+1} году, тогда:

$A_{i(прог)}(x_{n+1}) = ax_{n+1} + b$ (линейная модель

$A_{i(прог)}(x) = ax + b$). (13)

Определение параметров **a** и **b** линейной модели.

Сумма квадратов отклонений $F_i(a,b)$ значений линейной функции от выборочных значений числового показателя A_{ij} (j=1,…n) определяется соотношением:

$$F_i(a,b)= \sum_{j=1}^{n} (aX_j + b - A_{ij})^2. \quad (14)$$

Из условий равенства нулю частных производных функции $F_i(a,b)$

$\frac{\partial F_i(a,b)}{\partial a} = 0$ и $\frac{\partial F_i(a,b)}{\partial b} = 0$ получим следующие расчетные соотношения для определения параметров **a** и **b**:

$$a = \frac{n\sum_{j=1}^{n} x_j A_{i,j} - \sum_{j=1}^{n} x_j \sum_{j=1}^{n} A_{i,j}}{n\sum_{j=1}^{n} x_j^2 - (\sum_{j=1}^{n} x_j)^2}; \quad (15)$$

$$b = \frac{(\sum_{j=1}^{n} x_j^2 \sum_{j=1}^{n} A_{ij}) - (\sum_{j=1}^{n} x_j \sum_{j=1}^{n} x_j A_{i,j})}{n\sum_{j=1}^{n} x_j^2 - (\sum_{j=1}^{n} x_j)^2}. \quad (16)$$

На основе реальных данных из хранилища, проведено исследование различных классов аппроксимационных моделей, позволяющих прогнозировать значения числовых показателей годичной периодичности.

Исследовались модели 1-ой (линейные), 2-ой (параболические), 3-ей, 4-ой, 5-ой, 6-ой и более высоких степеней, в которых использовались аппроксимирующие полиномы (интерполяционные полиномы) соответствующей степени. Показано, что использование полиномов 3-ей, 4-ой, 5-ой, 6-ой и более высоких степеней для хранилища данных большой размерности не является оправданным, так как в большинстве случаев приводит к существенным ошибкам прогнозирования. Причем ошибки прогнозирования на основе многополиномиальных моделей превышают в большинстве случаев ошибки прогнозирования на основе линейных и параболических моделей.

Заключение. Были получены следующие основные результаты:

1. Проведена обобщенная характеристика различных классов числовых показателей годичной периодичности. Предложено все числовые показатели разбить на два класса – показатели федерального, окружного и регионального уровней (1 класс), показатели уровня города, промышленного предприятия и технического объекта (2 класс).

2. Разработан оригинальный методический подход грубого и уточненного контроля значений числовых показателей годичной периодичности 1-го и 2-го классов.

3. Получены расчетные формулы для прогнозирования значений числовых показателей годичной периодичности на основе линейных интерполяционных моделей, линейных аппроксимационных моделей (метод наименьших квадратов) и интерполяционных параболических моделей.

СПИСОК ЛИТЕРАТУРЫ

1. *Бронштейн И.Н. Семендяев К.А.* Справочник по математике для инженеров и учащихся втузов. 13-е изд. исправленное. – М.: Наука. Гл. ред. физ.-мат. лит., 1986. – 544 с.

2. *Гмурман В.Е.* Теория вероятностей и математическая статистика. Учебное пособие для втузов. Изд. 5-е, перераб. и доп. – М.: Высшая школа, 1977. – 479 с.

3. *Корн Г., Корн Т.* Справочник по математике для научных работников и инженеров. – М.: Наука. Гл. ред. физ.-мат. лит., 1973. – 832 с.