ОБНАРУЖЕНИЕ ЭМПИРИЧЕСКИХ ЗАКОНОМЕРНОСТЕЙ
(Вычислительные системы)
1999 год Выпуск 166
УДК 519.95

ТАКСОНОМИЯ И РАСПОЗНАВАНИЕ В λ ПРОСТРАНСТВАХ С ИСПОЛЬЗОВАНИЕМ КОНЦЕПТОВ[1]

Н.Г. Загоруйко, И.А. Борисова, И.Н. Сунина

§1. Таксономия в λ-пространствах

При разработке критерия качества таксономии было обнаружено, что человек делает классификацию, основываясь не только на оценке евклидовых расстояний [1]. При одном и том же евклидовом расстоянии между двумя точками он может считать их "близкими" и включить в один таксон, но может считать и "далекими" и отнести их к разным таксонам. Все зависит от того, как расположены по отношению к этим точкам соседние с ними точки.

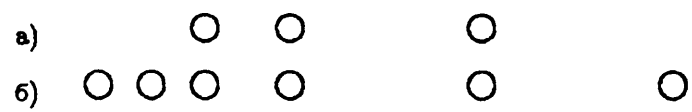

Рис. 1.

[1]Работа выполнена при частичной поддержке гранта РФФИ № 99-01-00582, гранта Миннауки № 0201.05.283 и гранта программы ИНТЕГРАЦИЯ.

Так, при разделении объектов на два таксона, средний объект на рис.1,а человек обычно относит к левому таксону, а тот же объект на рис.1,б к правому. Зрительный анализатор человека, решая задачу таксономии, работает не в евклидовом пространстве, а в λ–пространстве, свойства которого состоят в следующем.

Соединим объекты ребрами кратчайшего незамкнутого пути. Выделим ребро, соединяющее два объекта a и b, и все смежные ему ребра. Среди смежных, найдем ребро минимальной длины $\beta_{\min}$. Если евклидово расстояние между точками a и b равно α, то λ расстояние между ними является функцией от двух аргументов α и величины $t = \alpha/\beta_{\min}$, характеризующей локальный скачок плотности точек: $\lambda = f(\alpha, t)$. В экспериментах со зрительным восприятием выяснилось, что эти аргументы человек учитывает с разными весами. Так, если длины всех ребер графа и величины t нормировать по отношению к их самым большим значениям, то величина λ расстояния определяется как $\lambda = \alpha * t^2$. Дополнительно к этому человек стремится сделать разбиение множества объектов на k таксонов одинаковой мощности. Если количество объектов в j–м таксоне равно l_j и общее число объектов $L = \sum_{j=1}^{k} l_j$, то равномощность оценивается величиной $h = k^k * \prod_{j=1}^{k} \frac{l_j}{L}$, принимающей значения в пределах от 0 до 1. Алгоритм таксономии λ–KRAB [1] максимизирует критерий качества такого вида: $F = \alpha * t^2 * h^4$. При этом получаются таксоны произвольной формы, граница между таксонами может быть сколь угодно сложной. Важно то, что в тех случаях, когда машинное решение можно сравнить с решением, полученным человеком "вручную", т.е. в пространстве малой размерности, эти решения обычно совпадают. В пространстве большой размерности, когда человек не может пользоваться зрительным восприятием, он переходит на примитивные методы таксономии и формирует таксоны в виде гиперпараллелепипедов, указывая минимальную и максимальную границы по каждой координате. Программа же, опираясь на приведенный выше критерий качества, как бы распространяет уникальные свойства зрительного анализатора на пространство любой размерности. Примеры решения некоторых двумерных задач концептуальной таксономии показаны на рис. 2,3 и 4.

