

Основные понятия и определения

- Объект (образ) и признак (свойство)
- Класс (кластер)
- Представление исходной информации
- Представление результатов классификации

В силу чисто исторических причин класс задач распознавания связан с понятием “образа”. В свое время не обратили внимания, что в заимствованном из англоязычных работ термине “pattern recognition” термин “pattern”, кроме значения “образ”, имеет еще значение “модель”, “стиль”, “режим”, “закономерность”, “образ действия”. В современном распознавании и особенно искусственном интеллекте его употребляют в самом широком смысле, имея в виду, что “образ” – это некоторое структурированное приближенное (обратите внимание – “приближенное”!) описание (эскиз) изучаемого объекта, явления или процесса.

То есть, частичная определенность описания является принципиальным свойством образа.

Основное назначение описаний (образов) – это их использование в процессе установления соответствия объектов, то есть при доказательстве их идентичности, аналогичности, подобия, сходства и т.п., которое осуществляется путем сравнения (сопоставления). Два образа считаются подобными, если удастся установить их соответствие. Можно, в частности, считать, что имеет место соответствие, если достигнута их идентичность.

Сопоставление образов представляет собой основную задачу распознавания и играет существенную роль в информатике в целом. Эта задача возникает, в частности, в различных разделах искусственного интеллекта, например в понимании естественного языка компьютером, символьной обработке алгебраических выражений, экспертных системах, преобразовании и синтезе программ ЭВМ.

Теперь отметим следующий важный момент, что в различных задачах образу придается различный смысл. Это определяется часто тем, какие характеристики объекта входят в описание образа, какой аппарат используется для представления этих характеристик. Именно отсюда и можно понять, почему образ является приближенным описанием объекта. Чем большее число свойств и качеств объекта отражено на принятом языке в образе рассматриваемого объекта, тем полнее это описание, тем полнее этот образ характеризует описываемый объект. Однако в любом случае мы имеем дело с описанием, а не с самим объектом, который всегда богаче описания. Итак, любой образ представляется некоторым набором признаков. Поэтому вполне допустимо наряду с выражением “распознавание образов” применять выражение “отождествление некоторых наборов описаний объектов”.

Пусть S множество, элементы которого S_i называются допустимыми объектами (образами), т.е. $S = \{S_i\}$. Для каждого S_i определено описание объекта $I(S_i) = (i_1, i_2, i_3, \dots, i_n)$, где i_j – значение признаков измеренных в соответствии выбранному пространству признаков.

Определение 1.1. Объект (образ) – это сущность, содержание которого должно быть определено набором характеризующих признаков (свойств).

Объект обладает: состоянием (совокупностью характеризующих признаков); индивидуальностью (способом выделения объекта среди других – может быть подобных).

Таким образом, сущность задается совокупностью признаков, а их значения определяют состояния объекта.

Под объектом могут отображаться множества объектов, процессов или явлений материального и нематериального происхождения.

Другими словами объект (образ) – это наименование области в пространстве признаков.

В зависимости от цели можно выбрать разные признаковые пространства или разные подобласти в этих признаковых пространствах, в результате чего одна и та же пара объектов может быть отнесена как к одному и тому же, так и к различным классам.

В геометрической интерпретации под объектом понимается точка n - мерного пространства, вдоль координатных осей которого отложены значения признаков (параметров).

Определение 1.2. Признаком называется количественное или качественное значение того или иного свойства (параметра) исследуемого объекта (образа) или явления.

Признаки делятся на качественные и количественные.

Качественные признаки отражают наличие, степень наличия или отсутствие того или иного свойства у изучаемого объекта или явления. Выражаются обычно они в двоичной форме “да” или “нет”. Область значений признаков пространства при такой системе описания представляет собой n - мерный двоичный куб, расстояние между вершинами которого равно числу несовпадающих разрядов, соответствующих n - разрядных двоичных векторов.

Количественные признаки, измеряемые приборами со шкалами “порядка”, “отношении”, “интервалов” отражают числовую характеристику степени проявления свойств объекта или явления, измеренных в соответствующих единицах измерения.

В некоторых задачах при определении объектов встречается сочетание как качественных, так и количественных признаков.

