«Оптика атмосферы», **1**, № 9 (1988) УДК 519.92

В.А. Алдын-Херель, С.В. Сапожников

АЛГОРИТМ КЛАССИФИКАЦИИ РЕЗУЛЬТАТОВ ДИСТАНЦИОННОГО ЗОНДИРОВАНИЯ ПОДСТИЛАЮЩЕЙ ПОВЕРХНОСТИ

На примере решения задачи зондирования загрязняющих пятен на взволнованной водной поверхности с летательного аппарата формулируется задача кластерного анализа и предлагается эффективный алгоритм, основанный на преобразовании исходных данных в пространство позиционных координат с нумерацией типа Z-развертки и расстоянием Хемминга в качестве метрики.

При обработке результатов дистанционного зондирования подстилающей поверхности с летательных аппаратов часто требуется разделить исходный набор экспериментальных данных на непересекающиеся классы или группы, руководствуясь некоторыми параметрическими или непараметрическими критериями. Если можно произвести предварительное обучение обрабатывающей системы путем введения обучающих выборок, каждая из которых заведомо содержит данные только одного из известных классов, то задача решается методом дискриминантного анализа. При невозможности проведения этапа обучения ставится задача кластерного анализа.

Рассмотрим в качестве примера задачу поиска загрязняющих пятен (например, нефтяных) на взволнованной водной поверхности методом дистанционного зондирования, безразлично активного или пассивного. Следует отметить, что невооруженным глазом такие загрязнения обнаружить практически невозможно. Пусть исследуемый район водной поверхности имеет прямоугольную форму и разбит на целое число K квадратных участков, таких что их изображение фиксируется бортовой аппаратурой как результат единичного измерения с фиксированным номером i ($i = 1,\ldots,K$). Таким образом, объем выборки результатов зондирования равен $K = M \cdot N$, где N — количество галсов летательного аппарата, а M — количество изображений, фиксируемых при проходе одного галса. Каждое из этих изображений подвергается некоторой предварительной обработке, в результате чего вырабатывается вектор информативных признаков, адекватно, в некотором смысле, представляющих особенности данного изображения. Методы получения таких признаков достаточно полно описаны в литературе (см., например, [4]) и в данной работе рассматриваться не будут. Обычно для таких задач вектор признаков содержит от 5 до 15 элементов.

Итак, результаты зондирования представляются на обработку в виде набора K векторов, длиной p, где p — количество информативных признаков (размерность признакового пространства):

$$X_i = \{x_{i1},\ x_{i2},\ldots x_{ip}\},\ i = 1,\ldots,\ K. \qquad (1)$$

Необходимо разбить эту последовательность, по меньшей мере, на две группы, одна из которых содержит номера изображений чистых участков, а вторая — номера участков поверхности, загрязненных нефтяной пленкой. Для простоты полагаем, что выполнены условия квазистационарности волнения в исследуемом районе. Вполне понятно, что объективно может быть обнаружено более 2-х классов хотя бы за счет изображений, на которых только часть площади покрыта пленкой, например, на границах загрязняющего пятна.

Результаты классификации выдаются в виде так называемого кадра решений, представляющего собой карту исследуемого района, каждый из K квадратов которой помечен определенным образом, например, раскрашен своим цветом. В результате должны проявиться загрязняющие пятна, контуры которых определяются с точностью до размера фиксируемого квадратного участка.

Сформулированная выше задача является типичной задачей обработки данных на основе кластерного анализа. Характерной чертой всех методов кластерного анализа, как отмечено в работе [1], является резкий контраст между простотой описаний алгоритмов и чрезвычайной сложностью получаемых программ для ЭВМ. Кроме того, эти программы требуют для своей работы значительных ресурсов — оперативной памяти и времени вычислений на ЭВМ. Достаточно сказать, что практически для всех методов необходимо вычислять и хранить матрицу межэлементных попарных расстояний, треугольник которой содержит $K \cdot (K - 1)/2$ чисел для выборки длиной K. Для нашего примера небольшие размеры исследуемого района (20×20 квадратов) дадут выборку с $K = 400$, число попарных расстояний в этом случае — около 80 тыс., что почти в 4 раза превышает объем доступной для программ оперативной памяти ЭВМ «Электроника–60». Предлагается алгоритм, который свободен от перечисленных недостатков. Этот алгоритм опирается на пирамидально–рекурсивное представление данных [2, 3]. Без потери общности можно считать, что все компоненты вектора измерений