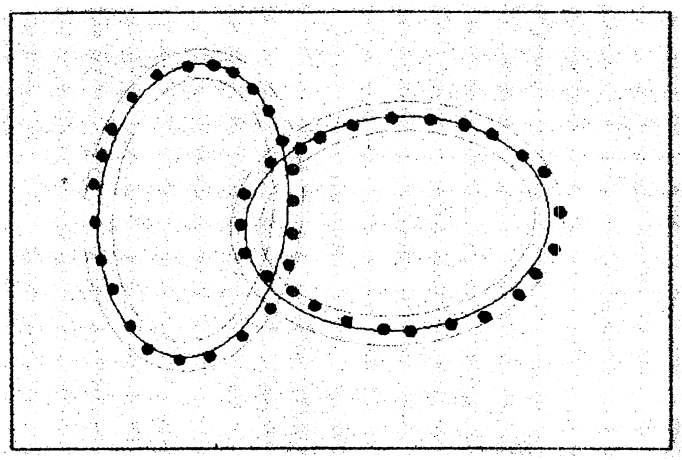

Рис. 2

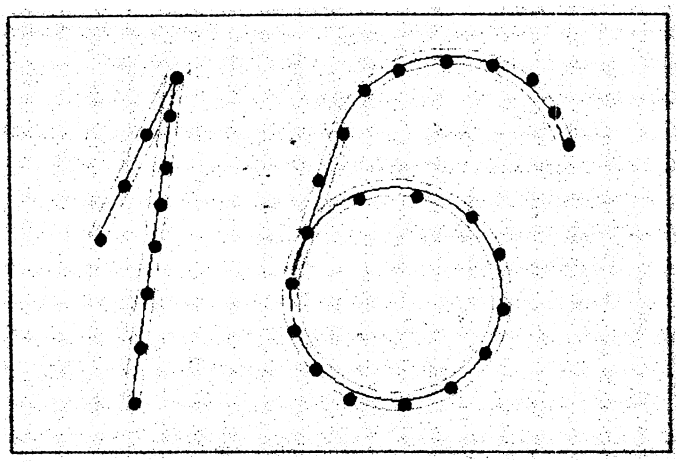

Рис. 3

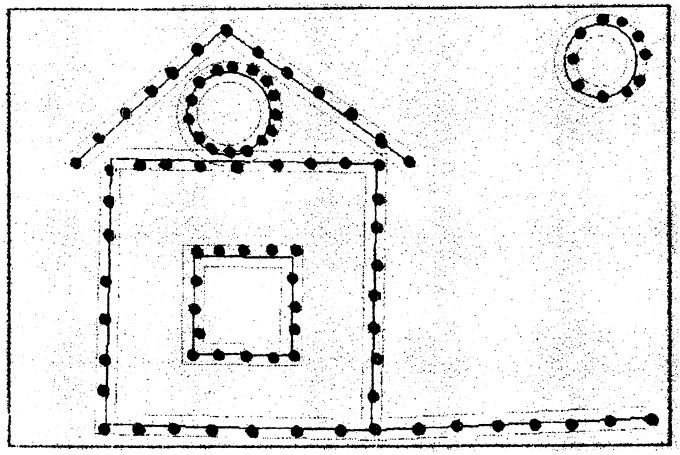

Рис. 4

Алгоритм естественным образом обобщается на многомерный случай, где в качестве полосы также рассматривается множество точек, отстоящих от n–мерной фигуры на расстоянии не более h.

§2. Таксономия с использованием концептов

Дальнейшее сближение машинной таксономии с человеческой достигается при использовании для описания таксонов так называемых "концептов" [2]. Под "концептом" понимается элемент языка, предназначенного для обобщенного описания некоторого множества объектов. Это описание должно быть хорошо знакомо человеку, легко им восприниматься и запоминаться. При таксономии в метрическом пространстве человеку удобно оперировать "концептами" в виде простых для восприятия геометрических объектов: точек, прямых линий, сфер, эллипсоидов, поверхностей первого или второго порядков, регулярных решеток и т.п. Элементы языка геометрических фигур могут быть упорядочены по мере нарастания сложности. Сложность геометрического объекта будем характеризовать числом точек, достаточным для

однозначной идентификации объекта. Так, точка характеризуется сама собой, отрезок прямой определяется двумя точками, расположенными на его концах. Концепты "окружность" и "круг" можно задать двумя точками, лежащими на концах их радиусов. Чтобы выделить множество точек, распределение которых подчиняется нормальному закону с единичной матрицей ковариаций, достаточно указать два параметра: координату математического ожидания и значение дисперсии. Для построения на плоскости концептов "треугольник", "эллипс", "прямоугольник", "правильный n–угольник" и кривая второго порядка достаточно знать координаты трех характерных точек.