В качестве признаков медицинских объектов служат различные симптомы заболевания, показатели биохимического анализа, в объектах геологии - различные параметры-свойства, характеризующие подземный грунт, в социологии - показатели социально-экономического развития, производственных отношений.

Определение 1.3. Под классом образов (объектов) понимается некоторая категория, определяющаяся рядом свойств, общих для всех ее объектов.

Класс определяется некоторыми общими свойствами, присущими всем его объектам, и характеризуется перечнем входящих в него объектов. Таким образом, класс – множество объектов, связанных общностью структуры и состоянием. Часто синонимом класса используются термины: кластер, сгусток, скопление, таксоны или однородность.

В геометрической интерпретации класс представляет собой некоторое подмножество n – мерного евклидова пространства, объединяющее некоторые близкие (однородные) свойства объектов, где по осям координат отложены значения признаков.

Представление исходной информации

Пусть S – множество допустимых объектов. Объекты $S_i \in S$ задаются значениями некоторых признаков x_i , $i=1,2,...,n$ (этот набор признаков всегда один и тот же для всех объектов, рассматриваемых при решении определенной задачи). Совокупность значений признаков x_i определяет описание $I(S)$ объекта S . Каждый из признаков может принимать значения из различных множеств допустимых значений признаков, рассматриваемых в §1.10 данного учебника.

Определение 1.4. Описание $I(S)=(x_1(S), x_2(S),..., x_n(S))$ называют стандартным, если $x_i(S)$ принимает значение из множества допустимых значений.

Если задана обучающая информация $I(K_1, K_2,..., K_l)$ о классах, т.е. классы K_j описаны с помощью объектов, входящих в него, тогда априорная информация в задаче распознавания с непересекающимися классами часто задается в виде так называемой таблицы обучения $T_{n,n}$ (табл. 1.1).

Очевидно, что объекты $S_1, S_2, ..., S_{t_1}$ принадлежат классу K_1 , объекты $S_{t_1+1}, S_{t_1+2}, ..., S_{t_2}$ принадлежат классу K_2 и объекты $S_{t_m}, S_{t_m+1}, ..., S_m$ – классу K_l .

Таблица 1.1

Стандартная форма задания обучающей информации $T_{n,n}$

Объекты	Признаки и их значения						Классы
	1	2	...	i	...	n	
S_1	$x_{1,1}$	$x_{1,2}$	...	$x_{1,i}$	...	$x_{1,n}$	K_1
S_2	$x_{2,1}$	$x_{2,2}$		$x_{2,i}$	...	$x_{2,n}$	

			...				
...	...	...	...	...	...	...	
S_{t1}	$t_{1,1}$	$t_{1,2}$	...	$t_{1,j}$	...	$t_{1,n}$	
...	...	...	...	...	...	...	
S_{t1+1}	$t_{1+1,1}$	$t_{1+1,2}$	...	$t_{1+1,j}$	...	$t_{1+1,n}$	K_2
S_{t1+2}	$t_{1+2,1}$	$t_{1+2,2}$	...	$t_{1+2,j}$	...	$t_{1+2,n}$	
...	...	...	...	...	...	...	
S_{t2}	$t_{2,1}$	$t_{2,2}$	...	$t_{2,j}$	...	$t_{2,n}$	
...	...	...	...	...	...	...	
S_{tm}	$tm,1$	$tm,2$	...	tm,j	...	tm,n	K_l
S_{tm+1}	$tm+1,1$	$tm+1,2$	...	$tm+1,j$	...	$tm+1,n$	
...	...	...	...	...	...	...	
S_m	$m,1$	$m,2$	...	m,j	...	m,n	
S'	1	2	...	j	...	n	$K-?$

Последняя строка таблицы означает, что к распознаванию предъявлен новый объект S' , и требуется определить к какому классу принадлежит новый объект.

Если рассматривается задача классификации (кластерного анализа), то стандартная форма задания априорной информации будет аналогичной, но в таблице должен отсутствовать столбец “классы” с указанием классов, а также объекты должны представляться сплошным потоком, неразделенным на классы. Здесь перестановка местами объектов не должна влиять на сущность априорной информации.

Другими словами, стандартное представление априорной информации сводится в таблицу, строкой которой являются описания, столбцами – значения признаков в этих описаниях.

