基于信息熵的特征子集选择启发式算法的研究*

钱国良 舒文豪 陈 彬 权光日

(哈尔滨工业大学计算机科学与工程系 哈尔滨 150001)

E-mail: glqian@mlg.hit.edu.cn

摘要 特征子集选择问题是机器学习和模式识别中的一个重要问题.最优特征子集选择问题已被证明是NP难题.然而,目前的特征子集选择的启发式算法是基于正反例一致的,没有考虑到实际应用中的噪音数据影响,使得选择一个较好的特征子集非常困难.首先从统计学的角度分析了噪音对特征子集选择的影响,给出含有错误率的一致特征子集概念,然后利用信息熵和拉普拉斯错误估计函数构造了特征子集选择启发式算法EFS(entropy based feature subset selection).将该算法应用于两个实际领域的学习问题,并与GFS(greedy feature subset selection)算法进行了比较.实验结果表明,EFS选择的特征子集更具有代表性,较为有效地解决了实际应用中的噪音影响.

关键词 特征子集选择,机器学习,扩张矩阵,信息熵,噪音.

中图法分类号 TP18

特征子集选择FSS(feature subset selection)是机器学习和模式识别中非常困难而有意义的一个问题[1~3],FSS问题是指从一个大的候选特征集合中选择一个较好的、有代表性的子集,一致地描述已知例子集.由于在实际应用中,过多的特征会严重地影响归纳学习的质量,一些不必要的特征会使训练例子集噪音增大,损害所生成规则的精度.[4]因此,一个特征子集越小,所选特征可能越具有代表性,从而由此产生的规则质量越高.然而现已证明,最优(最小)特征子集选择OFSS(optimal feature subset selection)问题是NP困难问题.[1]因此,寻找一个较好的启发式算法是必然选择.然而,由于现有的特征子集选择启发式算法[1~4]是基于正反例一致的条件的精确学习方法,没有考虑到实际应用中的噪音数据影响,使得在具体应用中,很难选择出较好的特征子集.

噪音问题是实际应用中经常遇到的问题.通常在学习问题中,将正反例集中相同的例子定义为噪音[5],然而这种定义具有一定的片面性.因为在实际应用中,一些由于测量、记录、特征提取的错误所造成的数据偏差,不一定使正反例集中出现相同的例子,但都有可能给学习带来很大困难,而且特征的个数越多,产生噪音的机会越大.[6]本文所说的噪音不仅仅是指学习问题中定义的正反例集中存在相同例子的情况,而且从聚类角度来讲[7],那些使得生成的聚类概念所包含的非常少的例子(有的概念甚至只覆盖一个例子),都有可能是噪音数据.Michalski[8]发现:在进行学习之后得到的规则集中,去掉一些覆盖例子很少的复杂规则之后,识别的精度并没有减少.而值得注意的是,在那些被去掉的复杂规则中,包含有较多的特征(属性),其中某些特征只在这些规则中出现过.这种现象说明,噪音数据的存在,有可能对选择特征子集造成较大困难,不易找到较优的特征子集.

本文以扩张矩阵[9,10]为工具,从统计学的角度提出放宽对特征子集选择的一致性问题的限制,给出含有一定错误率条件下的一致性特征子集定义,并将信息熵和拉普拉斯错误估计函数应用到特征子集选择中,构造了一种新的特征子集选择启发式算法EFS(entropy based feature subset selection).我们将该算法应用于手写数字识别、手写汉字识别等实际问题中,并与GFS(greedy feature subset selection)算法[1]进行了比较.实验结果表明,EFS选择的特征子集比较优化,具有代表性,有效地解决了噪音问题带来的影响.

---

* 本文研究得到国家自然科学基金、国际合作项目彩色匹配基金和哈尔滨工业大学科技基金资助.作者简介:钱国良,1971年生,博士生,主要研究领域为机器学习,彩色匹配,模式识别.舒文豪,1932年生,教授,博导,主要研究领域为模式识别,汉字识别,智能控制.陈彬,1968年生,博士,讲师,主要研究领域为机器学习.权光日,1962年生,副教授,主要研究领域为神经网络与机器学习.

本文通讯联系人:钱国良,哈尔滨 150001,哈尔滨工业大学 319 信箱

本文 1997-09-10 收到原稿,1997-12-18 收到修改稿

1 特征子集的相关概念

下面首先介绍特征子集的基本概念,然后提出从统计学角度,为解决噪音问题而给出含有错误率的一致特征子集的定义.

1.1 一致特征子集[1]

设 PE 和 NE 分别是正例集和反例集,$PE=\{e_1^+,e_2^+,\ldots,e_k^+\}$,$NE=\{e_1^-,e_2^-,\ldots,e_l^-\}$. 其中,例子 $e=(v_1,\ldots,v_n)$ 是 n 维有穷离散向量空间 $D_1\times D_2\times\ldots\times D_n$ 的元素,D_j 是第 j 个特征 X_j 的值域. $X=\{X_1,\ldots,X_n\}$ 是候选特征集合.

