应用倒排文件进行模式匹配

阎 敬 伟

摘要 对模式识别技术的匹配方法提出了一种对参考模式特征数据的存贮结构，并给出了识别过程的算法。

一、引 言

随着科学技术的发展，模式识别已成为与高技术发展研究相联系的新兴学科之一，并与计算机技术相结合，出现了越来越多的新方法、新技术、与计算机软件技术、数据库系统，计算机硬件交织在一起，使模式识别技术的应用范围越来越广，对新技术，新方法的探求越来越重要。

二、模 式 与 识 别

所谓模式，按其性质可分为两类，一是抽象模式，如概念，心里等，这一类模式的识别问题通常属于人工智能的范畴，而这里要讨论的是另一类属具体模式，如图像，字符等，而模式的识别就是将未知模式正确地归类到已知的参考模式当中，采用匹配算法，其识别过程可用图 1 表示。

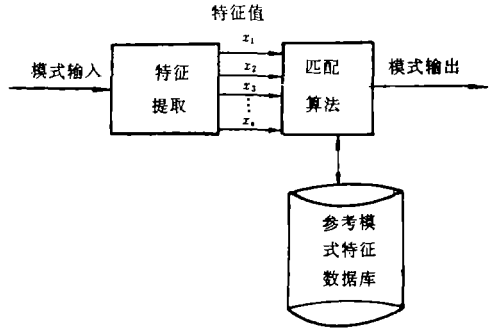

图 1

在参考模式的特征数据库中，存有正确反映参考模式的特征数据，当未知模式送入到计算机后，经过特征提取，取出与数据库中有相同属性的特征数据，检索数据库与参考模式进行匹配，找出未知模式应属类别。很显然，在匹配算法已经确定的情况下，高效准确地识别出未知模式，还取决于特征数据的存贮组织结构，使之能正确地检索出未知模式应归属的参考模式类别。故此，数据存贮结构的设计同样成为有效识别的关键技术之一。

三、倒排文件组织

在数据库的概念中，所谓倒排文件就是数据文件的非关键字索引文件，设

$$P = (A_1, A_2, \cdots, \cdots, A_n)$$

其中A_i $i = 1, 2, \cdots$， n是模式P的某一特征值，而 $(A_1, A_2, \cdots, A_n)$ 就形成了模式的特征向量，将所有的参考模式存到计算机中，就形成了一个关系型的数据文件。因为每种模式是唯一的，故可以将模式类别作为文件的主关键字，如表 1

该输入模式为 p，经特征提取后取得其特征向量，即有

$$p = (a_1, a_2, \cdots, a_n)$$

其中a_i的值域是A_i

表 1

P	A_1	A_2	……	A_n
			⋮	
P_i	A_{i1}	A_{i2}	……	A_{in}
			⋮	

识别时，即是要找出与 p 最接近的参考模式p_j，为寻找 p_j，我们须要用特征向量的分量 a_i 来检索数据库，这个过程就是用特征数据文件的次关键字A_i来检索数据文件，找出与 p 最接近的所有候选参考模式，因为这样的参考模式可能不只一个，而且如果参考模式有很多，将未知模式与文件中所有参考模式相匹配会很繁琐，而且数据运算量会很大，特别是在实时系统中将很不适合，所以对于不可能的参考模式不予考虑，选择候选模式是合理的，为此要对每一次关键字建立索引文件，即倒排文件。用倒排文件能够进行多码检索和用次关键字检索主关键字的优点，把对记录的查寻转化为对记录地址的集合运算，在识别算法中这不仅能提高识别速度，减少识别时的运算，而且如果用每一次关键字逐个限定候选参考模式的范围，能够恰当地选取候选参考模式。

为此，对每一次关键字A_i建立倒排索引文件，其结构如表 2：

表 2

Ai_{min}	Ai_{max}	P
⋮	⋮	⋮
B_i	B_{i+1}	$P_{i1}\ P_{i2}\cdots P_{im}$
B_{i+2}	B_{+i2}	$P_{j1}\ P_{j2}\cdots P_{jn}$
⋮	⋮	⋮