Информация о вхождении объекта S в класс K_i кодируется символами “1” ($S K_i$), “0” ($\bar{S} K_i$), “2” – неизвестно принадлежит ли объект S классу K_i или нет и записывается в виде так называемого информационного вектора

$$\bar{a}(S) = (a_1(S), a_2(S), \dots, a_l(S)), \quad (1.1)$$

где $a_i \in \{0, 1, 2\}$ и l – число классов.

Определение 1.5. Стандартной информацией $I(K_1, K_2, \dots, K_l)$ называют совокупность множеств $(I(S_1), I(S_2), \dots, I(S_m))$ и $(\bar{a}(S_1), \bar{a}(S_2), \dots, \bar{a}(S_m))$. Предполагается, что информационные вектора являются истинными.

Определение 1.6. Вектор $\bar{a}(S) = (a_1(S), a_2(S), \dots, a_l(S))$ называется истинным для S , если $a_j(S) \in \{0, 1\}$ и значение предиката определяется $P_j(S) = a_j(S)$, $j = 1, 2, \dots, l$.

Результаты классификации могут быть представлены в виде таблицы, аналогичной табл. 1.1 или могут выражаться через информационный вектор результатов.

Определение 1.7. Вектор

$$\bar{a}(S) = (a_1(S_1), a_2(S_2), \dots, a_m(S_m)) \quad (1.2)$$

называется информационным вектором результатов классификации.

Содержательно, значение $a_i(S_i)$ показывают номера классов, к которым принадлежат объекты S_i . Предполагается, что все значения элементов информационного вектора определены в рамках нумерации полученных классов.

Пусть в пространстве информационных векторов задана функция $\rho(\tilde{\alpha}, \tilde{\beta})$, обладающая всеми свойствами расстояния, за исключением, может быть, аксиомы треугольника. Задана также последовательность числовых функций $f_1(x_1), \dots, f_q(x_1, \dots, x_q), \dots$, причем:

- 1) $f_q(x_1, \dots, x_q)$ определена и неотрицательна при $x_i \geq 0$;
- 2) f_q монотонно не возрастает по каждой из переменных x_i ;
- 3) f_q достигает абсолютного максимума на наборе $(0, \dots, 0)$.

Пусть опять $S'_1, \dots, S'_q$ – произвольная конечная выборка из множества допустимых объектов, $\tilde{\alpha}(S'_i)$ – истинный для S'_i , $\alpha^A(S'_i)$ – информационный вектор S'_i в алгоритме A.

Определение 1.8. Функционалом качества алгоритма A на $S'_1, \dots, S'_q$ называется величина

$$f_q(\rho(\tilde{\alpha}(S'_1), \alpha^A(S'_1)), \dots, \rho(\tilde{\alpha}(S'_q), \alpha^A(S'_q))). \quad (1.3)$$

Функционал качества алгоритма A будем обозначать в дальнейшем через $\phi(A)$ или $\phi(A, S'_1, \dots, S'_q)$. В практических задачах рассматриваются обычно следующие виды $\phi(A)$.

1. Пусть для каждой пары (S'_i, K_j) , $S'_i \in M$, $1 \leq j \leq l$ определены числа $y_{ij}(\cdot, \cdot)$, $y_{ij}(\cdot, \cdot) \in \{0, 1\}$, причем $y_{ij}(\cdot, \cdot) = y_{ij}(\cdot, \cdot)$ при $y_{ij}(\cdot, \cdot) = y_{ij}(\cdot, \cdot)$, $y_{ij}(\alpha, \Delta) \geq y_{ij}(\tilde{\alpha}, \tilde{\beta})$, если $\alpha \in [0, 1]$, $\beta \in [0, 1]$, $\tilde{\alpha} \neq \beta$. Функционал $\phi(A)$ называется линейным [26], если

$$\phi(A) = \frac{1}{q \times l} \sum_{i=1}^q \sum_{j=1}^l y_{ij}(\alpha_{ij}, \alpha_{ij}^A). \quad (1.4)$$

2. Если $y_{ij}(1, 1) = y_{ij}(0, 0) = 1$ и все остальные $y_{ij} = 0$, то функционал $\phi(A)$ называется долей правильных прогнозов.