Если вокруг линии, образующей геометрическую фигуру, выделить коридор шириной h, то образуется "полоса" площадью S. Точки, попавшие в эту полосу, считаются входящими в концепт данной фигуры. Понятно, что чем меньше S, тем точнее содержание данного концепта описывается лежащей в его основе фигурой. В качестве самого грубого варианта будем рассматривать концепт с шириной полосы h_{max}, равной диаметру выборки D, где D — расстояние между двумя самыми далекими друг от друга точками выборки.

Если j–м концептом, $j = 1, 2, \ldots, m$, покрыто подмножество, состоящее из l_j точек, то с уменьшением h_j площадь фигуры S уменьшается, плотность точек и качество концептуального описания повышается, так что количественное значение этого качества (назовем его "качеством концепта") можно измерять величиной $Q = \frac{l_j}{S + c}$, где c — константа. В идеальном случае, когда все L точек выборки лежат на линиях, образующих фигуру некоторого одного концепта, его качество достигает максимального значения $Q = \frac{l_j}{c}$. В другом крайнем случае, когда $h_j = D$, площадь фигуры максимальна и качество описания Q достигает минимума. Если множество из L объектов (точек) разделено на k таксонов (покрыто k концептами), то качество такой таксономии можно определять величиной $Q = (\sum_{j=1}^{k} Q_j)/k$. Используя эти обозначения, опишем схему алгоритма концептуальной таксономии (алгоритм ConTax).

Возможно два подхода к решению задачи концептуальной таксономии. Первый подход (переборный алгоритм $C\lambda$-KRAB) состоит в следующем.

1. Случайным образом выбирается n_i точек в качестве характерных для концепта заданного (i-го) типа.

2. Оценивается качество Q_{1i} покрытия этим концептом подмножества l_1 точек при разных значениях ширины полосы h_1.

3. Выявляется концепт с максимальным значением качества.

4. Процедуры 1–3 повторяются для всех m типов концептов.

5. Фиксируется концепт с максимальным значением критерия качества Q_{1i}. Точки, покрытые этим концептом, объявляются принадлежащими 1-му таксону i-го типа и из дальнейшего рассмотрения исключаются.

6. На оставшихся точках пп.1–5 повторяются до полного исчерпания всех L точек.

Получение заданного количества k таксонов можно обеспечить, меняя предельное значение ширины полосы ($h_{\max}$). Чем больше $h_{\max}$, тем меньше число получаемых таксонов.

Трудоемкость самого сложного первого прохода по пп.1–5 при максимальной сложности концепта $n_{\max}$ имеет порядок величины $T_1 = w*v*m*L^{n_{\max}}$, w — количество разных значений ширины полосы, а v — размерность пространства признаков.

Второй подход (последовательный алгоритм $C\lambda$-KRAB1) состоит в предварительном поиске варианта таксономии с помощью алгоритма λ-KRAB и последующем описании полученных таксонов подходящим набором концептов. Этот подход ориентирован на большое число объектов L, так как трудоемкость описания k таксонов концептами меньше трудоемкости поиска концептуальных таксонов в $k^{n_{\max}}$ раз. Затраты на таксономию по алгоритму λ-KRAB приблизительно равны $T_2 = v*(k*L)^2$, следовательно суммарные затраты на концептуальную последовательную таксономию равны $T_3 = T_2 + \dfrac{T_1}{k^{n_{\max}}}$.

Анализ отношения затрат $R = \dfrac{T_1}{T_3}$ показывает, что при $n_{\max} = 1$ затраты машинных ресурсов на переборный алгоритм меньше, чем на последовательный. При $n_{\max} = 2$ эти алгоритмы имеют приблизительно одинаковую трудоемкость. Если же

$n_{max} > 2$, то целесообразно пользоваться последовательным алгоритмом.

