Поступила в редколлегию 14.01.81.

УДК 681.327.120.888

Г. Я. ШЕВЧЕНКО, канд. техн. наук, *А. Н. ПЕРКИН*,
А. М. ПРЯНИЦКИЙ, канд. техн. наук, *В. А. ЧИКИНА*, канд. техн. наук

СТРУКТУРНО-ЛОГИЧЕСКИЙ МЕТОД РЕШЕНИЯ ЗАДАЧИ КЛАСТЕРИЗАЦИИ

В большинстве случаев решение задач классификации проводится с использованием ЭВМ. Поэтому широкое развитие получили методы, учитывающие, помимо прочего, особенности реализации этих методов на универсальных ЭВМ. Например, предложено решать задачу кластеризации на основе анализа связанности структуры данных [1]. Однако связанные структуры в общем случае могут быть довольно причудливой формы, что не согласуется с гипотезой компактности [2]. С другой стороны, необходима интерпретация выделенной структуры данных для человека, что наилучшим образом осуществляется в терминах исчисления высказываний, при этом, как и в [3, 4], будем считать ценность высказывания тем больше, чем оно короче. При геометрической интерпретации высказываний они представляются в виде совокупности гиперпараллелепипедов в пространстве признаков. Очевидно, что чем короче высказывание, тем меньше число гиперпараллелепипедов, представляющих его, и в пределе самое короткое высказывание — два гиперпараллелепипеда.

С учетом сказанного выше целесообразно ввести некоторые ограничения на возможные структуры данных, сформулировать их и использовать далее в работе для решения задачи кластеризации на основе метода, который назван в данной работе структурно-логическим.

Сформулируем ряд определений, на основании которых дадим формальную постановку задачи распознавания образов с самообучением, т. е. задачи кластеризации. Предполагается, что признаки измерены в шкалах порядка, а если нет — то сведены к таким путем квантования количественных шкал и упорядочения значений шкал и значений шкал наименований [1].

Пусть $X=\prod_{i=1}^{n} x_i$ — множество всевозможных значений экспериментальных данных, $x_i=\{0, 1, .., k_i-1\}$ — упорядоченное множество значений признака $i(i=\overline{1,n})$, n — количество признаков, k_i — значность признака i. Обозначим через $\tilde{X}$ множество экспериментальных данных. Ясно, что $\tilde{X}\subseteq X$.

Дадим следующие определения.

Определение 1. *Элементы* $x=(x_1 x_2 \dots x_n)$ *и* $x'=(x'_1 x'_2 \dots x'_n)$ *множества* X *называются соседними, если* $|x_i-x'_i|\in\{0,1\}$ *для* $i=\overline{1,n}$.

Определение 2. *Элементы* $x, x'\in X$ *называются* l_1*-соседними* ($l_1\in\{0, 1, \dots, n\}$), *если они являются соседними и* $l_1=n-\sum_{i=1}^{n}|x_i-x'_i|$.

Определение 3. *Множество* $X^*\subseteq X$ *называется* l_2*-связным* ($l_2\in\{0, 1, \dots, n\}$), *если для любых* $x, x'\in X^*$ *существует последовательность* ($x, \dots, x'', \dots, x'$); *каждый элемент которой принадлежит* X^* *и в которой любые два рядом стоящих элемента не менее, чем* l_1*-соседние.*

Определение 4. *Множество* $X^*\subseteq X$ *называется связным, если* X^* *является 0-связным и называется максимально связным, если* X^* *является* $(n-1)$*-связным.*

Определение 5. *Множество* $X^*\subseteq X$ *называется выпуклым, если* X^* *максимально связно и если для любых* $x, x'\in X^*$ *не существует* $x''\in X\setminus X^*$, *для которого и* x *и* x' *является* $(n-1)$*-соседними.*

Определение 6. *Множество* $X^*\subseteq\tilde{X}$ *называется* l_3*-изолированным* ($l_3\in\{0, 1, .., n\}$), *если для любого* $x\in X^*$ l_1*-соседними являются только* $x'\in X^*\cup(X\setminus\tilde{X})$.

