УДК 519.25

А.И. Новиков, М.Е. Ильин

АЛГОРИТМЫ ВОССТАНОВЛЕНИЯ НЕСОСТОЯТЕЛЬНЫХ И ПРОПУЩЕННЫХ ДАННЫХ

Приведено описание алгоритмов импутации (восстановления) данных, вошедших в автоматизированную систему обработки результатов Всероссийской сельскохозяйственной переписи 2006 года (ВСХП 2006).

Введение. В 2006 году впервые в истории современной России проведена Всероссийская сельскохозяйственная перепись. Объектами переписи являлись сельскохозяйственные структуры всех форм собственности. Одной из важнейших проблем любой переписи и, в частности сельскохозяйственной, является достоверность данных, полученных в ходе ее проведения. Актуальность восстановления (технологический процесс, направленный на обнаружение и обработку ошибочных или пропущенных данных) истинных данных не вызывает сомнения. Обычно под редактированием понимают проведение формального логического контроля, направленного на обнаружение ошибок в статистических данных. Такой контроль показателей на непротиворечивость проводит к балансовым соотношениям, заложенным в самой структуре исходных данных:

– уравнениям и неравенствам алгебраического типа для количественных показателей внутри отдельных групп или между несколькими группами; чаще всего это линейные уравнения и неравенства;

– логическим высказываниям для качественных (атрибутивных) показателей.

В случае истинных данных эти балансовые соотношения должны выполняться тождественно. Ошибочные данные, в том числе и пропущенные, могут нарушать уравнения и неравенства. Если ошибочные данные не нарушают балансовые соотношения, то они считаются достоверными. В случае же их нарушения возможны две ситуации.

В первой – пропущенные данные однозначно определяются из балансовых соотношений и, следовательно, их решение позволяет восстановить недостающие данные. Такой технологический процесс называется автокоррекцией. Он выполняется в рамках одной группы данных.

Во второй – балансовые уравнения выделяют некоторое множество, которому должны удовлетворять пропущенные или ошибочные данные. Поэтому требуется сформулировать правила однозначного выбора из такого допустимого множества. Для вывода таких правил необходимо знание статистических свойств всей совокупности. Такой процесс называется импутацией. В дальнейшем в статье описывается именно этот процесс.

Алгоритмы импутации. Пусть результаты наблюдения над статистической совокупностью представлены в виде матрицы X

$$X = \begin{pmatrix} x_{11}\ x_{12}\ \cdots x_{1n} \\ x_{21}\ x_{22}\ \cdots x_{2n} \\ \cdots\cdots\cdots\ \cdots \\ x_{k1}\ x_{k2}\ \cdots x_{kn} \end{pmatrix} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_k \end{bmatrix}, \qquad (1)$$

где X – матрица исходных данных некоторого наблюдения за статистической совокупностью. При этом вектор–строка $X_i = (x_{i1}, x_{i2}, ..., x_{in})$ является вектором состояния i-го экземпляра наблюдения (хозяйства) с заполненными n позициями; вектор $X_j = (x_{1j}, x_{2j}, ..., x_{kj})^T$ – вектор-столбец значений j-го показателя в k экземплярах статистической совокупности. Элементы матрицы X – суть целые неотрицательные числа. Таким образом, здесь и далее:

– n – число наблюдаемых данных в каждом объекте (порядка 500– 1000);

– k – объем статистической совокупности (количество наблюдений) в исходной базе данных (порядка $2 \cdot 10^5 - 10^6$) по каждому региону Российской федерации.

Для каждого j-го показателя X_j вычисляются оценки математического ожидания и СКО по формулам:

$$\hat{m}_{x_j} = \frac{1}{k} \cdot \sum_{s=1}^{k} x_{sj}, \qquad \hat{\sigma}_j = \sqrt{\frac{\sum_{s=1}^{k} \left(x_{sj} - \hat{m}_{x_j}\right)^2}{k}}. \qquad (2)$$

Все наблюдения в (1) разобьем на три части. В первую отнесем те, в которых значение хотя бы одного показателя является нетипичным. Показатели из этой части наблюдений не подлежат восстановлению вследствие экстремальности их значений. Для этой цели по каждому показателю X_j вычисляются нижняя $\underline{x}_j$ и верхняя $\overline{x}_j$ границы "типичности", с помощью которых формируется отрезок $\left[\underline{x}_j;\overline{x}_j\right]$ допустимых значений j-го показателя статистической совокупности (1):

$$\underline{x}_j = \begin{cases} \hat{m}_{x_j} - \alpha_j \cdot \hat{\sigma}_j, & \text{если } \hat{m}_{x_j} - \alpha_j \cdot \sigma_j > 0, \\ 0, & \text{если } \hat{m}_{x_j} - \alpha_j \cdot \sigma_j \le 0, \end{cases}$$

$$\overline{x}_j = \hat{m}_{x_j} + \alpha_j \cdot \hat{\sigma}_j, \qquad (3)$$

где $\alpha_j = 3$.

