ВЫЧИСЛИТЕЛЬНЫЕ СИСТЕМЫ
1971 г. вып. 44

УДК 621.391:519.2

ОБЩИЕ СВОЙСТВА ЗАДАЧ РАСПОЗНАВАНИЯ ОБРАЗОВ

Н.Г. Загоруйко

Основными задачами распознавания в соответствии с [1] мы считаем задачу поиска решающих функций (D), информативной системы признаков (X) и группировки (таксономии) объектов (S). К настоящему времени разработано большое число специальных алгоритмов, каждый из которых, как правило, предназначен для решения задач только одного из этих трех типов. Насколько оправдана такая специфичность ? Нельзя ли, отправляясь от общих свойств разных задач распознавания, сделать алгоритмы более универсальными ?

Для поиска ответа на эти вопросы рассмотрим на частном модельном примере процедуры, которым подвергаются исходные данные в процессе решения разных задач распознавания.

Исходные данные предоставляются обычно в виде прямоугольной таблицы, строками в которой служат изучаемые объекты $\mathcal{X}$ (предметы, люди, явления и т.д.), а столбцами - признаки или свойства этих объектов X. Элементами матрицы являются значения a_{ij} признака x_i у объекта $\mathcal{X}_j$. Способ записи величин a_{ij} безразличен - это могут быть десятичные, двоичные числа, буквы и т.д. Важно знать, чему соответствуют эти обозначения в реальной действительности и отсюда уметь пользоваться ими для установления отношения между совокупностями символов a_{ij}, например, измерять расстояния между объектами или признаками, устанавливать степень "похожести" той роли, которую играют объекты в разных системах, и т.д. [2] . Пусть признаки в таблице I таковы, что система эмпирических отношений между объектами не нарушается,

если значения признаков записывать в двоичной форме, а расстояния между совокупностями a_{ij} измерять в метрике L_1.

При решении задачи таксономии делается сокращение таблицы путем уменьшения числа строк. Из исходного множества $\mathcal{Z}$ объектов можно оставить их ("базовое") подмножество S, достаточно хорошо (в определенном смысле) представляющее всю совокупность строк, а остальные строки можно вычеркнуть, как это делается в алгоритме "принадлежности" [3] или "вычеркивания" [4]. Иногда объекты базового подмножества не точно совпадают с исходными объектами, а синтезируются по определенным правилам из объектов, представителями которых они будут служить в дальнейшем. Так делается, например, в алгоритмах таксономии типа "Форэль" [5] или "группировка" [6]. В том и другом случаях ясно, что не все варианты базовых подмножеств одинаковы. При одинаковой степени сокращения таблица 3 лучше отражает содержание таблицы I, чем таблица 2 *).

Таблица I

	X_1	X_2	X_3	X_4	X_5	X_6
z_1	I	0	0	I	0	0
z_2	0	0	I	0	I	0
z_3	0	I	0	0	0	I
z_4	0	0	I	0	I	0
z_5	0	I	0	0	0	I
z_6	I	0	0	I	0	0
z_7	0	0	I	0	I	0

$F_x = 0{,}222$; $F_{\bar{x}} = 0{,}267$;

$F_{x\bar{x}} = 0{,}059$.

Таблица 2

	X_1	X_2	X_3	X_4	X_5	X_6
S_{12}	$\frac{1}{2}$	0	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	0
S_{34}	0	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$	$\frac{1}{2}$
S_{567}	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

$F_s = 0{,}333$; $F_{\bar{x}} = 0{,}167$;

$F_{s\bar{x}} = 0{,}056$.

Таблица 3

	X_1	X_2	X_3	X_4	X_5	X_6
S_{16}	I	0	0	I	0	0
S_{35}	0	I	0	0	I	0
S_{247}	0	0	I	0	0	I

$F_s = 0{,}667$;

$F_{\bar{x}} = 0{,}267$;

$F_{s\bar{x}} = 0{,}178$.

*) Таксоны представлены в этих таблицах координатами математических ожиданий объектов, вошедших в данный таксон.

В разных алгоритмах используются разные критерии качества таксономии. Наиболее часто стремятся минимизировать расстояние (ρ) между объектами внутри одного таксона и максимизировать расстояние (d) между таксонами. Будем в наших примерах считать качество таксономии F_s прямо пропорциональным среднему расстоянию ($d_{ср}$) между таксонами, причем d будем измерять по кратчайшему незамкнутому пути [7]. По этому критерию оценка варианта таксономии для таблицы 2 равна 0,333, а таблицы 3 - 0,667.

