ЭМПИРИЧЕСКОЕ ПРЕДСКАЗАНИЕ И РАСПОЗНАВАНИЕ ОБРАЗОВ
(Вычислительные системы)
1976 год Выпуск 67

УДК 681.3.06:621.391

МЕТОД ОБНАРУЖЕНИЯ ЛОГИЧЕСКИХ ЗАКОНОМЕРНОСТЕЙ НА ЭМПИРИЧЕСКИХ ТАБЛИЦАХ

Г.С.Лбов, В.И.Котиков, Ю.П.Машаров

В данной работе рассматривается задача поиска определенного класса закономерностей на основе анализа эмпирических таблиц. Под эмпирической таблицей будем понимать таблицу $\{ x_{ij} \}$ размером $(N \cdot n)$, элементы которой есть результаты замеров признаков $X = \{ X_1, \ldots, X_j, \ldots, X_n \}$ для каждого из N объектов. Каждая реализация (строка таблицы) есть элемент пространства событий

$$D = D_1 \cdot D_2 \cdot \ldots \cdot D_j \cdot \ldots \cdot D_n ,$$

где D_j – область значений признака X_j $(j = 1, \ldots, n)$.

Метод поиска закономерностей на таблицах рассматривается для решения следующих типов задач:

а) распознавания образов;

б) аппроксимации эмпирических таблиц.

Указанные задачи решаются с одновременным выбором наиболее информативной подсистемы признаков из X.

Необходимость разработки рассматриваемого ниже метода была вызвана существованием целого класса эмпирических таблиц, характерной особенностью которых является наличие признаков, замеренных в шкалах разных типов *). Методы анализа таблиц должны учитывать типы шкал, в которых замерены признаки, и быть инвариантными по отношению к допустимым преобразованиям этих шкал [1].

*) Признаки, замеренные в шкалах интервалов и отношений, часто называют количественными; признаки, замеренные в шкале порядка, – качественными, ранговыми; признаки, замеренные в шкале наименований, – классификационными, номинальными и т.д.

Эмпирические таблицы с разнотипными признаками встречаются, как правило, при решении задач из области социологии, экономики, медицины, геологии. Кроме того, решение задач из указанных областей усложняется следующими особенностями:

1) признаки взаимозависимы;

2) количество признаков велико;

3) объем выборки мал;

4) таблицы имеют "пропуски" (ряд признаков у некоторых объектов не замерен).

Существующие методы анализа эмпирических таблиц с разнотипными признаками предполагают предварительное сведение всех типов признаков к одному типу, что связано с искажением информации. В связи с этим были сформулированы новые подходы к анализу рассматриваемого класса эмпирических таблиц [2-4].

Введем необходимые определения. Под логическим высказыванием $S(X')$ на множестве значений некоторых признаков $X' \subset X$ будем понимать конъюнкцию следующего вида:

$$S(X') = \bigwedge_{j=1}^{m} I_j,$$

где m – число признаков в подмножестве X' ($m \le n$); величина I_j есть либо какое-то конкретное значение $X_j \in X'$, если X_j – признак, замеренный в шкале наименований, либо некоторый интервал значений X_j, если X_j – признак, замеренный в шкале отношений, интервалов и порядка. В случае шкалы порядка под интервалом понимаем объединение соседних значений признака.

Будем говорить, что логическое высказывание $S(X')$ выполняется на некоторой i-й реализации (объекте) $x_i=(x_{i1},\dots,x_{in})$, если имеют место либо $x_{ij} = I_j$ (I_j – значение), либо $x_{ij} \in I_j$ (I_j – интервал) для каждого $X_j \in X'$. Логическое высказывание $S(X')$ выполняется на некоторой подтаблице исходной таблицы, если оно выполняется на всех реализациях, входящих в эту подтаблицу.

Будем предполагать, что на рассматриваемом множестве признаков существуют логические высказывания, удовлетворяющие определенному критерию информативности. Такие высказывания будем называть закономерностями.