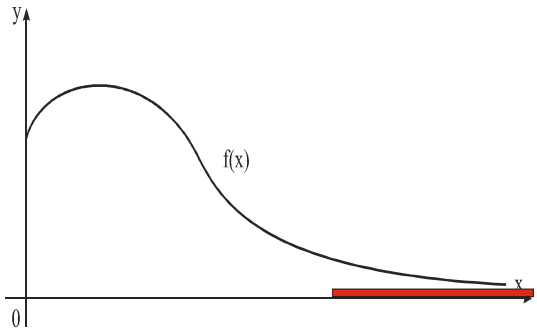
Если хотя бы одно значение показателя экземпляра X_i не принадлежит отрезку допустимых значений $\left[\underline{x}_j;\overline{x}_j\right], j = 1,2,...,n$, то наблюдение X_i помечается как нетипичное.

Во вторую часть (доноры) включим те наблюдения, значения всех показателей которых принадлежат отрезкам допустимых значений $\left[\underline{x}_j;\overline{x}_j\right], j = 1,2,...,n$. Значения показателей из этой части статистической совокупности являются типичными в том смысле, что они в большей степени определяют статистические свойства наблюдений, и они будут использованы для восстановления пропущенных или ошибочных данных.

Замечание. Непосредственное применение формулы (3) к некоторым полям может привести к ситуации, когда множество $\left[\underline{x}_j;\overline{x}_j\right]$ не содержит целых значений. В этом случае используется множество $\left[m_{x_j} - d; m_{x_j} + d\right]$, где d полагается равным 2.05 и заведомо содержащее, по крайней мере, 4 целых значения, наиболее близких к средней.

В третью часть (реципиенты) включаются записи, в которых некоторые данные являются несостоятельными (не отвечают определенным ограничениям) или пропущены. В матрице реципиентов особое положение занимают записи, в которых не заполнены почти все позиции. Такие записи появляются в результате отказа респондента отвечать на вопросы переписного листа.

Необходимость такого деления матрицы X (1) на три матрицы (доноров, реципиентов и нетипичных объектов) объясняется тем, что, как показал мировой опыт ведущих стран в области обработки статистических данных (США, Канада, Нидерланды др.), распределения (функции плотности распределения) подавляющего большинства экономических количественных показателей обладают ярко выраженной правосторонней асимметрией (см. рисунок). Как восстановление, так и использование данных из «хвоста» распределения (на рисунке заштриховано) в качестве донорских значений сопряжено с возможностью получения больших ошибок (завышенных значений) при импутации данных. Для исключения подобных ситуаций и повышения доли корректно восстановленных значений целесообразно разделить массив, состоящий из записей, которые не содержат пропусков, на две части: доноры и данные с нетипичными значениями отдельных показателей.



Типичное распределение экономического показателя

Таким образом, $X^{(\partial)}_{m\,x\,n}$ - матрица доноров порядка $m \times n$, а $X^{(p)}_{r\,x\,n}$ - матрица реципиентов порядка $r \times n$, где $r << \mathrm{m} < \mathrm{k}$, и $r + \mathrm{m} < \mathrm{k}$.

Общая идея восстановления опирается на предположение, что для каждой строки из матрицы реципиентов $X^{(p)}_{r\,x\,n}$ может быть найден сосед из матрицы доноров $X^{(\partial)}_{m\,x\,n}$. Это значит, что должна быть определена мера близости или сходства между строками (1).

Предварительный статистический анализ близости наблюдаемых данных показал наличие определенных однородных классов в матрице $X^{(\partial)}_{m\,x\,n}$. Поэтому первоначально была выполнена кластеризация. Для этой цели в матрице $X^{(\partial)}_{m\,x\,n}$ были выделены экспертным путем показатели, позволяющие провести многомерную классификацию объектов. Число таких компонент оказалось равным 39 (из 500), что позволило существенно сократить размерность факторного пространства без значимой потери его информативности.

