УДК 681.3.06:621.391

ЧАСТЬ I. ПРОГРАММЫ ТАКСОНОМИИ

В настоящей главе опубликован ряд программ для решения задач таксономии (группировки объектов, выделения групп близких точек). Под задачей таксономии обычно понимают задачу поиска наилучшего в некотором смысле разбиения S множества Q объектов (реализаций) q_i на таксоны (формальные элементы) S_m в заданном пространстве признаков X с помощью решающих функций $\mathcal{D}$ определенного типа. При решении задачи таксономии, как правило, предполагается, что пространство признаков X и способ измерения расстояния ("близости") между точками выбраны (на основании знаний, опыта и интуиции исследователя) так, что точки, соответствующие "близким", "похожим" объектам - объектам одного класса - образуют изолированные области, достаточно удаленные от образов объектов других классов.

Данные алгоритмы не используют сведения об априорном распределении точек по таксонам. Все построения для заданного множества точек делаются только на основании предположения об изолированности областей, соответствующих различным образам.

По алгоритму "Фораль-I" ("Формальный элемент - I") объединяются в один таксон точки, близкие в евклидовом пространстве признаков. Граничная область отдельного таксона представляет собой гиперсферу. В данном сборнике публикуются программы "Фораль-I" для ЭВМ БЭСМ-6, М-20 (М-220) и Минск-22. Описание алгоритма и его применений можно найти в работах [1, 2, 3]. В программах предусмотрена возможность нормализации исходных данных по дисперсиям и, как правило, программы должны использоваться в режиме "с нормализацией".

Эти программы могут быть использованы, в основном, для ре-

мения задач с количественными или ранжированными качественными признаками. Применять эти программы для чисто двоичных кодов нецелесообразно ввиду нерационального использования машинного времени и памяти. Для двоичных кодов разработаны специальные алгоритмы (группа "Форэль-5") и программы. Логика алгоритмов группы "Форэль-5" та же, что и алгоритмов "Форэль-I".

Программа "Форэль-5 а" [I, 3] позволяет осуществлять разбиение множества на таксоны при условии равноценности (или почти равноценности) разрядов, как по полному набору признаков, так и по любому их подмножеству без искажения массива исходных данных. С этой целью в информации к программе задается маска, размерность которой совпадает с размерностью исходного двоичного кода. В маске на местах тех признаков, по которым должна быть проведена классификация, стоят единицы. Если нужна классификация по всему набору признаков, то во всех разрядах должны стоять единицы. Так как при счете исходный массив данных не искажается, то на одном и том же массиве данных можно без дополнительного ввода просчитать несколько вариантов задачи.

В алгоритме "Форэль-5 б", помимо задания признакового подпространства с помощью маски, предусмотрена возможность введения "весов" для признаков и отдельных разрядов. Это бывает необходимо в том случае, когда признаки неравноценны или закодированы существенно различным количеством двоичных разрядов.Для возможности учета весовых коэффициентов в алгоритме "Форэль-5б" производится сравнение двоичных кодов $y_k(\xi_{k1},\dots,\xi_{kn})$ и $y_e(\xi_{e1},\dots,\xi_{en})$ по "взвешенному" хэммингову расстоянию:

$$d(y_k, y_e) = \sum_{m=1}^{n} (\xi_{km} + \xi_{em}) \bmod 2 \cdot W_m ,$$

где W_m - "вес", приписываемый каждому разряду двоичного кода. Разряды, относящиеся к одному и тому же признаку x_i , имеют равные "веса" W_i . Эти веса W_i должны быть заданы перед началом работы программы. В общем случае каждому двоичному разряду ξ_{ip} $(p = 1,\dots,\alpha_i)$ приписывается "вес" $W_i = \beta_i/\alpha_i$, где β_i - "вес" признака x_i , α_i - количество двоичных разрядов, отведенных под признак x_i . В случае, когда нет оснований выделять какие-либо признаки, т.е. $\beta_i = 1$, тогда $W_i = 1/\alpha_i$. Если и все разряды равнозначны,то матрица весов состоит из единиц. В этом случае обе программы дают один и тот же результат, но "Форэль-5 а" требует меньше машинного времени.