Приведем пример из области распознавания. Логическое высказывание "Рабочий имеет дело с определенного вида инструментом, стаж его работы более пяти лет, он имеет профессию клепальщика, артериальное давление в пределах I50-I80" может выступать в качестве закономерности. Критерием информативности в этом случае, например, можно выбрать вероятность того, что рабочий, для которого выполняется это высказывание, болен определенной болезнью. Указанное высказывание будет закономерностью в том случае, если значение указанной вероятности больше некоторого порога.

Заметим, что закономерности такого вида обладают важным, на наш взгляд, свойством. Они достаточно просты для интерпретации.

Формально на множестве значений признаков из X может быть сформулировано практически неограниченное количество логических высказываний. Нам необходимо при анализе эмпирической таблицы выбрать те из высказываний, которые являются закономерностями.

При решении задачи поиска логических закономерностей возникают вопросы обоснования критерия информативности и разработки алгоритма поиска закономерностей.

В данной статье изложен алгоритм, позволяющий обнаружить характерные для таблицы закономерности в виде логических высказываний. Кроме того, в работе предлагается критерий информативности для решения задач аппроксимация таблиц.

Алгоритмическая сторона метода одинакова для задачи распознавания образов и для задачи аппроксимации таблиц.

§I. Задача распознавания образов

Рассмотрим поиск закономерностей для некоторого образа Q. Для каждого логического высказывания S можно определить две величины n_S и m_S (n_S – число реализаций образа Q, а m_S – число реализаций остальных образов, на которых выполняется высказывание S). Закономерностью образа Q будем считать логическое высказывание, для которого выполнено следующее условие:

$$\left.\begin{aligned} n_S &\ge \delta, \\ m_S &\le \beta. \end{aligned}\right\} \qquad (1)$$

Для простоты изложения сначала опишем алгоритм поиска закономерностей для признаков, замеренных в шкале наименований.

На первом этапе рассматриваются всевозможные элементарные высказывания (т.е. все значения признаков). Для каждого такого высказывания проверяется выполнение условия $n_S \geq \delta$. Если условие не выполняется, то данное высказывание исключается из дальнейшего перебора. В противном случае проверяется выполнение условия $m_S \leq \beta$. Если это условие выполняется, то высказывание считается закономерностью, выдается на печать и исключается из дальнейшего перебора. Иначе высказывание некоторым образом отмечается и сохраняется в памяти ЭВМ.

На втором этапе рассматриваются всевозможные парные высказывания. Эти высказывания являются конъюнкциями отмеченных на первом этапе значений различных признаков. Если для какого-либо парного высказывания S $n_S \geq \delta$ и $m_S \leq \beta$, то данное высказывание считается закономерностью и выдается на печать; если $n_S \geq \delta$ и $m_S > \beta$, то оба элементарных высказывания отмечаются и сохраняются в памяти ЭВМ. Если элементарное высказывание не вошло ни в одно парное высказывание, для которого $n_S \geq \delta$, то оно исключается из дальнейшего рассмотрения.

Далее рассматриваются всевозможные тройные высказывания с аналогичной проверкой указанных условий и т.д. Такой поэтапный процесс формирования логических высказываний продолжается до тех пор, пока в памяти не останутся отмеченные элементарные высказывания.

Легко видеть, что предлагаемый алгоритм перебора высказываний обнаруживает все закономерности. Действительно, для любых высказываний S_1 и $S_2 (S_1 \neq S_2)$ из того, что $n_{S_1} < \delta$ и (или) $n_{S_2} < \delta$, следует $n_S < \delta$, где $S = S_1 \wedge S_2$. Поэтому, не рассматривая при формировании на очередном этапе всех тех высказываний $S = S_i \wedge S_j$, для которых $n_{S_i} < \delta$ и (или) $n_{S_j} < \delta$, мы сокращаем перебор, но не исключаем ни одно из высказываний, из которых могут сформироваться закономерности.

Время, требуемое для выбора закономерностей, зависит от информативности признаков и величин δ и β. Память для хранения промежуточных результатов пропорциональна числу элементарных высказываний, сформированных на первом этапе.

