

МЕТОДЫ КЛАССИФИКАЦИИ СОСТОЯНИЙ ОБЪЕКТА В ЭКСПЕРТНЫХ СИСТЕМАХ

А.М. Курносов

(представил д.т.н., проф. Е.И. Бобыр)

Проведен сравнительный анализ методов кластерного анализа, используемых при классификации состояний объекта в экспертных системах.

Состояние всякого объекта или предметной области можно охарактеризовать значениями некоторых признаков или параметров. Если множество состояний объекта обладает общими свойствами или значениями признаков описания состояний объекта по их схожести, то процесс классификации можно представить в виде двух этапов: изучение и установление общих свойств состояний одного и того же класса или выделение класса; принятие решения, устанавливающего принадлежность нового, ранее в процессе классификации не рассматриваемого, входного состояния к данному классу путем сравнения значений признаков (свойств) этого состояния с уже известным состоянием, присущим данному классу.

Существует несколько методов задания классов состояний объекта, в частности, метод перечисления членов класса, метод общности свойств, метод кластеризации. Задание класса перечислением состояний, входящих в его состав, заключается в реализации процедуры сравнения с эталоном. Однако в тех случаях, когда классы состоят из большого числа элементов и каждый элемент описывается значениями большого числа признаков, реализация этого принципа сильно затруднена.

Задание класса при помощи свойств, общих для всех состояний данного класса, заключается в выделении признаков описания состояний объекта и работы с ними. Поскольку значения признаков, характеризующих класс, находятся в некотором диапазоне, метод сопоставления признаков допускает некоторое отличие значений признаков отдельных состояний, чего не допускает процедура сравнения с эталоном. Исключительно трудно, если невозможно найти для некоторого класса полный набор различающих признаков, что делает применение данного метода достаточно сложным.

Когда состояние объекта можно представить в виде вектора чисел, определяющего геометрическое расположение состояний в пространстве, координатами которого являются признаки описания состояния, говорят о кластеризации состояния. Кластером называют группу состояний объекта, образующих в пространстве описаний компактную в не-

котором смысле область. Реализация метода кластеризации в экспертных системах (ЭС) определяется взаимным пространственным расположением кластеров в пространстве. Если кластеры, соответствующие разным классам, разнесены достаточно далеко друг от друга, то можно воспользоваться классификацией по какой-либо из метрик. Если кластеры перекрываются, что происходит в результате неполной информации о состоянии объекта, то необходимо применять более сложные методы разбиения пространства состояний.

Рассмотрим методы кластерного анализа. Пусть состояние объекта имеет размерность N и представляет собой точку в N -мерном пространстве. Каждая координата в пространстве соответствует одному из признаков описания состояния объекта.

Одной из основных задач классификации является определение многомерных областей в пространстве состояний, в которых содержатся состояния, принадлежащие к различным классам. Пространство описаний состояний может быть преобразовано из N -мерного в пространство меньшей размерности с определенной потерей информации. Этот переход осуществляется при помощи взвешивания каждого признака. Для этого всем используемым признакам приписывается вес, больший нуля, а неиспользуемым – вес, равный нулю. Меры сходства, используемые в кластерном анализе, как правило, имеют вид

$$d(S_i, S_j) = \left(\sum_{k=1}^n |a_{ik} - a_{jk}|^r / n \right)^{1/r},$$

где S_i, S_j - возможные состояния объекта; a_{ik}, a_{jk} - значения признака описания состояний S_i, S_j ; n - число признаков; r - положительное целое число.

Расстояние является функцией, отвечающей интуитивному пониманию экспертом сходства объектов. Тем не менее, определение подобия через расстояние между состояниями не является единственно возможным. Главное заключается в том, чтобы подобрать такую меру, при помощи которой можно принять правильное (с точки зрения эксперта) решение о принадлежности или не принадлежности состояния некоторому классу. В качестве меры сходства, применяемой в кластерном анализе, может быть использован, например, коэффициент общего сходства Гауэра

$$S_{jk} = \sum_{i=1}^n W_{ijk} S_{ijk} / \sum_{i=1}^n W_{ijk},$$

где $0 \leq S_{ijk} \leq 1$ - значение сходства между значениями признака i для состояний j и k ; W_{ijk} - вес, приписываемый признаку; n - число признаков.

При классификации состояний в качестве меры сходства, помимо рассматриваемых, могут использоваться и другие меры, например, Махаланобиса, Танимото и другие. Среди параметров, используемых в различных методах кластерного анализа, можно выделить следующие: плот-

ность состояний в пространстве описаний состояний объекта; объем, занимаемый кластером; связность элементов определенного кластера; промежутки между соседними кластерами в сравнении с их диаметрами.

Критерием оптимальности произведенной классификации состояний может служить функция качества

$$\sum_k \sum_t \sum_n (x_{ij} - \overline{x_{iM}})^2,$$

где k - число кластеров; t_m - число состояний в кластере M ; n - число признаков; x_{ij} - значение i -го признака j - го состояния; $\overline{x_{iM}}$ - среднее значение i - го признака для кластера M .