В качестве алгоритма разбиения на классы использовался метод «К - средних» [1]. Это позволило отказаться от построения матрицы расстояний. Статистический анализ результатов разбиения показал их значимость и статистическую устойчивость к выбору первоначальных центров кластеров. В результате все доноры были разделены на несколько классов (10 – 16). Количество классов выбиралось априори. Центры этих классов рассматривались в качестве типичного представителя, а их значения вычислялись по формуле среднего арифметического.

При импутации количественных данных в качестве основного метода восстановления пропущенных значений был использован «Метод ближайшего соседа». Он основан на двукратном поиске. Первоначально выбирался класс, центр которого наиболее близок к реципиенту: выполнялся поиск по 39 основным компонентам донора. Выбор класса означал, что в дальнейшем реципиента следует искать среди представителей только этого класса. Такой подход позволил не перебирать при поиске всех доноров. Заметим, что центр класса не обязательно удовлетворял балансовым ограничениям. Далее среди доноров в выбранном классе производился поиск наблюдения $X_{i_0}^{(\partial)}$, наиболее близкого к данному $X_i^{(p)}$ из матрицы $X^{(p)}$ реципиентов. Пропущенное поле (или набор полей) в векторе $X_i^{(p)}$ замещается числовыми значениями из вектора-донора $X_{i_0}^{(\partial)}$

$$\underbrace{\left(\mathrm{x}_{\mathrm{i}_0 1}^{(\mathrm{д})} \quad \mathrm{x}_{\mathrm{i}_0 2}^{(\mathrm{д})} \quad \ldots \quad \mathrm{x}_{\mathrm{i}_0 \mathrm{s}}^{(\mathrm{д})} \quad \ldots \quad \mathrm{x}_{\mathrm{i}_0 \mathrm{n}}^{(\mathrm{д})}\right)^{\mathrm{T}}}_{\mathrm{X}_{\mathrm{i}_0}^{(1)}} \rightarrow$$

$$\rightarrow \underbrace{\left(\mathrm{x}_{\mathrm{il}}^{(\mathrm{p})} \quad \mathrm{x}_{\mathrm{i2}}^{(\mathrm{p})} \quad \ldots \quad \mathrm{x}_{\mathrm{is}}^{(\mathrm{p})} \quad \ldots \quad \mathrm{x}_{\mathrm{in}}^{(\mathrm{p})}\right)^{\mathrm{T}}}_{\mathrm{X}_{\mathrm{i}}^{(2)}}.$$

Вектор-донор $X_{i_0}^{(\partial)}$ находится из условия

$$i_0 = \arg\min_{j \in \{1,\ldots m\}} \rho\left(X_i^{(p)}, X_j^{(\partial)}\right),$$

$$\rho(X,Y) = \sqrt{\sum_{s=1}^{n} (x_s - y_s)^2}. \qquad (4)$$

Для непосредственной реализации предложенного алгоритма необходимо выбрать метрику, позволяющую оценить меру близости. На данном этапе использовалась евклидова метрика. Несоизмеримость компонент вектора не позволяет применить непосредственно формулу расстояния (4). Поэтому первоначально проводилась нормировка компонент в матрице $X_{r\,x\,n}^{(p)}$ реципиентов:

$$\tilde{x}_{ij}^{(p)} = \begin{cases} \dfrac{x_{ij}^{(p)}}{x_{j,\max}^{(\partial)}}, & \text{если} \quad x_{j,\max}^{(\partial)} \neq 0 \\ 0, & \text{если} \quad x_{j,\max}^{(\partial)} = 0, \end{cases}$$

где $x_{j,\max}^{(\partial)}$ - максимальное значение j-го показателя в матрице доноров. Аналогично нормировались и величины в матрице доноров.

Из нормированной матрицы реципиентов $X_{r\,x\,n}^{(p)}$ последовательно выбирались векторы $\tilde{X}_i^{(p)}$, и для каждого вектора составлялся список импутируемых полей, содержащих пропущенные или ошибочные данные. Находился класс, ближайший к вектору $\tilde{X}_i^{(p)}$, а затем и вектор $\tilde{X}_{t_0}^{(\partial)}$ из класса, ближайший к вектору $\tilde{X}_i^{(p)}$, значения полей которого могут быть приписаны реципиенту с соблюдением всех правил редактирования его компонент. Если найден не единственный донор, то вычислялись полные нормы всех таких доноров; доноры ранжировались по возрастанию их норм, и из упорядоченного множества доноров выбирался донор с медианной нормой. В импутируемые поля вектора-реципиента $\tilde{X}_i^{(p)}$ записывались данные из одноименных полей вектора-донора $\tilde{X}_{t_0}^{(\partial)}$. В заключение осуществлялись восстановление абсолютных значений показателей в векторе $\tilde{X}_i^{(p)}$ по формуле $x_{ij}^{(p)} = \tilde{x}_{ij}^{(p)} \cdot x_{j,\max}^{(\partial)}$ и проверка правил редактирования по всем восстановленным значениям (полям), входящим в группы взаимосвязанных показателей.