Для оценки качества кластеризации, полученной в результате работы соответствующего алгоритма обычно, как например в [5], формально определяют, с одной стороны, меру близости объектов, принадлежащих одному кластеру, а, с другой стороны, вводят меру взаимной удаленности кластеров. В данной работе предлагается оценивать близость r_1 объектов одного кластера на основе одновременного учета степени связности и выпуклости множества, образующего кластеры:

$$r_1=\frac{(l_2/n)+v}{2},$$

где

1, если множество выпукло,
0, в противоположном случае.

Взаимная удаленность r_2 кластеров оценивается как степень изолированности множеств, образующих эти кластеры: $r_2=l_3/n$.

Кроме этих двух показателей качества классификации будем пользоваться еще одним показателем r_3, отражающим полноту охвата кластерами множества экспериментальных данных, и равным отношению суммы мощностей множеств, представляющих кластеры к мощности множества экспериментальных данных. Целью анализа является нахождение кластеров, максимизирующих векторный критерий качества кластеризации $R=(r_1, r_2, r_3)$.

Однако такая оптимизационная задача не может быть поставлена на множестве исходных данных, так как эти данные, с одной стороны, содержат

ошибки измерений, копирования и первичной обработки данных, а, с другой стороны, исходных данных зачастую бывает недостаточно, чтобы из них построить кластеры с высоким значением критерия качества, поэтому необходимо преобразование исходных данных. Это необходимо для того, чтобы устранить ошибки, уменьшающие степень связности и выпуклость кластеров, а также устранить изолированность элементов, принадлежащих одному кластеру и сделать изолированными те кластеры, которые из-за ошибок оказались связными.

Предлагаемые преобразования исходных данных в общем виде функции двух типов:

$$x^t = F'(x^{t-1}, X_x^{t-1}); \quad X_x^t = F''(x^{t-1}, X_x^{t-1}),$$

где x^t, x^{t-1} — элементы множества X в моменты времени t и $(t-1)$ соответственно; X_x^t, X_x^{t-1} — некоторые подмножества множества элементов, соседних с x в моменты времени t и $(t-1)$ соответственно.

Общая схема алгоритма преобразования исходных данных состоит в следующем. Из введенных выше функций выбирается набор функций, каждая из которых влияет отдельно на связность, выпуклость и изолированность множества. Алгоритм преобразования состоит в применении этих функций в некотором сочетании различное число раз. После каждой группы преобразований вычисляется векторный критерий качества. Число применений набора функций ограничено тем, что либо в результате преобразований все данные исчезнут, либо данные займут все признаковое пространство. Наступление этих событий означает останов алгоритма. Далее из всех полученных векторных критериев качества выбираются те, которые превосходят остальные по всем трем параметрам, и выдается сообщение о кластерах, соответствующих этим критериям. Таких различных кластеризаций может быть несколько. В случае нескольких типов кластеризаций пользователь выбирает искомую из дополнительных соображений.

Немаловажным аспектом в структурных методах распознавания является представление полученных образов. Обычно это осуществляется снижением размерности исходного пространства до двух- или одномерного пространства. Однако такое преобразование невозможно без потери информации о связности структуры образов. Чтобы избежать такой потери информации, в данной работе предлагается представлять образы в виде логических зависимостей между значениями признаков, которые легко интерпретируются человеком.

Список литературы: 1. *Александров В. В., Горский Н. Д.* Алгоритмы и программы структурного метода обработки данных. Л., 1983. 208 с. 2. *Браверман Э. М., Мучник И. Б.* Структурные методы обработки эмпирических данных. М., 1983. 464 с. 3. *Бонгард М. М.* Проблема узнавания. М., 1967. 297 с. 4. *Лбов Г. С.* Методы обработки разнотипных экспериментальных данных. Новосибирск, 1981. 159 с. 5. *Загоруйко Н. Г., Елкина В. Н., Лбов Г. С.* Алгоритмы обнаружения эмпирических закономерностей. Новосибирск, 1985. 109 с.

Поступила в редколлегию 09.01.89