Например, объединением кластеров, дающих наименьший вклад в получаемое значение качества, достигается оптимальность данного критерия для сгруппированных таким образом кластеров (алгоритм Уорда).

Целью всех алгоритмов формирования кластеров (без учителя) помимо собственно формирования кластеров является нахождение для каждого кластера так называемого центра, являющегося самым представительным эталоном кластера. При описании алгоритмов в качестве меры сходства используем евклидово расстояние. Одним из самых простых алгоритмов кластеризации является следующий.

Пусть задано множество состояний $S_1, S_2, S_3, \dots, S_n$. Центром первого кластера принимается любое состояние S_i из заданного множества. Пороговой величиной, относительно которой можно судить о принадлежности состояния к тому или иному кластеру, является положительное число T из интервала $(0,1)$. Для простоты полагаем центром первого кластера состояние S_1 . Вычисляем степень сходства между центром первого кластера S_1 и состоянием S_2 . Если полученное значение меньше T (в том случае, когда в качестве меры сходства используется мера расстояния, иначе – больше), то состояние S_2 включается в первый кластер. Иначе состояние S_2 становится центром нового (второго) кластера. Такая же процедура производится и для остальных состояний из заданного множества. Если значение меры сходства очередного состояния S_i и каждого определенного ранее центра кластера превышает значение T , то S_i становится центром нового кластера. Иначе состояние S_i включается в состав кластера с наиболее сходным к состоянию S_i центром.

Недостатком данного алгоритма является зависимость получаемой классификации состояний от выбора первого центра кластера, а также от значения порога T . Известны другие алгоритмы, например, максиминного расстояния, внутригрупповых средних, Изодата. Эти алгоритмы, как и некоторые другие, представляют собой итерационные процедуры.

Рассмотрим возможность классификации текущих состояний объекта в ЭС. Пусть в ЭС при помощи одного из алгоритмов кластерного анализа определено K кластеров с центрами Z_1, Z_2, \dots, Z_k . Тогда процесс классификации текущего состояния объекта S_0 , подаваемого на

вход ЭС, заключается в сопоставлении S_0 с каждым из этих центров и принятии решения о принадлежности S_0 тому кластеру, расстояние с которым минимально. При этом данное расстояние не должно превышать некоторого порога T , определяемого заранее. Если минимальное расстояние превышает T , то вводится новый кластер с центром, совпадающим с входным состоянием. Это указывает на то, что обучающая выборка, сформированная для задания кластера на этапе обучения, была не полна и не включала состояния, сходные поступившему на вход ЭС состоянию. Можно отметить, что задание порога T может существенным образом повлиять на результаты классификации.

Пусть в процессе кластерного анализа было построено нечеткое покрытие пространства состояний объекта нечеткими кластерами F_1, \dots, F_k и на вход ЭС подается текущее состояние объекта S_0 . В процессе распознавания S_0 определяется множество значений $\mu_{01}, \dots, \mu_{0k}$, характеризующее соответствие состояния S_0 каждому кластеру F_i , $i = \overline{1, k}$. Если получаемые значения μ_{0i} приводят к выполнению соотношения $\mu_{01} + \dots + \mu_{0k} > 1$ или $\max_i \mu_{0i} \geq T$, где i - номер кластера, T - некоторый порог, то принимается решение о создании нового нечеткого кластера F_{k+1} , центром которого является S_0 . Если в процессе классификации текущего состояния объекта S_0 был сформирован новый кластер $k + 1$ и каждому i - му кластеру ($i = \overline{1, k}$) экспертом было сформировано управляющее воздействие на объект, то состояние S_0 предъявляется эксперту для определения управляющего воздействия, соответствующего данному состоянию.

Учет неполноты информации о состоянии объекта (ситуации, в которых находится объект) заключается в следующем. Пусть для признаков описания состояния объекта X_1, \dots, X_q неизвестны их значения в некоторой ситуации S_i . Тогда соответствующий элемент a_i^k , характеризующий степень сходства ситуаций S_i и центра кластера Z_j имеет значение 0.5, что означает неопределенность в отношении этого признака при сравнении S_i и Z_j . Чем больше число неопределенных признаков q и выше вес признаков ω_k , тем ниже качество классификации состояний.

ЛИТЕРАТУРА

1. Ту Дж., Гонсалес Р. Принципы распознавания образов. – М.: Мир, 1978. – 415 с.
2. Сокал Р.Р. Кластер-анализ и классификация: предпосылки и основные направления // Классификация и кластер. – М.: Мир, 1980. – С. 52 - 73.
3. Дуда Р., Харт П. Распознавание образов и анализ сцен. – М.: Мир, 1976. – 328 с.

Поступила в редколлегию 21.05.2001