Для признаков, замеренных в шкалах отношений, интервалов и порядка, в множество $\{I_j\}$ включаются значения и объединения

Таблица I

Реализация	Признак				
	I	2	3	4	5
1	2	5	0,25	0	0,9
2	2	5	0,3	1	0,6
3	2	5	0,1	0	0,4
4	2	5	0,37	1	0,2
5	2	4	0,75	0	0,9
6	2	4	0,8	1	0,6
7	2	4	0,9	0	0,4
8	2	4	0,65	1	0,2
9	3	4	0,31	0	0,9
10	3	4	0,37	1	0,6
11	3	4	0,23	0	0,4
12	3	4	0,15	1	0,2
13	3	5	0,81	0	0,9
14	3	5	0,9	1	0,6
15	3	5	0,7	0	0,4
16	3	5	0,62	1	0,2
17	2	5	0,85	0	0,9
18	2	5	0,93	1	0,6
19	2	5	0,72	0	0,4
20	2	5	0,65	1	0,2
21	2	4	0,34	0	0,9
22	2	4	0,02	1	0,6
23	2	4	0,23	0	0,4
24	2	4	0,15	1	0,2
25	3	4	0,68	0	0,9
26	3	4	0,95	1	0,6
27	3	4	0,83	0	0,4
28	3	4	0,72	1	0,2
29	3	5	0,3	0	0,9
30	3	5	0,2	1	0,6
31	3	5	0,15	0	0,4
32	3	5	0,02	1	0,2

Образ А

Образ В

соседних значений (если признак X_j замерен в шкале порядка) и интервалы значений (если X_j замерен в шкалах интервалов и отношений). Во втором случае рассматривается интервал $[x_j^Q{}_{min}, x_j^Q{}_{max}]$, где $x_j^Q{}_{min}$ и $x_j^Q{}_{max}$ – минимальное и максимальное значения из множества значений признака X_j для объектов образа Q. Этот интервал разбивается на ряд подынтервалов, каждый из которых содержит не меньше чем $[\varepsilon_1 \cdot \delta]$ значений признака X_j для объектов образа Q ($\frac{1}{\delta} \leq \varepsilon_1 \leq 1$). В качестве множества элементарных высказываний $\{I_j\}$ для признака X_j выбираем интервалы, содержащие не меньше чем δ реализаций признака X_j для объектов образа Q и являющиеся всевозможными объединениями указанных подынтервалов. При этом не рассматриваем те интервалы, на которых число значений признака X_j для объектов остальных образов $m_S \geq \varepsilon_2 \cdot N_{\tilde{Q}}$, где $N_{\tilde{Q}}$ – число объектов всех образов за исключением образа Q ($\frac{1}{N_{\tilde{Q}}} \leq \varepsilon_2 \leq 1$). Выбор величин ε_1 и ε_2 определяет число рассматриваемых элементарных высказываний. При $\varepsilon_1 = \frac{1}{\delta}$ и $\varepsilon_2 = 1$ алгоритм

обнаруживает все закономерности для образа Q ,но при $\varepsilon_1 \to \frac{1}{\delta}$ и $\varepsilon_2 \to 1$ число рассматриваемых высказываний резко возрастает.В конкретной реализации данного алгоритма выбраны $\varepsilon_1 = \frac{1}{2}$ и $\varepsilon_2 = \frac{1}{3}$.В остальном схема алгоритма совпадает со схемой,изложенной для случая признаков, замеренных в шкале наименований.

На основе полученных закономерностей можно сформулировать решающее правило для распознавания образов.Для этого выбирается из полученных для каждого образа закономерностей такое их подмножество, которое выполняется на возможно большем числе объектов рассматриваемого образа Q. При этом число объектов, на которых одновременно выполняется несколько закономерностей, должно быть минимально. Выбирая такое подмножество закономерностей, мы стремимся к тому, чтобы любая контрольная реализация, принадлежащая образу Q , удовлетворяла одной и только одной закономерности из этого подмножества (требование I).