Существует реализация описанного алгоритма $C\lambda$-KRAB на языке Java в виде апплета. В библиотеке концептов представлены геометрические концепты с числом характерных точек n=1 и n=2. В двумерном случае программа выдает решения, "естественные" для человека, и успешно решает задачи, которые приводятся в качестве примеров задач, якобы неразрешимых для алгоритмов таксономии [3]. Алгоритм естественным образом обобщается на многомерный случай, где в качестве полосы также рассматривается множество точек, отстоящих от n-мерной фигуры на расстоянии не более $\frac{h}{2}$.

§3. Таксономические решающие правила

Задача распознавания образов в традиционной постановке состоит в поиске решающего правила D, которое обеспечивало бы минимум функции затрат N, зависящих от стоимости ошибок распознавания R и сложности реализации решающего правила C. Начальные условия включают в себя фиксированный набор признаков X, алфавит распознаваемых образов S и конечную обучающую выборку A_0. После добавления к этой информации предположений H о законе распределения генеральных совокупностей, о характере зависимости между признаками, о степени представительности обучающей выборки и т.д. выбирается некоторый класс решающих функций F и находится наиболее подходящее правило из этого класса: $f = \arg\min_{f \in F} N(D_f)/X, S, A_0, H.$

И впоследствии, какие бы реализации ни предъявлялись для распознавания, выбранное решающее правило остается неизменным.

Между тем, информация о свойствах объектов контрольной выборки A_K могла бы дополнить представление о генеральной совокупности, сформированное обучающей выборкой, и улучшить качество распознавания. При этом решающая функция должна определяться из следующих условий: $f = \arg\min_{f \in F} N(D_f)/X, S, A_0, A_K, H.$

Задача в этой постановке решается с помощью таксономических решающих функций [1,3], сущность которых состоит в следующем. При распознавании k образов смесь обучающих и контрольных объектов подвергается таксономии на k таксонов с помощью того или иного алгоритма таксономии. Если в некотором таксоне оказались точки обучающей выборки только одного i-го образа, то все контрольные точки, попавшие в этот таксон, относятся также к образу i. Если в таксоне есть точки из k' разных образов, то возможны два варианта дальнейших действий.

При первом варианте считается, что все объекты обучающей выборки должны быть распознаны без ошибок и потому наличие представителей разных образов в одном таксоне недопустимо. Такой таксон разбивается на k' более мелких таксонов. Эта процедура продолжается до тех пор, пока в каждом таксоне не окажутся обучающие точки только одного образа. Сложность полученного решающего правила (т.е. число таксонов K) не принимается во внимание.

Второй вариант поведения основан на учете хорошо известного факта, состоящего в том, что чрезмерное усложнение решающего правила в погоне за правильным распознаванием обучающей выборки часто ведет к росту числа ошибок при распознавании контрольной выборки. В связи с этим таксоны со смесью реализаций разных образов не всегда подвергаются дальнейшему дроблению. Так, если в некотором таксоне количество объектов обучающей выборки i-го образа существенно больше чем объектов других образов, то все обучающие и контрольные объекты, попавшие в этот неоднородный таксон, считаются принадлежащими образу i. Ясно, что чем однороднее будут таксоны, тем меньше ошибок распознавания. Исходя из этого, процесс таксономии смеси обучающей и контрольной выборок должен управлять критерием, одним из аргументов которого является однородность таксонов.

Введем меру однородности таксона q. Если в j-м таксоне среди N_j объектов обучающей выборки есть n_i представителей каждого из k образов ($N_j = \sum_{i=1}^{k} n_i$), то, принимая доли p_i каждого

образа в этой смеси равными $p_i = \frac{n_i}{N_j}$, найдем меру неоднородности (энтропию) таксона: $E_j = -\sum_{i=1}^{k} p_i * \log p_i$.