(1) являются численными и масштабированы в полуинтервале [0, 1]. Тогда каждый элемент исследуемой выборки можно представить точкой в p-мерном пространстве R_p, ограниченном единичным гиперкубом. Точки, принадлежащие каждому классу, образуют некоторые сгущения или облака произвольной формы. Метод, положенный в основу алгоритма, предусматривает вычисление с требуемой точностью кусочно-линейной аппроксимации замкнутых гиперповерхностей, ограничивающих эти облака. Для этого исходное пространство — единичный гиперкуб в R_p — разбивается на q^p меньших гиперкубов — квантов первого уровня разбиения, где q — произвольное натуральное число. Каждый квант первого уровня, в свою очередь, может быть разбит на q^p квантов второго уровня разбиения и т.д., вплоть до уровня, обеспечивающего требуемую точность аппроксимации ограничивающих гиперповерхностей. Кванты каждого уровня нумеруются определенным и одинаковым для всех уровней образом. Полный номер некоторого j-го кванта m-го уровня разбиения можно записать в виде

$$V_j^{(m)} = \{v_j^1,\ v_j^2,\ldots,\ v_j^m\}, \tag{2}$$

где числа $\upsilon_j^l (l = 1,\ldots,\ m-1)$ — номера квантов предыдущих уровней разбиения, включающих рассматриваемый квант с номером υ_j^m. Вектор (2) определяет так называемые позиционные координаты кванта m-го уровня разбиения в пространстве D_m. Существует взаимно–однозначное отображение

$$F_q : R_p \to D_m, \tag{3}$$

преобразующее декартовы координаты вектора $x_j \in R_p$ в позиционные координаты $V_j \in D_m$. Выбор $q = 2$ значительно упрощает вычисление позиционных координат на ЭВМ с двоичной системой исчисления. Пример, иллюстрирующий разбиение двумерного пространства ($p = 2$) на два уровня квантования с нумерацией квантов типа Z–развертки и $q = 2$, приведен на рисунке.

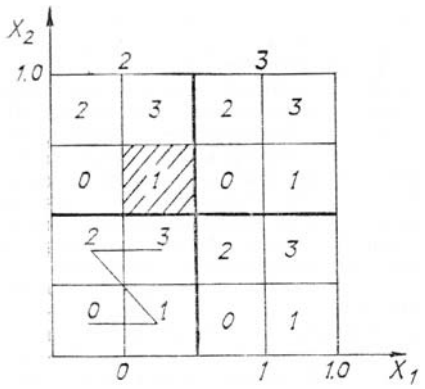
Полные позиционные координаты заштрихованного кванта равны {2, 1}. Легко видеть, что позиционные координаты некоторого кванта m-го уровня разбиения равны m-м двоичным разрядам соответствующих декартовых координат, и явное выражение для оператора (3) имеет вид

$$F_2 : v_i = \sum_{j=1}^{p} x_j^{(i)} \cdot 2^{p-j},\ i = 1,\ m, \tag{4}$$

где $x_j^{(i)} \in \{0,1\}$ — i-й двоичный разряд j-й декартовой координаты. Обратное преобразование получается простой заменой m на p:

$$F_2 : x_i = \sum_{j=1}^{m} v_i^{(j)} \cdot 2^{m-j},\ i = 1,\ldots,\ p, \tag{5}$$

где $\upsilon_j^{(i)} \in \{0,1\}$ — i-й двоичный разряд j-й позиционной координаты.



Разбиение двумерного пространства на два уровня. Заштрихованный квант имеет позиционные координаты 2,1

Исходная выборка (1), преобразованная в позиционные координаты по правилу (4), образует некоторую иерархическую рекурсивную структуру, напоминающую пирамиду, откуда и следует название — «пирамидально–рекурсивное представление данных». В памяти ЭВМ эта структура хранится в виде связанного списка или дерева. Вершинам дерева соответствуют кванты определенных уровней. В процессе обработки данных вычисляется вес вершин, равный количеству элементов ис-

ходной выборки, попавших в соответствующий квант. Такой список уже некоторым образом отражает структуру исходных данных. Теперь достаточно произвести кластеризацию квантов одного уровня обычным агломеративным методом, что представляет собой существенно более простую задачу, т.к. квантов много меньше, чем элементов в исходной выборке. Кроме того, в пространстве позиционных координат D_m более простой является процедура проверки критерия принадлежности к классу. Очевидно, что связанные структуры данных должны образовывать в D_m связанные цепочки — кластеры соседних квантов одного уровня разбиения. Кванты одного уровня разбиения являются соседними, если имеют общую грань порядка $p-1$. При этом возможны два случая:
— оба соседних кванта принадлежат одному кванту предыдущего уровня разбиения;
— кванты принадлежат разным соседним квантам предыдущего уровня.