Из разных возможных таких подмножеств выбирается подмножество, которое содержит минимальное число закономерностей (требование 2). Последнее требование теоретически обосновывается в §3. В результате получаем для всех образов набор подмножеств, включающих некоторый список закономерностей $\{ S_1, \dots, S_t, \dots, S_M \}$. В общем случае не всегда удается удовлетворить требованию I,и поэтому на реализации могут выполниться несколько закономерностей.

Сформулируем решающее правило. Делаем проверку выполнения на данной контрольной реализации всех закономерностей из списка. Определяем количество объектов, на которых выполнились эти закономерности из разных образов. Реализацию относим к тому образу, количество объектов которого оказалось наибольшим.

В качестве модельного примера была рассмотрена таблица I (пять признаков, тридцать две реализации).

Признаки X_3 и X_5 замерены в шкале отношений,остальные - в шкале наименований. В таблице I содержатся следующие закономерности:

$$\left.\begin{aligned} S_1 &= (X_1=2) \wedge (X_2=4) \wedge (0.95 \geq X_3 \geq 0.62) \\ S_2 &= (X_1=2) \wedge (X_2=5) \wedge (0.02 \leq X_3 \leq 0.37) \\ S_3 &= (X_1=3) \wedge (X_2=4) \wedge (0.02 \leq X_3 \leq 0.37) \\ S_4 &= (X_1=3) \wedge (X_2=5) \wedge (0.95 \geq X_3 \geq 0.62) \end{aligned}\right\} \text{Образ A}$$

$$\left.\begin{aligned} S_5 &= (X_1=2) \wedge (X_2=4) \wedge (0.02 \leq X_3 \leq 0.37) \\ S_6 &= (X_1=2) \wedge (X_2=5) \wedge (0.95 \geq X_3 \geq 0.62) \\ S_7 &= (X_1=3) \wedge (X_2=4) \wedge (0.95 \geq X_3 \geq 0.62) \\ S_8 &= (X_1=3) \wedge (X_2=5) \wedge (0.02 \leq X_3 \leq 0.37) \end{aligned}\right\} \text{Образ B}$$

Все закономерности удовлетворяют требованиям $\delta \geq 4$, $\beta=0$. Других закономерностей,удовлетворяющих этим требованиям,в таблице I нет. Предлагаемый алгоритм обнаружил все указанные закономерности.

ЗАМЕЧАНИЕ. При решении прикладных задач нами было замечено, что, как правило,максимальное число элементарных высказываний, входящих в закономерность, $m \leq 7$ и число выбранных закономерностей не превышает пяти для каждого образа.

§2. Задача аппроксимации эмпирических таблиц

Под задачей аппроксимации таблицы понимается установление соответствия между данной эмпирической таблицей и набором из небольшого числа логических закономерностей, каждая из которых выполняется на некоторой подтаблице.

Считается, что закономерные связи между объектами и признаками отсутствуют, если исходная таблица данных получена с помощью случайного механизма, т.е. ее можно рассматривать как выборку из пространства D с равномерным законом распределения.

В этом случае для любого логического высказывания S можно определить вероятность его выполнения P_S на такой "случайной" таблице. Вероятность выполнения высказывания S на N_S объектах из общего числа N равна

$$\mathcal{P}(N_S) = C_N^{N_S} P_S^{N_S} (1 - P_S)^{N - N_S}$$

Выбор критерия предпочтения одного логического высказывания другому при поиске закономерностей для решения указанной задачи аппроксимации таблицы основан на следующей гипотезе: чем меньше величина $\mathcal{P}(N_S)$ для высказывания S , тем больше оснований рассматривать это высказывание в качестве закономерности.

Число реализаций N_S, на которых выполняется высказывание S на "случайной" таблице, в среднем равно $N \cdot P_S$. Для целей аппроксимации таблицы будем рассматривать только те высказывания, для которых $N_S > N \cdot P_S$.