Иногда сокращению подвергаются не строки, а столбцы. Содержательно это трактуется как поиск наиболее информативной подсистемы признаков. В одних случаях часть признаков из исходного набора оставляется в качестве искомой подсистемы X_n, а остальные $q-n$ признаков вычеркиваются (в алгоритме "СПА", например, [8]). В других случаях признаки группируются, и в сокращенной таблице остаются так называемые "вторичные" признаки ("факторы" "синдромы") Y, синтезированные из первичных по определенным правилам. При одном и том же методе факторизации качество факторов будет зависеть от того, какие первичные признаки сгруппированы в один фактор. Как и в случае таксономии, качество факторизации определяется используемым критерием. Смысл многих применяемых критериев сводится к следующему: набор тех факторов считается лучшим, в пространстве которых объекты (строки) максимально отличаются друг от друга. Так что, как и при таксономии, качество факторов можно оценивать по критерию $F_Y = d_{ср}$, причем d измеряются по кратчайшему пути между строками (группами строк).

В таблицах 4 и 5 представлены два варианта факторизации, которым соответствуют оценки качества $F_Y = 0{,}321$ и $F_Y = 0{,}643$.

Можно пойти ещё дальше и попытаться сократить исходную таблицу за счет одновременного уменьшения строк и столбцов. В этом случае мы будем иметь дело с задачей комбинированного типа [9]. Качество такого сжатия таблицы можно, вероятно, оценивать по критерию $F_{sY} = F_s \cdot F_Y$. По этому критерию таблица 6 оценивается величиной $F_{sY} = 0{,}009$, а таблица 7 - $F_{sY} = 0{,}445$.

Общим для всех случаев является следующее:

I. Фиксируется вид закономерности, которую мы хотели бы обнаружить в структуре множества (строк или столбцов).

2. Выбирается критерий для количественной оценки качества сокращения таблицы, который достигает экстремального (*max*) зна-

Таблица 4

	Y_{12}	Y_{34}	Y_{56}
x_1	1/2	1/2	0
x_2	0	1/2	1/2
x_3	1/2	0	1/2
x_4	0	1/2	1/2
x_5	1/2	0	1/2
x_6	1/2	1/2	0
x_7	0	1/2	1/2

$F_s = 0{,}111$; $F_Y = 0{,}321$;
$F_{sY} = 0{,}036$.

Таблица 5

	Y_{14}	Y_{35}	Y_{26}
x_1	1	0	0
x_2	0	1	0
x_3	0	0	1
x_4	0	1	0
x_5	0	0	1
x_6	1	0	0
x_7	0	1	0

$F_s = 0{,}222$; $F_Y = 0{,}643$;
$F_{sY} = 0{,}143$.

Таблица 6

	Y_{12}	Y_{34}	Y_{56}
S_{12}	1/4	1/2	1/4
S_{34}	1/4	1/4	1/2
S_{567}	1/3	1/3	1/3

$F_s = 0{,}111$
$F_Y = 0{,}083$
$F_{sY} = 0{,}009$

Таблица 7

	Y_{14}	Y_{35}	Y_{26}
S_{16}	1	0	0
S_{35}	0	1	0
S_{247}	0	0	1

$F_s = 0{,}667$; $F_Y = 0{,}667$;
$F_{sY} = 0{,}445$.

чения тогда, когда структура близкого подмножества полученным способом соответствует ожидаемой закономерности структуры исходной таблицы.

3. Исходное множество строк (столбцов) разбивается на некоторое число таксонов (факторов) так, что в один таксон объединяются строки (столбцы), структура отношений между которыми наилучшим образом соответствует ожидаемой закономерности.

Именно наличие закономерности в структуре множества эмпирических данных позволяет эффективно сокращать описание этого множества.

Самым распространенным видом закономерности, используемым в алгоритмах распознавания, является так называемая "компактность" [10] . Гипотеза компактности для рассмотренных примеров может быть сформулирована так: "среди множества объектов найдутся группы "похожих" объектов, расположенных в пространстве признаков геометрически близко ("компактно") друг к другу и максимально удаленных от объектов из других групп".

Объединение объектов таких "компактных" групп в таксоны позволяет синтезировать краткое описание (эталон, базовый объект), которое мало отличается от описания каждого объекта данного таксона и сохраняет важные отличия от описания объектов из других таксонов.

