УДК 62-5:007:621.391:519.2

ОДНОВРЕМЕННЫЙ ПОИСК ЭФФЕКТИВНОЙ СИСТЕМЫ ПРИЗНАКОВ И НАИЛУЧШЕГО ВАРИАНТА ТАКСОНОМИИ (алгоритм "SX")

Н.Г. Загоруйко

Задачи распознавания образов возникают тогда, когда нужно по результатам единичного эксперимента определить, с явлением или объектом какого типа мы имеем дело; выбрать свойства (признаки), по которым объекты одного типа отличаются от других; разделить некоторый экспериментальный массив на группы похожих (в каком-то смысле) объектов.

В соответствии с терминологией, используемой в /1/, задачу поиска информативной системы признаков X будем обозначать A_X, а задачу поиска наилучшего в смысле некоторого критерия качества F разбиения S исходного множества Z на K групп ("таксонов") будем называть задачей таксономии и обозначать A_S.

В литературе по распознаванию образов основное внимание уделяется решению задач таких основных типов, как задача A_D (поиск решающей функции D и задачи A_X и A_S. В последнее время в связи с рассмотрением многоуровневых (иерархических) распознающих систем /2/, все больше осознается необходимость разработки алгоритмов для одновременного поиска более чем одного из процедурных элементов (D, X и S).

Решение одной из таких задач комбинированного типа – задачи A_{DX} – описано в книге Г. Себестиана /3/.

Проработка этого направления показывает, что в многоуровневых системах имеют смысл все возможные задачи комбинированного типа – A_{DX}, A_{DS}, A_{SX} и A_{DSX}. Опишем две разновидности задачи типа A_{SX}. Алгоритм для решения этой задачи мы и назовем алгоритмом "SX".

Задача A_{SX} является одной из наиболее интересных и важных для приложения задач комбинированного типа. Необходимость одновременного поиска удобной для некоторой цели группировки (таксономии) и системы информативных признаков возникает довольно часто – при выборе промежуточного алфавита в многоуровневой системе распознавания речевых сигналов, при работе с незнакомой геологической или биологической коллекцией, при анализе объектов и явлений в социальных исследованиях и т.д.

Для решения этой задачи необходимо задать следующие элементы:

D – тип решающих функций, с помощью которых определяется принадлежность точки к тому или иному таксону;

N_0 – предельно допустимые затраты, складывающиеся из "стоимости процедуры" решения задачи (сюда входят затраты на память, число машинных операций, вес, габариты устройства и т.д.) и "стоимости потерь" (здесь – потери информации при группировке);

Z – исходное множество реализаций (точек), подлежащих таксономии;

K – желательное число таксонов ($K < Z$).

В итоге решения нужно найти сочетание такого варианта таксономии S_Ω (из всех Ω возможных вариантов) и такой системы признаков X_n (из исходной системы описания X_q, $n < q$), при котором суммарные затраты N были бы минимальными и меньше допустимой величины N_0.

Сказанное выше можно записать в следующем виде:

$$n, \Omega = \underset{\substack{n \in q \\ \Omega \in \Omega}}{\arg\min} N(X_n, S_\Omega) \Big/ D, Z, K, N_0 \; ^{*)}.$$

Процесс решения комбинированной задачи можно практически свести к последовательному применению алгоритмов решения составляющих её задач основных типов. Интерес представляет именно порядок применения этих алгоритмов и правила перехода от одного

*) Использование термина "$\arg$" (от слова "аргумент" и по аналогии с обозначением обратных тригонометрических функций "$\arc$") предложено В.А. Ковалевским.

алгоритма к другому.

В данном случае, на этапе поиска системы признаков будем использовать алгоритм случайного поиска с адаптацией ("СПА" /4/), а на этапе таксономии – один из алгоритмов поиска формальных элементов алфавита (например, алгоритм "Краб" /5/).

Первый вариант алгоритма "SX" предназначен для решения задачи A_{SX} в одноступенчатом устройстве или в режиме самообучения", т.е. когда результат группировки множества Z оценивается по некоторому критерию качества F, отражающему такие требования, как "близость" точек внутри одного таксона друг от друга и т.д.

Этот вариант алгоритма "SX" состоит в следующем:

1. Отрезок вещественной оси (0–1) разбивается на q равных отрезков. Ширина каждого отрезка ($\frac{1}{q}$) равна априорной вероятности P_q включения каждого из q признаков в систему X, состоящую из n признаков (X_n), где $n < q$.