Кроме того, нас интересуют высказывания, выполняющиеся на исходной таблице не менее чем δ раз. Таким образом, высказывания должны удовлетворять условиям:

$$\left.\begin{aligned} N_S &> N \cdot P_S, \\ N_S &\geq \delta. \end{aligned}\right\} \qquad (2)$$

Для простоты вычислений при определении порядка предпочтения на множестве высказываний будем использовать величину

$$\gamma(S) = -\frac{(N_S - N \cdot P_S)^2}{P_S(1 - P_S)},$$

которая получается при аппроксимации биномиального распределения $\mathcal{P}(N_S)$ нормальным приближением

$$f(N_S) = \frac{1}{\sqrt{2\pi N P_S(1 - P_S)}} e^{\frac{\gamma(S)}{2N}}.$$

Ясно, что, чем меньше величина $\mathcal{P}(N_S)$, тем меньше и величина $\gamma(S)$.

Рассмотрим высказывания S_1 и S_2, выполняющиеся на данной таблице N_{S_1} и N_{S_2} раз, соответственно. Будем считать, что высказывание S_1 предпочтительнее S_2, если $\gamma(S_1) < \gamma(S_2)$.

Алгоритм формирования высказываний аналогичен алгоритму, изложенному в §I. Отличие состоит в том, что при формировании высказываний вместо условий (I) проверяются условия (2). В качестве закономерностей принимаются те из высказываний $\{S\}$, удовлетворяющих условию (2), из которых на следующем этапе алгоритма уже не могут быть сформулированы высказывания с меньшей, чем у них, величиной γ.

Для модельного примера была предложена таблица 2.

Признаки X_1 и X_8 замерены в шкале интервалов, остальные – в шкале наименований. В таблице 2 содержатся следующие закономерности, удовлетворяющие условию $\delta \geq 10$:

$S_1 = (67 \leq X_1 \leq 91)$,

$S_2 = (X_2 = 2)$,

$S_3 = (X_4 = 11) \wedge (0.13 \leq X_8 \leq 0.42)$,

$S_4 = (X_7 = 1) \wedge (0.13 \leq X_8 \leq 0.42) \wedge (X_9 = 1)$,

$S_5 = (X_3 = 4) \wedge (X_4 = 11) \wedge (X_5 = 1) \wedge (X_6 = 9)$.

Таблица 2

Реализация	Признак									
	I	2	3	4	5	6	7	8	9	I0
1	67	2	4	11	1	9	0	0,2	1	0
2	79	2	4	11	0	5	0	0,37	0	1
3	84	2	4	11	1	9	0	0,81	0	0
4	69	2	4	11	0	5	1	0,28	1	1
5	91	2	4	11	1	9	1	0,15	1	0
6	73	2	4	11	0	5	1	0,13	1	1
7	82	2	4	11	1	9	0	0,61	0	0
8	43	3	4	11	1	9	0	0,78	1	1
9	22	3	4	11	1	9	0	0,82	0	0
10	38	2	4	11	1	9	1	0,18	1	1
11	74	3	4	11	1	9	1	0,29	1	0
12	86	3	4	11	1	9	1	0,33	1	1
13	36	3	4	11	1	9	1	0,42	1	0
14	30	3	10	17	0	5	1	0,17	1	1
15	15	2	10	17	0	5	1	0,25	1	0
16	68	3	10	17	0	9	1	0,36	1	1
17	27	2	10	11	0	5	1	0,40	0	0
18	33	2	4	17	0	5	1	0,63	0	1

Других закономерностей, удовлетворяющих условию $\delta \geq 10$, в таблице 2 нет. Предлагаемый алгоритм обнаружил все указанные закономерности.

ЗАМЕЧАНИЯ. I. Если из выбранного множества закономерностей (различных высказываний с минимальными значениями критерия γ) удается выделить такое подмножество, чтобы ему соответствовал набор подтаблиц непересекающихся и полностью покрывающих исходную таблицу, то тогда можно решать задачу аппроксимации таблицы, сформулированную в работе [6]. В указанной работе эта задача решается только для случая, когда таксономия объектов

осуществляется для заранее выбранных подсистем признаков. Сняв указанное ограничение, можно улучшить качество аппроксимации таблицы с точки зрения критерия, используемого в [6].

Кроме того, если потребовать, чтобы закономерности выделяли подтаблицы, которые не имели бы общих строк, то эти подтаблицы представляли бы собой таксоны объектов. В общем случае может оказаться, что каждый из этих таксонов будет задан на некоторой подсистеме признаков.