При сокращении столбцов мы руководствовались гипотезой о том, что "среди исходного множества признаков найдутся такие группы похожих признаков, расположенных в пространстве объектов геометрически близко ("компактно") друг к другу и максимально удаленных от признаков из других групп".

Сочетание слов "пространство объектов" использовано здесь по аналогии с более распространенным сочетанием "пространство признаков". В определенном смысле эти сочетания равноправны : об объекте мы судим по результату его взаимодействия с конечным набором приборов, а о приборе - по результату его взаимодействия с конечным числом объектов (набору эталонов, например). Кстати, результаты таксономии строк позволяли нам синтезировать новые объекты. Очевидно, что применение алгоритмов таксономии к столбцам позволяет формулировать требования к новым приборам. Нужно, однако, отметить, что вопрос о метрике пространства объектов не всегда имеет простой ответ.

Следует обратить внимание на то, что часто при сокращении

признаков руководствуются не расстояниями между признаками (F_x), а так же, как и при таксономии,–расстояниями между объектами (F_y). Из таблицы 8 видно, что F_x и F_x очень сильно коррелированы, и выбор варианта сокращения таблицы, возможно, не будет зависеть от того, какая из величин – F_y или F_x – максимизируется.

Таблица 8

Номер матрицы	F_x	F_Y	F_y	F_y	$F_y \cdot F_x$	Вид сокращения
I.	0,267		0,222		0,0593	исходная матрица
2.	0,I67			0,353	0,0555	сокращение
3.	0,267			0,67	0,I78	строк
4.		0,32I	0,III		0,0355	сокращение
5.		0,643	0,222		0,I43	столбцов
6.		0,083		0,III	0,009	одновременное сокращение строк и столбцов
7.		0,67		0,67	0,46	

Напрашивается также вывод о том, что процедуры, предназначенные для сокращения строк, могут применяться для сокращения столбцов и наоборот.

Легко представить себе применение метода СПА для выбора базового подмножества строк, а алгоритма "Краб" – для выделения факторов [*)] или выбора наиболее информативных признаков. Большой интерес представляет сравнение эффективности разных алгоритмов при решении одной и той же задачи.

Может оказаться, что при решении задачи комбинированного типа для сокращения столбцов и строк будет использоваться один и тот же алгоритм вместо используемой в [9] пары "Краб" – "СПА" окажется более целесообразной пара "Краб" – "Краб" или "СПА" – "СПА".

Такая возможность основана на том, что, как было показано выше, все варианты сокращения таблицы основаны на выявлении и использовании одинаковых закономерностей в структуре отношений между её элементами.

[*)] Алгоритмы самообучения уже используются для выделения факторов [II].

Обратимся теперь к задаче построения решающей функции D. Пусть "обучающая выборка" строк разбита на K непересекающихся подмножеств ("образов"). Выделим особый столбец, в который занесем информацию о принадлежности данной строки тому или иному образу (признак X_D в таблице 9). Предъявим теперь новую строку, содержащую все элементы, кроме a_{xD}.

Таблица 9

	X_I	X_2	X_3	X_4	X_5	X_6	X_D
x_I	I	I	I	0	0	0	0
x_2	I	I	I	0	0	0	0
x_3	I	I	I	0	0	0	0
x_4	0	0	0	I	I	I	I
x_5	0	0	0	I	I	I	□
x_6	0	0	0	I	I	I	I
x_D	I	I	□	0	0	0	

Задача состоит в том, чтобы правильно указать значение признака X_D для этой строки. Строгих формальных оснований для выбора того или иного значения пропущенного элемента в столбце X_D нет. Наиболее оправданной из используемых сейчас эвристических процедур является такая: выдвигается гипотеза G о закономерности в структуре отношений на множестве строк, и распознаваемая строка относится к той группе строк, в составе которой она наилучшим способом удовлетворяет этой закономерности. Как в предыдущих, так и в этой задаче можно пользоваться одной какой-нибудь гипотезой (например, "компактности"), однако, более подходящим, по нашему мнению, является метод последовательного опробования разных закономерностей и выбора простейшей закономерности, описывающей массив данных с заданной степенью точности [I2, 2].

Аналогичные рассуждения можно привести и для задачи распознавания принадлежности нового признака тому или иному из ранее выделенных факторов. В этом случае мы говорили бы о поиске закономерности для наилучшего предсказания пропущенно-

го элемента в особой строке, содержащей информацию о принадлежности признаков факторам (строка x_0 в таблице 9). Подход к этим задачам как к комбинированным позволяет переформулировать их следующим образом: нужно так заполнить пустую клеточку в таблице, чтобы закономерность в структуре как строк, так и столбцов становилась наиболее простой и четкой.