很显然这是对次关键字的分段索引，其 P_{ij} $j = 1, \cdots$是在此数据段内的所有参考模式，在实际存贮时，P_{ij}可以换成其在数据文件中该参考模式所在记录的地址。为提高识别的准确性，可以使 $B_{i+1} > B_{i+2}$，使段之间交叉在一起，这样可以利用合理的识别过程扩大候选的参考模式集合，不致于由于特征提取时的误差将真实的模式丢掉而产生误识别。

四、识别方法与识别过程

模式识别在匹配过程中，主要的两个方面就是匹配方法及识别过程，模式识别作为一个新学科分支发展至今天，其识别技术已较为成熟，而且方法多种多样，在此不予赘述，但象本文所述的数据结构一般适用于非参数决策方法，如最常用最小距离方法，在此方法中设

$$A_i = (A_1, A_2, \cdots, A_n)$$

表示与模式P_i相联系的特征向量，

$$a = (a_1, a_2, \cdots, a_n)$$

表示经特征提取后的未知模式的特征向量，此向量可以考虑成是 n 维特征空间的一个点，对此，设有 m 个候选的与未知模式最接近的参考模式，计算

$$D_i = |a - A_i| \qquad i = 1 \cdots m$$

取$D_i = \min\{D_i\}$ $i = 1 \cdots m$ 则D_j就是所要将未知模式归属的类。当然还有一些公式用来识别如线性识别公式，多项式识别公式等等，但其原理都是如此。

用此识别公式可以产生一个识别过程，设计此过程可以做如下考虑。

在接收到待识别的未知模式特征向量之后，要考虑到特征提取时产生的误差，为此在识别时先用每一单个的特征分量进行选择候选参考模式并尽可能扩大候选范围，然后再由每一

分量确定的所有参考模式集的交集运算减少那些不必要的参考模式，从而减少运算量，即是把部分识别运算转化为对数据文件的检索和集合的交集运算。

进行交集运算时，可能由于误差的原因使交集之后的集合产生空集，为此在利用特征向量的过程中首先要选用误差小，可信度高的特征值检索数据文件，当用特征值限定的参考模式达到一定的数量时，就不再用余下的特征值继续检索参考模式了，所以在查寻参考模式时，未必每个特征向量全部用到。

其识别过程可采用如下算法

1．输入特征向量 a；2．$j=1, i=2$；3．用第j个分量查索引文件，找到 a_j 所在的索引记录，并把它所包含的所有参考模式取到M_1中；4．a_j 同样在下个索引记录中吗？如果不是，则转到6；5．下个索引记录中所有参考模式与进行并集运算，结果存取到M_1中；6．$j=j+1$；7．用第j个特征分量检索索引文件，并将a_j 所在的索引记录中包含的参考模式取到M_2中；8．a_j在下个索引记录中吗？如果不在则转10；9．下个索引记录中所有参考模式与M_2取并集运算，结果仍送到M_2中；10．$M=M_1\cap M_2$；11．$|M|\leqslant N$吗？如果是则转15；12．$M_1=M$，$j=j+1$；13．$i=n$吗？如果是则转17；14．$i=i+1$转7；15．$|M|=0$吗？如果是则转17；16．将M中所有的参考模式取出其特征值与输入模式进行匹配，取最小值为识别结果，转18；17．将M_1中的所有参考模式取出其特征值与输入模式进行匹配运算，取最小值为识别结果；18．输出识别结果；19．继续输入识别吗？如果是则转1；20．退出。

五、算法分析

本文所述的内容已在 AST80286 系统机上实现。系统的时钟频率是10MHz，在此条件下，模拟雷达目标识别过程，其对单个目标的识别时间为0.06s，在不大于2%的误差下，效率可达93.3%以上，可以看出用此识别方法在速度和准确程度上都是令人满意的。其正确识别主要看参考模式间不差异，如果其差异较大，则对特征分量允许的误差也较大，而且识别的准确度高，而且算法也能顺利地应付误差较大的情况，对候选模式的选择也较为恰当。

参考文献

[1] 傅京孙；《模式识别》，北京大学出版社，1990

[2] 萨师煊、王珊；《数据库系统概论》，高等教育出版社，1985年

Applying Reverse File Sorting for Model Matching

Yan Jingwei

Abstract

This paper discurses the feature data structure of reference pattern in pattern recognition technology and an algorithm of identificatior is desigred.