2. Датчиком случайных чисел выбирается система $X_n^{(1)}$.

3. В пространстве $X_n^{(1)}$ с помощью одного из алгоритмов таксономии (например, алгоритма "Краб" /5/) множество Z группируется в K таксонов так, чтобы качество таксономии F достигало максимальной величины.

4. Ищется описание полученных таксонов с помощью набора решающих функций заданного типа D.

5. Оцениваются суммарные затраты

$$N^* = N(S_\Omega) + N(X_n),$$

где $N(S_\Omega)$ – сумма затрат, необходимых для использования набора решающих функций типа D, описывающего K таксонов, и затрат, связанных с потерями R информации вследствие перекодировки, а $N(X_n)$ – затраты на измерение n выбранных признаков.

6. Процедуры по п.п. 2,3,4 и 5 повторяются m раз.

7. Затем осуществляется "поощрение" и "наказание" признаков путем увеличения вероятности P_q для признаков, входивших в систему $X_n^{(i)}$, обеспечившую наименьшее значение N^*, и уменьшения P_q для признаков из системы $X_n^{(i)}$, соответствующей наибольшему значению N^*. Величина поощрения и наказания

зависит от N^* и состоит в увеличении или уменьшении ширины соответствующих отрезков $\frac{1}{2}$ на величину $h << \frac{1}{2}$.

8. Процедуры по п.п. 5 и 6 повторяются до останова по критерию, используемому в СПА, т.е. тогда, когда на нескольких шагах подряд выбирается одна и та же система признаков X_n.

В итоге выбирается пространство X_n, наиболее удобное для хорошей таксономии множества Z на K таксонов.

Если K не строго фиксировано, а задано в диапазоне значений от K_1 до K_2, то блоком 3 выбирается такое значение $K_1 < K < K_2$, при котором функция F максимальна.

Второй вариант алгоритма "SX" предназначен для решения задачи A^*_{SX} в устройстве, состоящем более чем из одной ступени, например, в двухступенном распознающем автомате.

Здесь множество Z представляет собой список элементов промежуточного алфавита S. Максимальное сокращение алфавита S за счет объединения его элементов в группы (таксоны) может быть осуществлено алгоритмом типа "группировка" /6/. В результате алфавит S преобразуется в алфавит S_Ω, элементы которого представляют некоторые скопления элементов алфавита S. Лучшим считается тот вариант группировки, при котором выполняются условия достаточности и необходимости (в смысле /6/) и величина N_Σ суммарных затрат на устройство в целом достигает минимального значения. Отсюда, в пункте 5 выше описанного алгоритма оценивается не величина N^*, а $N_\Sigma = N^*_i + N_{i+1}(D)$, где i - номер иерархического уровня, алфавит и система описания на котором подвергаются сокращению, а $N_{i+1}(D)$ - стоимость классификатора, принимающего решения об элементах алфавита S_{i+1} по последовательности элементов алфавита S_i.

Использование аналогичного подхода может оказаться полезным и при решении других задач комбинированного типа.

ЛИТЕРАТУРА

1. Н.Г. ЗАГОРУЙКО. Классификация задач распознавания образов. Тр. ИМ СО АН СССР. - "Вычислительные системы", вып.22, Новосибирск, 1966.
2. Н.Г. Загоруйко. Современное состояние проблемы распознавания образов. Тр. ИМ СО АН СССР, -"Вычислительные системы", вып. 28, Новосибирск, 1967.
3. Г.С. СЕБЕСТЬЯН. Процессы принятия решений при распознавании образов. Пер. с англ. Изд-во "Техника", г.Киев, 1965.
4. Г.С. ДЕОВ. Выбор эффективной системы зависимых признаков. Тр. ИМ СО АН СССР, - "Вычислительные системы", вып.19, Новосибирск, 1965.
5. Н.Г. ЗАГОРУЙКО, В.Н. ЕЛКИНА. Количественные критерии качества таксономии и их использование в процессе принятия решений. (данный сборник).
6. Н.Г. ЗАГОРУЙКО, В.Н. ЕЛКИНА. Алфавит с минимальной избыточностью. Тр. ИМ СО АН СССР, - "Вычислительные системы", вып. 28, Новосибирск, 1967.

Поступила в редакцию
6.I.1969г.