Можно рискнуть заполнить не один пропущенный элемент в таблице, а несколько. Пример такой задачи, которую после работы Себестьяна [13] многие считают безнадежной, приведен в таблице 10.

Таблица 10

	X_1	X_2	X_3	X_4	X_5	X_0
x_1	1	1	1	0		1
x_2	1		1	0	0	1
x_3		0	0	1	1	0
x_4	0	0		1	1	0
x_0	0	0	0		1	a_{00}

Все строки и столбцы имеют хотя бы по одному пропущенному элементу, однако общая закономерность структуры таблицы улавливается легко, что позволяет считать наиболее оправданным значение элемента a_{00}, равное нулю. Ясно, что надежность таких прогнозов тем меньше, чем больше пропусков содержит таблица $\mathfrak{X}X$.

Отметим, что главным элементом как процедуры сокращения таблиц, так и распознавания (угадывания) пропущенных элементов является обнаружение (простейшей) закономерности в структуре отношений, зафиксированных таблицей. Отсюда следует возможность применения одних и тех же алгоритмов для решения задач любого типа. В частности, в работе [7] описаны методы принятия решений, использующие алгоритм таксономии "Краб" ("Таксономические решающие функции"). Очевидна возможность применения алгоритмов таксономии для задачи упрощения решающих функций, рассмотренной в [14] .

На обнаружении и использовании закономерностей в эмпирических массивах основаны также задачи прогнозирования, анализа тенденций, планирования эксперимента и т.д., т.е. практически все задачи, решаемые научным методом.

Отсюда можно сделать вывод, что та естественно-научная область, которую мы называем "распознаванием образов", является, по-видимому, основной частью складывающегося сейчас научного направления, связанного с формализацией метода научного познания.

Литература

1. Загоруйко Н.Г. Классификация задач распознавания образов. "Вычислительные системы", Новосибирск, № 22, 1966.
2. Гаврилко Б.П., Загоруйко Н.Г., Самохвалов К.Ф. Уточнение гипотезы простоты. "Вычислительные системы", Новосибирск, № 37 , 1969.
3. Турбович И.Т. Об оптимальном методе опознания образов при взаимнокоррелированных признаках. Опознание образов. Теория передачи информации. Москва, АН СССР, ИППИ, Изд-во "Наука", 1965.
4. Ёлкина В.Н., Загоруйко Н.Г. Об алфавите объектов распознавания. "Вычислительные системы", Новосибирск, № 22,1966,
5. Ёлкина В.Н., Загоруйко Н.Г. Алгоритмы таксономии. В книге : "Распознавание образов в социальных исследованиях". Новосибирск, Изд-во "Наука", 1968.
6. Загоруйко Н.Г., Ёлкина В.Н. Алфавит с минимальной избыточностью. "Вычислительные системы", Новосибирск, № 28 , 1967.
7. Ёлкина В.Н., Загоруйко Н.Г. Количественные критерии качества таксономии и их использование в процессе принятия решений. "Вычислительные системы", Новосибирск, № 36 , 1969.
8. Лбов Г.С. Выбор эффективной системы зависимых признаков. "Вычислительные системы", Новосибирск, № 19, 1965.
9. Загоруйко Н.Г. Одновременный поиск эффективной системы признаков и наилучшего варианта таксономии (алгоритм "ΣX") "Вычислительные системы", Новосибирск, № 36, 1969.
10. Браверман Э.М. Опыты по обучению машины распознаванию зрительных образов. "Автоматика и телемеханика". т.XXI, № 3, 1962.

II. Браверман Э.М. Методы экстремальной группировки параметров и задача выделения существенных факторов. "Автоматика и телемеханика" № I, 1970.
I2. Загоруйко Н.Г., Самохвалов К.Ф. Природа проблемы распознавания образов. "Вычислительные системы", Новосибирск, № 36, 1969.
I3. Себестиан Г.С. Принятие решений при распознавании образов. Пер. с англ. Киев, изд-во "Техника", 1965.
I4. Загоруйко Н.Г. Линейные решающие функции, близкие к оптимальным. "Вычислительные системы", Новосибирск, № I9, 1965.

Поступила в редакцию
24.I2.I970.