Правило установления соседства в первом случае очевидно: кодовое расстояние Хемминга между позиционными координатами этих квантов должно быть равно единице:

$$h_{ij} = \sum_{s=1}^{p} i^{(S)} \oplus j^{(S)} = 1; \tag{6}$$

где $i^{(s)}$, $j^{(s)}$ — S-е двоичные разряды позиционных координат сравниваемых i-го и j-го квантов; $\oplus$ — операция поразрядного сложения по mod 2. Для второго случая это правило определяется аналогично:

$$h_{i_{m-1}j_{m-1}} = 1, \tag{7а}$$

$$h_{i_m j_m} = 1; \tag{7б}$$

$$i^{(s)}_{m-1} \oplus j^{(s)}_{m-1} = i^{(s)}_{m} \oplus j^{(s)}_{m}; \; s = 1,\ldots, \; p. \tag{7в}$$

Введение условия (7 в) объясняется тем, что кванты, соседние на m-м уровне разбиения должны располагаться вдоль той же оси (5), что и содержащие их кванты $m-1$-го уровня.

Следует заметить, что условиям (7 а)—(7 в) удовлетворяют также пары противолежащих квантов уровня m, которые не могут быть соседними. Поэтому, введя еще одно условие, проверим на соседство только те кванты, для которых справедливо:

$$\begin{aligned} &i_m < j_m, \text{ если } i_{m-1} > j_{m-1}; \\ &i_m > j_m, \text{ если } i_{m-1} < j_{m-1}. \end{aligned} \tag{7г}$$

Итак, на втором этапе процесса кластеризации для заданного уровня разбиения проверяются условия (6), (7 а)—(7 г) для всех пар квантов этого уровня и вершины дерева, соответствующие квантам, для которых выполнены эти условия, помечаются номером соответствующего класса.

Процедуру проверки условий соседства (6), (7) можно значительно упростить, если перейти обратно от позиционных к декартовым координатам, которые будут теперь целочисленными величинами с m двоичными разрядами:

$$Y_i = \{y_{i1}^{(m)}, \ldots, \; y_{ip}^{(m)}\}. \tag{8}$$

Условием соседства в этих координатах будет равенство единице расстояния Хемминга:

$$H_{ij} = \sum_{s=1}^{p} |y_{iS} - y_{jS}| = 1. \tag{9}$$

Кроме того, метрику H_{ij} можно использовать для определения центров классов, межклассовых относительных расстояний и т.п.

На третьем этапе процесса кластеризации необходимо определить номера элементов исходной выборки, принадлежащих каждому классу. Если позволяет наличная память ЭВМ, можно хранить эти номера вместе с каждой вершиной дерева. Если же память ограничена, то исходная выборка снова подвергается преобразованию (4) и каждому элементу приписывается тот номер класса, которым помечена вершина дерева с позиционными координатами, равными позиционным координатам классифицируемого элемента.

В заключение необходимо сделать замечание о выборе номера уровня разбиения m, соответствующего объективной кластеризации исходной выборки. Вполне понятно, что невозможно выработать единый формальный критерий объективности полученных результатов кластеризации без привлечения дополнительной информации об исследуемом объекте типа «ожидаемого количества классов».

При отсутствии информации для оценки «объективного» уровня m можно использовать следующий эмпирический прием [2]. Если представить зависимость количества определенных классов от параметра m в виде графика, то можно заметить, что при некотором значении $m = m_0$ этот график имеет излом, после которого количество классов резко возрастает. Это значение m_0 можно взять в качестве оценки номера уровня разбиения, приблизительно соответствующего объективной кластеризации.

Программа, реализующая обсуждаемый алгоритм, была написана на языке Фортран для ЭВМ «Электроника–60» с операционной системой Рафос (RT–11) и показала при эксплуатации неплохие результаты. Для приведенного примера с объемом выборки $K = 420$ время счета составило около 50 с, а обрабатываемые данные в структурированном виде удалось разместить в массиве размером 512×5 ячеек памяти ЭВМ «Электроника–60». Реализация традиционных алгоритмов кластерного анализа для ЭВМ этого класса является весьма трудной задачей из-за ограниченного объема оперативной памяти.

1. Статистические методы для ЭВМ./Под ред. К. Энслейна, Э. Рэлстона, Г.С. Уилфа. Пер. с англ. Под ред. М.Б. Малютова. М.: Наука, 1986. 464 с.
2. Александров В.В., Горский Н.Д. Алгоритмы и программы структурного метода обработки данных. Л.: Наука. 1983. 208 с.
3. Александров В.В., Горский Н.Д. Представление и обработка изображений: Рекурсивный подход. Л.: Наука. 1985. 192 с.
4. Дуда Р., Харт П. Распознавание образов и анализ сцен/Пер. с англ. Под ред. В.Л. Стефанюка. М.: Мир. 1976. 511 с.

Институт оптики атмосферы
СО АН СССР, Томск

Поступила в редакцию
15 марта 1988 г.

V.A. Aldin-Kherel, S.V. Sapozhnikov. **Data Classification Algorithm for Remote Sounding of Underlying Surface.**

The cluster analysis problem is formulated. As an example, the solution to the problem of airborne sounding of pollution spots on the swelling sea surface is considered. An efficient classification algorithm is proposed based on the source data conversion into a position coordinate system with the numbering of the Z-scanning type and the Hamming distance used as a metric.