Д И С К Р Е Т Н Ы Й А Н А Л И З

Сборник трудов

1970 г. Института математики СО АН СССР Выпуск 17

УДК 519.95

ИСПОЛЬЗОВАНИЕ ДЛИН ТУПИКОВЫХ ТЕСТОВ ПРИ ОБРАБОТКЕ ТАБЛИЦ

А.Н. Дмитриев

Цель настоящей заметки – дополнить тестовый метод [2] обработки таблиц бинарных символов учетом длин тупиковых тестов. Таблицы содержат информацию о реальных объектах и составлены из нулей и единиц, указывающих на отсутствие или наличие признаков, характеризующих эти объекты. Обработка таблиц употребляется при распознавании образов, классификации, упорядочивании объектов по заданному критерию и в других подобных задачах [3–5]. Длины тупиковых тестов (как было обнаружено в конкретных практических задачах) позволяют строить новые величины, уточняющие и дополняющие решение задач указанного профиля. Содержание заметки состоит из небольшого введения, числовых характеристик таблиц, строк, столбцов, связанных с длинами тупиковых тестов и некоторых рекомендаций по использованию введенных величин.

1. Напомним, что тестом для таблицы T называется такая ее часть t, которая получается из T удалением тех или иных столбцов, и в которой в с е с т р о к и р а з л и ч н ы; тупиковым называется такой тест, в котором никакая его часть не является тестом [1]; длиной $\ell(t)$ теста t называется число содержащихся в нем столбцов.

Например, в таблице

$$T = \begin{bmatrix} 000\ldots00 \\ 100\ldots01 \\ 010\ldots10 \\ 001\ldots11 \end{bmatrix}$$
$$\underbrace{\qquad}_{t_1} \quad \underbrace{\qquad}_{t_2}$$

ее части t_1 и t_2 образуют тупиковые тесты, причем $\ell(t_1) = 3$, а $\ell(t_2) = 2$. В силу того, что в тупиковом тесте в с е строки различны, его можно интерпретировать как носителя распознавания строк таблицы, а в силу несжимаемости такого теста его можно рассматривать как первичный элементарный носитель распознавания. Для данной таблицы T и теста t разность $\ell(T) - \ell(t)$ можно принимать за меру его выразительности, то есть чем больше $\ell(T) - \ell(t)$, тем больше "экономичность" распознавания строк T с помощью t (по сравнению с другими тестами той же таблицы). Эти соображения составляют основу построений во втором пункте.

2. Пусть $T = \{a_{ij}\}$ – бинарная (т.е. $a_{ij} = 0, 1$) таблица из m строк ($i = 1, 2, \ldots, m$) и n-столбцов ($j = 1, 2, \ldots, n$). В задачах указанного типа о с н о в н у ю роль играют следующие величины, разработанные на тестовой основе [2,3]:

а) $P_j = \frac{\kappa_j}{\kappa}$ – информационный вес j-го столбца, где

κ – общее число тупиковых тестов; κ_j – нагрузка j-го столбца, или число вхождений j-го столбца в тупиковые тесты t таблицы T.

б) $P(\sigma) = \sum_{i=1}^{n} P_i \sigma_i$ – информационный вес строки $\sigma = (\sigma_1, \ldots, \sigma_n)$ таблицы T.

Как видно из определения для P_j, информация о длинах тупиковых тестов t в операции вычисления P_j не участвует, поскольку все t в T с ч и т а ю т с я р а в н о п р а в н ы м и. Указанная выше основа (п. а) и п. б)) сохраняется, но в дополнение к ней теперь предлагается некоторое р а с ч л е н е н и е информационного веса столбцов, учитывающее длины тупиковых тестов, а именно: для j-го столбца таблицы T ($j = 1, \ldots, n$) определим

$$R_j = P_j\left(1 - \frac{\ell_j}{m}\right), \quad Q_j = P_j \cdot \frac{\ell_j}{m},$$

где $\bar{l}_i$ – средняя длина тупикового теста, содержащего i-й столбец, а m – число строк. Как и в п. б), для строки $\sigma=(\sigma_1,\dots,\sigma_n)$ определим числа:

$$R(\sigma)=\sum_{i=1}^{n} R_i\sigma_i \qquad \text{и} \qquad O(\sigma)=\sum_{i=1}^{n} O_i\sigma_i .$$

Числа R_i, R и O_i, O назовем соответственно р а з л и ч а ю щ и м и и о т о ж д е с т в л я ю щ и м и весами столбца или строки. По определению $R_i+O_i=P_i$, $R(\sigma)+O(\sigma)=P(\sigma)$.

В связи с этим расчленением и практическими нуждами возникает потребность в изучении р а с п р е д е л е н и я д л и н тупиковых тестов и в отыскании класса таблиц с заданным распределением тестов – таблиц, у которых преимущественно "короткие" тесты и т.п.; большое прикладное значение имеет сравнение максимальной длины тупикового теста таблицы со средней длиной тупиковых тестов той же таблицы. Отметим попутно, что последняя равна $\sum_{i=1}^{n} P_i$.

3. Прикладное значение величин различающих и отождествляющих информационных весов состоит в том, что с их помощью более тонко ф р а к ц и о н и р у ю т с я н а г р у п п ы столбцы и строки таблицы T.

Употребляя R_i и O_i при анализе столбцов, интерпретатор из общей совокупности признаков может выделить группу признаков, которая либо отождествляет, либо различает объекты в T. Это подразделение минимизирует число признаков в соответствии с целями обработки таблиц и обнаруживает качественный состав признаков.

Употребление величин R и O для фракционирования строк приводит к подразделению такого вида:

а) Оценка строк по величине R предпочтительна для процедуры у п о р я д о ч и в а н и я объектов по заданному критерию внутри данной таблицы.

б) Оценка строк по величине O предпочтительна для процедур обнаружения родственных групп исследуемых объектов одной или нескольких таблиц, то есть в задачах компоновки классов.

Данные, получаемые с помощью указанных величин, позволяют естествоиспытателю существенно корректировать этапы решения задач, связанных с обработкой нескольких таблиц, содержательно приближая их к цели исследования.

Л и т е р а т у р а

I. И.А. ЧЕГИС, С.В. ЯБЛОНСКИЙ. Логические способы контроля электрических схем. – Труды Математического института им. В.А. Стеклова, I958, т. 5I, стр. 270-360.

2. А.Н. ДМИТРИЕВ, Ю.И. ЖУРАВЛЕВ, Ф.П. КРЕНДЕЛЕВ. О математических принципах классификации предметов и явлений. – Дискретный анализ, Новосибирск, I966, вып. 7, стр. 3-I5.

3. А.Н. ДМИТРИЕВ. Некоторые табличные числа. – Дискретный анализ, Новосибирск, I968, вып. I2, стр. 22-26.

4. Сб. "Алгоритмы и программы вычислительной диагностики психических заболеваний". Новосибирск, I969, стр. 5I-76.

5. Сб. "Алгоритмы и программы решения геологических задач на ЭЦВМ "Минск-2" и "БЭСМ-3М", вып. 2, Алма-Ата, I969, стр.30-53.

Поступила в редакцию
23.9.I970 г.