Наибольшая неоднородность будет в случае, когда все образы представлены в таксоне одинаковым числом объектов. Тогда $E'_j = -\log\frac{1}{k}$. Если же в таксоне есть представители только одного образа, то $E_j = 0$. Однородность j–го таксона определяется следующим соотношением: $g_j = 1 - \frac{E_j}{E'_j}$. Однородность K таксонов (G) будем оценивать средневзвешенной мерой однородности: $G = (\sum_{j=1}^{K} g_j * l_j)/L$, где L — общее число объектов выборки, а l_j — число объектов в j–м таксоне.

С учетом сказанного на каждом шаге процесса таксономии, как и в алгоритме $C\lambda$–KRAB1, строится кратчайший незамкнутый путь между всеми объектами и находится то ребро, при разрыве которого получается два таксона с максимальным значением меры однородности G. Эти таксоны описываются на языке концептов с максимально возможным их качеством Q. Качество F полученного таксономического решающего правила оценивается величиной $F = f(Q * G)$. Затем из двух полученных таксонов выбирается тот, разбиение которого на два более мелких таксона обеспечивает наибольшее приращение величины F. Процедура дробления таксонов продолжается до выполнения одного из двух условий: либо число таксонов K достигло предельно допустимой величины K^*, либо на очередном шаге не удается найти вариант разбиения, который бы увеличивал значение критерия F.

Если обучающий материал представлен информацией о законах распределения генеральных совокупностей пересекающихся образов, то оптимальная (байесова) граница проводится по точкам с равной плотностью вероятности этих образов. При этом часть объектов в области пересечения распознается неправильно, но общие потери от ошибок минимальны. Второй вариант, описанный выше, фактически представляет собой дискретный вариант байесовой стратегии: редкие объекты любого образа, оказавшиеся в j–м таксоне в плотном окружении объектов i–го

образа, объявляются представителями образа i, если общая мера однородности G велика и качество концептуального описания полученных таксонов Q достаточно велики.

Можно отметить, что так мы действуем всегда, когда есть возможность использовать те или иные обобщающие понятия (концепты): индивидуальные особенности объектов игнорируются, если они противоречат легко и просто формулируемой закономерности. Яркой демонстрацией плодотворности использования гипотезы простоты служит периодический закон Менделеева, который утверждал наличие строгой периодической зависимости между свойствами химических элементов и их атомными весами, игнорируя тот факт, что эта зависимость нарушается в нескольких местах таблицы. После изучения внутренней структуры атомов стало ясно, что свойства элементов периодически зависят от числа электронов во внешних слоях их оболочек, и эта зависимость строго соблюдается именно при том расположении элементов в таблице, какое сделал Д.И. Менделеев.

В итоге работы описанного выше алгоритма построения Концептуальных Таксономических Решающих Функций (КТРФ) мы получаем перечень концептов с указанием того, к какому образу относятся объекты, входящие в состав каждого из них. Концепт, не содержащий объектов обучающей выборки, может быть присоединен к ближайшему концепту или считаться принадлежащим новому $(k+1)$-му образу, не представленному в обучающем материале.

Применение таксономических решающих правил повышает устойчивость решения по отношению к такому часто встречающемуся явлению, как непредставительность обучающей выборки. Концептуальная форма представления этих правил полностью соответствует требованиям, предъявляемым к современным методам анализа данных (Data Mining) [1,4], в соответствии с которыми результат решения задачи должен иметь вид, удобный для использования машиной и понятный человеку.

Литература:

1. ЗАГОРУЙКО Н.Г. Прикладные методы анализа данных и знаний. Новосибирск: Изд. ИМ СО РАН, 1999.

2. МИХАЛЬСКИЙ Р, СТЕПП Р (Michalski R.S., Stepp R.) Learning from observation: conceptual clustering // Machine learning: An artificial intelligence approach. Morgan Caufman. 1983.

3. ЁЛКИНА В.Н., ЗАГОРУЙКО Н.Г. Количественные критерии качества таксономии и их использование в процессе принятия решений //Вычислительные системы. Вып. 36. — Новосибирск, 1969. — С. 29–46.

4. МИКИ Д. (Michie D.) Machine learning in the next five year //EWLS-88; Proc. 3-th Europ. working session on learning. Glasgow; London: Pitman, 1988.

Поступила в редакцию
29 февраля 2000 года