2. Найденные закономерности могут быть использованы для заполнения пропусков в эмпирических таблицах [7].

3. Описанный метод обнаружения закономерностей можно использовать также для предсказания значения признаков, замеренных в шкалах отношений, интервалов и порядка. Значение целевого признака определяется с точностью до некоторого интервала предсказания. В этом случае в качестве критерия информативности рассматриваем величину интервала предсказания l на целевом признаке (в случае признака, замеренного в шкале порядка, интервалом является объединение соседних значений; величина интервала l – это количество значений, вошедших в объединение).

Необходимо минимизировать величину интервала предсказания l таким образом, чтобы существовало хотя бы одно высказывание S, удовлетворяющее условию

$$\left.\begin{aligned} n_{ls} &\geq \delta, \\ m_{ls} &\leq \beta, \end{aligned}\right\}$$

где n_{ls} (m_{ls}) – количество реализаций, удовлетворяющих высказыванию S и имеющих значение целевого признака в интервале l (вне интервала l).

4. Легко видеть, что предлагаемый метод поиска закономерностей как для цели распознавания, так и для цели аппроксимации является инвариантным по отношению к допустимым преобразованиям шкал. То есть данный метод для любых двух таблиц, описывающих один и тот же экспериментальный материал, но представленных в разной числовой форме, обнаруживает одни и те же закономерности, видоизмененные из-за разного представления таблиц.

Программы, реализующие описанный в данной работе метод обнаружения закономерностей, составлены на языке "Фортран-4".

§3. Об ε-оптимальном решающем правиле

При формировании решающего правила для распознавания возникает задача определения связи между объемом обучающей выборки и сложностью решающего правила.

Пусть получено множество закономерностей $B = \{S_1, \ldots, S_t, \ldots, S_M\}$, удовлетворяющее требованию I (см. §1). При этом если заданы распределения $\{P(\omega,x)\}$ ($\omega = 1,\ldots,k$) на множестве признаков X (ω – номер образа, x – точка n-мерного пространства признаков), то для произвольного множества B определены вероятности $P_{\omega t} = P(\omega) \cdot P(S_t/\omega)$, $\omega = 1,\ldots,k$; $t = 1,\ldots,M$, где $P(\omega)$ – априорная вероятность образа ω, $P(S_t/\omega)$ – вероятность того, что реализация образа ω удовлетворяет S_t. Можно говорить, что набор фиксированных вероятностей $C = \{P_{\omega t}\}$ задает стратегию природы ($C \in E$). Приписывая решение о принадлежности каждой закономерности тому или иному образу, мы тем самым задаем решающее правило из множества $R = \{r\}$ возможных правил, число которых равно k^M (k – число образов, M – число закономерностей).

Известно, что оптимальное решающее правило r_0, минимизирующее вероятность ошибки распознавания, формулируется так: $r_0(S_t) = \omega'$, если $P_{\omega' t} = \max\limits_{\omega = 1,\ldots,k} P_{\omega t}$ (реализацию относим к образу ω', если на ней выполняется закономерность S_t). Обозначим вероятность ошибки при r_0 через $\mathcal{P}_0$.

Так как распределения $\{P(\omega,x)\}$ неизвестны, решающее правило строится на выборке. Показано [5], что наилучшим в некотором смысле решающим правилом будет правило $r(S_t) = \omega'$, если $n_{\omega' t} = \max\limits_{\omega = 1,\ldots,k} n_{\omega t}$ ($n_{\omega t}$ – число реализаций образа ω, удовлетворяющих S_t).

Обозначим вероятность ошибки при выборочном решающем правиле r через $\mathcal{P}$. Ясно, что величина $\mathcal{P}$ будет случайной величиной с некоторым распределением $\varphi(\mathcal{P})$, причем $\mathcal{P} \geq \mathcal{P}_0$. Распределение $\varphi(\mathcal{P})$ и значение величины $\mathcal{P}_0$ зависят от стратегии природы.

