Алгоритм ZET

Пусть задана таблица экспериментальных данных $A = (a_{ij})$, $i = \overline{1,m}$, $j = \overline{1,n}$ типа „объект-свойство” (табл. 1),

Таблица 1

Таблица экспериментальных данных

Объект\Свойство	1	2	…	x	…	n
1	a_{11}	a_{12}	…	a_{1x}	…	a_{1n}
2	a_{21}	a_{22}	…	a_{2x}	…	a_{2n}
…	…	…	…	…	…	…
y	a_{y1}	a_{y2}	…	@	…	a_{yn}
…	…	…	…	…	…	…
m	a_{m1}	a_{m2}	…	a_{mx}	…	a_{mn}

где m – количество строк-объектов, n – количество столбцов-свойств, $a_{yx} = @$ - пропуск, причем количество таких пропусков в таблице может быть довольно большим.

Ставится задача восстановления отсутствующих (пропущенных) значений @ в таблице A.

Рассмотрим решение поставленной задачи с помощью алгоритма ZET [1]. Данный метод относится к локальным методам заполнения пробелов, так как использует для нахождения решения только некоторую локальную часть экспериментальных данных.

В основе его функционирования лежат три предположения (гипотезы):

1. ***Гипотеза избыточности***: предполагается, что в таблице A присутствует избыточность в строках (объекты могут быть похожи между собой) и столбцах (между свойствами могут быть зависимости). При отсутствии избыточности все строки и столбцы имеют одинаковый вес при прогнозировании и смысл локальности алгоритма теряется.
2. ***Гипотеза аналогичности***: предполагается, что если два объекта «похожи» по значениям $(n-1)$ свойств, то они «похожи» и по n-му свойству.
3. ***Гипотеза локальной компетентности***: предполагается, что избыточность строк и столбцов носит локальный характер, то есть для каждого пропущенного значения имеется только некоторое количество объектов – аналогов объекта с пропуском и свойств – аналогов свойства с пропуском. Поэтому предлагается использовать для прогнозирования только такие «компетентные» объекты и свойства, которые выбираются для каждого пропуска отдельно.

Основные этапы алгоритма ZET для обработки таблицы A с l пропусками:

1. Предварительная обработка начальных данных.
2. Прогнозирование пропуска - выполняется l раз:
   2.1. Формирование компетентной матрицы.
   2.2. Подбор параметров модели прогнозирования.
   2.3. Прогнозирование пропуска.

Рассмотрим подробнее каждый этап.

1. Вначале столбцы матрицы A нормируются по дисперсиям для приведения различных свойств объектов к единой шкале:

$$a_{ij} = \frac{a_{ij} - \overline{a_j}}{G_j}.$$

2. Следующие этапы выполняют l раз. Пусть координаты текущего элемента с пропуском x, y

2.1. Формирование компетентной матрицы

2.1.1. Задать размеры компетентной матрицы s_{ij}, $i = \overline{1, p}$, $j = \overline{1, q}$, $2 < p < m$, $2 < q < n$.

2.1.2. Выбрать $(p-1)$ компетентных строк для строки с пропуском.
Компетентность L строки i по отношению к строке с пропуском y определяется по формуле

$$L_{iy} = \frac{t_{iy}}{r_{iy}},$$

где t_{iy} - комплектность, то есть число значений известных для обеих строк i и y, r_{iy} - декартово расстояние между строками (элементы с пропусками не учитываются). Компетентная строка не должна содержать пропуска на x-й позиции.

2.1.3. Выбрать $(q-1)$ компетентных столбцов для столбца с пропуском.
Компетентность L столбца i по отношению к столбцу с пропуском x определяется по формуле

$$L_{iy} = |k_{ix}| \cdot t_{ix},$$

где t_{ix} - комплектность столбцов i и x, k_{ix} - коэффициент корреляции между столбцами i и x. При расчете k_{ix} используются только те значения столбцов, которые принадлежат к компетентным строкам. Компетентный столбец не должен содержать пропуск на y-й позиции.

2.2. Подбор параметров моделей прогнозирования αr (по строкам) и αc (по столбцам) – коэффициенты регулирующие влияние компетентности на результат предсказания.

2.2.1. Задаем пределы изменения коэффициентов αr и αc и шаг их изменения.

2.2.2. Находим оптимальные коэффициенты αr и αc для прогноза пропуска по строкам и по столбцам по следующему алгоритму (одинаков для строк и столбцов). Подавая значения коэффициента α ($\alpha = \alpha r$ для строк, $\alpha = \alpha c$ для столбцов) в указанных пределах и с указанным шагом минимизируем функцию

$$\sum_i | a_{ik} - b_{ik} | \to \min,\ i \neq @,$$

где a_{ik} - реальное значение элемента i строки (столбца) k с пропуском, b_{ik} - прогноз этого элемента с помощью компетентных строк (столбцов). b_{ik} рассчитываются по формуле

$$b_{ik} = \frac{\sum_{j=1}^{c-1} bl_{jk} \cdot L_{ij}^{\alpha}}{\sum_{j=1}^{c-1} L_{ij}^{\alpha}},$$

где $c = p$ для строк и $c = q$ для столбцов, bl_{jk} - прогноз для известных значений строки (столбца) с пропуском k с помощью i-й строки (столбца), рассчитывается с помощью линейной регрессии вида $y = ax + b$ по МНК.

2.3. Прогнозирование пропуска

2.3.1. Прогнозирование пропуска по столбцам выполняется по формуле

$$b_x = \frac{\sum_{i=1}^{q-1} bl_{ix} \cdot L_{ix}^{\alpha c}}{\sum_{i=1}^{q-1} L_{ix}^{\alpha c}}.$$

2.3.2. Прогнозирование пропуска по столбцам выполняется по формуле

$$b_y = \frac{\sum_{i=1}^{p-1} bl_{iy} \cdot L_{iy}^{\alpha r}}{\sum_{i=1}^{p-1} L_{iy}^{\alpha r}}.$$

2.3.3. Общий прогноз получается усреднением прогнозов по строкам и столбцам

$$b_{yx} = \frac{b_y + b_x}{2}.$$

Программы заполнения пробелов могут работать в одном из следующих режимов:

1. Заполнение всех пробелов в таблице по указанному алгоритму.
2. Заполнение только тех пробелов, ожидаемая ошибка для которых не превышает заданной величины. Для определения ожидаемой ошибки предсказания вычисляется дисперсия значений подсказок bl_{ij}, получаемых от всех q столбцов и p строк компетентной подматрицы.
3. Заполнение пробелов только на базе информации, имеющейся в исходной таблице.
4. Заполнение каждого следующего пробела с использованием исходной информации и прогнозных значений ранее заполненных пробелов.

Список использованных источников

1. Загоруйко Н.Г. Методы распознавания и их применение. – М.: Советское Радио, 1972.
2. http://math.nsc.ru/AP/oteks/index.html.