Как уже говорилось ранее, алгоритмы "Форэль-5 а" и "Форэль-5 б" предназначены для работы с данными, представленными двоичными кодами. Рассмотрим, какого типа информацию удобно представлять в таком виде. Прежде всего следует упомянуть о кодировании информации анкетного типа, например, такой как ответы испытуемого на те или иные вопросы, когда он должен сказать "да" или "нет". Признаки такого рода кодируются: I - "да", 0 - "нет". Двоичным кодом могут быть закодированы и количественные признаки, такие как вес, рост, размер и т.п. Представить число в виде двоичного кода можно различными способами. При подготовке данных для решения задач таксономии с помощью программ группы "Форэль-5" удобно пользоваться следующим методом. Диапазон изменения $(x_{i\max} - x_{i\min})$ конкретного числового признака x_i разбивается на некоторое число α_i + I градаций и соответственно под этот признак отводится α_i двоичных разрядов.Кодировать следует так, чтобы $x_{i\max}$ был представлен в виде набора из α_i единиц, а $x_{i\min}$ - из α_i нулей. Каждое значение признака $x_i^j (x_{i\min} \le x_i^j \le x_{i\max})$ кодируется соответствующим ему набором из α_i^j подряд идущих единиц и $\alpha_i - \alpha_i^j$ нулей. Например, для некоторого множества признак x_i принимает следующие значения: 5, 7, I7, I0, 8. Здесь $x_{i\max} = I7$, $x_{i\min} = 5$, $x_{i\max} - x_{i\min} = I2$. Пусть под этот признак отведено 5 разрядов. Тогда в закодированном виде приведенная выше последовательность значений признака имеет вид : $5 \sim 00000$; $7 \sim I0000$; $I7 \sim IIIII$; $I0 \sim III00$; $8 \sim II000$.

Выделение таксонов сложной формы осуществляется алгоритмом "Краб" [4, 5] . Этим алгоритмом производится поиск лучшего по критерию качества F (см. [4, 5]) разбиения множества на заданное число таксонов.

Функция F не зависит от числа точек L и количества таксонов K , что дает возможность не только оценивать различные разбиения одного и того же множества Q на равное число таксонов, но и определять наиболее предпочтительное для данного множества количество K таксонов S_m , а также сравнивать качество разбиений различных множеств Q_j на K_j таксонов .

Предварительно перед работой программы "Краб" все точки

исходного множества должны быть соединены в связный неориентированный граф без петель с минимальной суммарной длиной ребер [6, 7] (рис. I) по программе "Построение кратчайшего пути" (данный сборник).

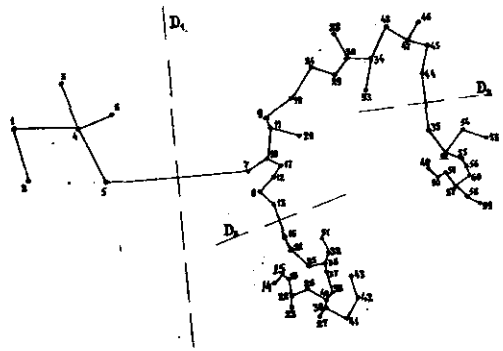

Рис. I. Соединение точек кратчайшими незамкнутыми путями.

Результаты работы программы ПКП являются исходными данными для программы "Краб".

Литература

I. Ёлкина В.Н., Загоруйко Н.Г. Об алфавите объектов распознавания. Сб. Вычислительные системы, вып. 22, Новосибирск, Изд-во "Наука", I966.

2. Ёлкина В.Н., Ёлкин Е.А., Загоруйко Н.Г. О возможности применения методов распознавания образов в палеонтологии. Геология и геофизика, № 9, I967.

3. Загоруйко Н.Г., Заславская Т.И., Ёлкина В.Н., Лбов Г.С. и др. Распознавание образов в социальных исследованиях. Новосибирск, Изд-во "Наука", I968.

4. Ёлкина В.Н., Загоруйко Н.Г. Количественные критерии качества таксономии и их использование в процессе принятия решений. Сб. Вычислительные системы, вып. 36, Изд-во "Наука", Новосибирск, I969.

5. Ёлкина В.Н. Выбор формальных элементов алфавита (алгоритмы таксономии). Автореферат диссертации. Новосибирск, I969.

6. Boruvka O., On a minimal problem. Prace moraske, Pridovedecke Spolecnosti, N 3, 1926.

7. Kruskal J.B., On a shortest spanning subtree of graph and the travelling salesman problem. Proc. Amer. Math.Soc., N 7, 1956.

В.Н.Ёлкина