Нас интересует следующий вопрос: при каком объеме обучающей выборки N (для фиксированного числа M) можно построить с большой достоверностью γ решающее правило, отличающееся от

оптимального не больше чем на ε в смысле вероятности ошибки независимо от стратегии природы? Другими словами, необходимо определить объем выборки N, при котором $P(\mathcal{P}-\mathcal{P}_0 \le \varepsilon) \ge \gamma$. Такое решающее правило назовем ε-оптимальным правилом.

Это условие должно выполняться для всех стратегий природы из E. При этом множество всевозможных стратегий E должно удовлетворять следующему ограничению: для каждой закономерности хотя бы одна из вероятностей $P_{\omega t}$ ($\omega = 1,\dots,K$) должна быть больше нуля, иначе число закономерностей будет меньше M.

Итак, задача состоит в том, чтобы связать между собой величины N, M, ε, γ. Ниже приводятся результаты моделирования для этой задачи при K = 2, $P(1) = P(2) = \frac{1}{2}$, вероятности $P(s_t/\omega)$ выбираются из множества {0; 0,1; 0,2; 0,3; 0,4; 0,5}. Значения величины ε даются в зависимости от объема выборки N и от числа M при $\gamma = 0{,}9$ (из-за того, что значения $P_{\omega t}$ задаются с точностью до 0,1, ε также принимает значения с точностью до 0,1). Значения ε приводятся в таблице 3.

Таблица 3

Объем выборки N	Число закономерностей M					
	2	3	4	5	6	7
10	0,2	0,2	0,2	0,2	0,2	0,2
20	0,1	0,1	0,2	0,2	0,2	0,2
30	0,1	0,1	0,1	0,1	0,1	0,1

Хотя при увеличении M величина ε в первой и третьей строках таблицы 3 остается постоянной из-за точности её определения, однако величина γ уменьшается. Для иллюстрации тенденции уменьшения γ с увеличением числа M приводим таблицу 4 значений величины γ в зависимости от N, M при $\varepsilon = 0{,}1$.

Таблица 4

Объем выборки N	Число закономерностей M					
	2	3	4	5	6	7
10	0,856	0,845	0,787	0,787	0,723	0,682
30	0,999	0,969	0,958	0,954	0,937	0,928

Решение задачи о связи между величинами N, M, ε, γ представляет практический интерес. Например, при поиске наилучшего множества B не имеет смысла перебирать множества с числом закономерностей, большим, чем некоторое число M_0, поскольку при заданном объеме выборки N и достоверности γ значение ε оказывается близким к единице.

Литература

1. СУППЕС П., ЗИНЕС Дж. Основы теории измерений. – В сб.: Психологические измерения. М., 1967, с. 1–89.

2. MICHALSKI R.S. A variable decision Spase approach for implementing a classification System. – In: Second International joint conference on pattern recognition, August 13–15, 1974, Copenhagen, Denmark, p.1–6.

3. ЛБОВ Г.С., КОТЮКОВ В.И., МАНОХИН А.Н. Об одном алгоритме распознавания в пространстве разнотипных признаков. –В кн.: Вычислительные системы. Вып. 55. Новосибирск, 1973, с. 108–110.

4. ЛБОВ Г.С., МАНОХИН А.Н. Об оценке качества решающего правила на основе малой обучающей выборки. –Там же, с.98–107.

5. МАНОХИН А.Н. Методы распознавания образов, основанные на логических решающих функциях. –Настоящий сборник, с.42–53.

6. БРАВЕРМАН Э.М., КИСЕЛЕВА Н.Е., МУЧНИК И.Б., НОВИКОВ С.Г. Лингвистический подход к задаче обработки больших массивов информации. –"Автоматика и телемеханика", 1974, № 11.

7. ЗАГОРУЙКО Н.Г., ЁЛКИНА В.Н., ТИМЕРКАЕВ В.С. Алгоритм заполнения пропусков в эмпирических таблицах (алгоритм "ZET"). –В кн.: Вычислительные системы. Вып. 61. Новосибирск, 1975, с. 3–27.

Поступила в ред.-изд.отд.
3 февраля 1976 года