В том случае, когда между некоторыми показателями существовали корреляционные связи, с заметными или высокими значениями показателя тесноты связи по шкале Чеддока, импутация проводилась по регрессионным уравнениям. Использовалась простейшая линейная модель регрессии:

$$y_i = a_0 + a_1 \cdot x_i + \varepsilon_i,$$

где ε_i – аддитивная случайная величина с математическим ожиданием, равным нулю, и конечной дисперсией. Коэффициенты a_0 и a_1 в уравнении регрессии оценивались методом наименьших квадратов:

$$\hat{a}_0 = \frac{\Delta_1}{\Delta_0};\ \hat{a}_1 = \frac{\Delta_2}{\Delta_0};\ \Delta_0 = m \cdot \left(\sum_{i=1}^{m} x_i^2\right) - \left(\sum_{i=1}^{m} x_i\right)^2;$$

$$\Delta_1 = \left(\sum_{i=1}^{m} y_i\right) \cdot \left(\sum_{i=1}^{m} x_i\right) - \left(\sum_{i=1}^{m} x_i\right) \cdot \left(\sum_{i=1}^{m} y_i \cdot x_i\right);$$

$$\Delta_2 = m \cdot \left(\sum_{i=1}^{m} y_i \cdot x_i\right) - \left(\sum_{i=1}^{m} y_i\right) \cdot \left(\sum_{i=1}^{m} x_i\right).$$

Восстановление значения $\hat{y}_i$ при известном значении x_i производится по формуле

$$\hat{y}_i = \hat{a}_0 + \hat{a}_1 \cdot x_i .$$

Значение восстановленного показателя $\hat{y}_i$ округлялось до ближайшего целого. Для повышения скорости импутации по регрессионным уравнениям целесообразно заранее построить оценки параметров всех уравнений регрессии по матрице доноров $X^{(\partial)}_{m\ x\ n}$ для каждой пары показателей, восстанавливаемых по регрессионным уравнениям.

Для получения логически непротиворечивых процедур автокоррекции и импутации параллельно с описанным алгоритмом разрабатывались правила редактирования. В этих правилах по каждому показателю переписного листа приведены описание возможных отклонений показателя от корректных состояний и описание предписываемых действий по исправлению некорректного состояния. Фрагмент таких правил по одному из 500 показателей приведен ниже.

Контроль ФЛК	Проверяемые условия	Действие в случае невыполнения	
		Автокоррекция	Импутация
Group7_18 >= Group7_19	Group7_18 = NULL And Group7_19 = NULL		Group7_18, Group7_19 методом «ближайшего соседа»
	Group7_18 > 0 And Group7_19 = NULL	Group7_19 = 0	
	Group7_19 > 0 And (Group7_18 = NULL Or Group7_18 < Group7_16)		Регрессия Group7_18 на Group7_19

Замечание. В случае равенства нулю величины Δ_0 (в пределах статистической совокупности величина x_i постоянна) коэффициенты в уравнении регрессии не могут быть вычислены. В этом случае восстановление показателя или группы показателей необходимо производить методом ближайшего соседа либо положить соответствующий показатель равным нулю.

Заключение. Всероссийская сельскохозяйственная перепись с автоматизированной обработкой результатов переписи проведена впервые в истории нашей страны. За короткий промежуток времени удалось создать работоспособный вариант национальной системы сбора и редактирования данных переписей. При разработке математического обеспечения автоматизированной системы учитывался опыт ведущих стран (США, Канада, Нидерланды и др.), имеющих длительную историю проведения переписей и обработки материалов переписей с помощью пакетов программ.

В мировой практике на разработку систем редактирования данных переписей и проведение процедур редактирования тратится до 40 % всех средств, выделяемых на проведение переписей. Временные ограничения не позволили исследовать эффективность некоторых алгоритмических решений и соответственно включить их в созданный вариант системы.

В будущем переписи в России планируется проводить с определенной периодичностью (один раз в 5 – 10 лет) и потому целесообразно продолжать исследования по созданию математического обеспечения процедур редактирования данных, а также систем генерации отчетов и визуализации итогов переписей.

Библиографический список

1. *Дюран Б., Оделл П.* Кластерный анализ: пер. с англ. – М., 1977.