定义 1.1.1. 如果 PE 和 NE 不含有公共元素,即 $PE\cap NE=\varnothing$,则称 PE 和 NE 在特征集 X 的表示下是一致的;同时,X 称为关于 PE 和 NE 的一致特征集合. 否则,称 PE 和 NE 不一致,其公共元素称为噪音. X 中具有最小基数的一致特征子集合称为最优特征子集.

定义 1.1.2. 设 $e^+=(v_1^+,\ldots,v_n^+)$ 和 $e^-=(v_1^-,\ldots,v_n^-)$ 分别是正例和反例,做一个扩张矩阵如下:

$$EM(e^+,e^-)=[r_1,\ldots,r_n],r_j=\begin{cases}v_j^-, & \text{if } v_j^+\neq v_j^-;\\ *, & \text{if } v_j^+=v_j^-;\end{cases}$$

其中 $*$ 称为死元素,$EM(e^+,e^-)$ 称为正例 e^+ 在反例 e^- 背景下的扩张矩阵.[7]

正例 e^+ 在反例集 NE 背景下的扩张矩阵:

$$EM(e^+,NE)=\begin{bmatrix}r_{11} & \ldots & r_{1n}\\ \ldots & \ldots & \ldots\\ \ldots & \ldots & \ldots\\ r_{m1} & \ldots & r_{mn}\end{bmatrix},\ r_{ij}=\begin{cases}v_j^-, & \text{if } v_j^+\neq v_{ij}^-;\\ *, & \text{if } v_j^+=v_{ij}^-.\end{cases}$$

定义 1.1.3. 做每一正例 $e_i^+\in PE$ 在反例集 NE 背景下的扩张矩阵 $EM(e_i^+,NE)$,并将其连接在一起,称为 PE 在 NE 背景下的连接扩张矩阵 $EM(PE,NE)$.

$$EM(PE,NE)=\begin{bmatrix}r_{11}^{(1)} & \ldots & r_{1n}^{(1)}\\ \ldots & \ldots & \ldots\\ \ldots & \ldots & \ldots\\ r_{m1}^{(1)} & \ldots & r_{mn}^{(1)}\end{bmatrix}\ldots\ldots\begin{bmatrix}r_{11}^{(k)} & \ldots & r_{1n}^{(k)}\\ \ldots & \ldots & \ldots\\ \ldots & \ldots & \ldots\\ r_{m1}^{(k)} & \ldots & r_{mn}^{(k)}\end{bmatrix}.$$

定理 1.1.1. 设 $EM(PE,NE)$ 是 PE 在 NE 背景下的连接扩张矩阵,则最优特征子集问题等价于在 $EM(PE,NE)$ 中找一条由非死元素组成的路,该路涉及最少数目的列(特征).

定理 1.1.2. 最优特征子集选择问题 OFSS 是 NP 难题.

1.2 含有错误率的一致特征子集

由于在实际应用领域的训练例子中存在噪音数据,对特征子集的选择产生较大影响[7],如果按照一致特征子集的概念用启发式算法寻找特征子集,将很难找到较好的特征子集. 传统的处理噪音方法是采用一些预处理技术对训练例子进行过滤[8],然而这种方法并不十分有效. 因此,我们从统计学的角度,对一致特征子集的概念进行扩充,放松限制,提出含有错误率的一致特征子集,以试图解决噪音问题.

定义 1.2.1. 如果 PE 和 NE 以很小的错误率含有公共元素,即在错误率 ε(ε 很小)下,$PE\cap NE\neq\varnothing$,则称 PE 和 NE 在特征集 X 的表示下是基于错误率 ε 一致的. 这里,ε 定义为使 $PE\cap NE\neq\varnothing$ 的例子个数占整个例子集总数的比例. 同时,X 称为关于 PE 和 NE 的含有错误率的一致特征集合. X 中具有最小基数的含有错误率的一致特征子集合称为含有错误率的最优特征子集.

显然,在定义 1.2.1 中,当错误率 $\varepsilon=0$ 时,含有错误率的一致特征子集就是一致特征子集的定义了. 因此,一致特征子集可以看作是含有错误率的一致特征子集的特殊形式. 显然,由定理 1.1.1,我们可以直接推出相应的定理 1.2.1.

定理 1.2.1. 设 $EM(PE,NE)$ 是 PE 在 NE 背景下的连接扩张矩阵,则含有错误率的最优特征子集问题等价于在 $EM(PE,NE)$ 中找一条由非死元素和以很小的错误率包含有死元素组成的路,该路涉及最小数目的列(特征).

虽然噪音数据是一种小概率事件,但往往给特征选择带来困难.[11]因此,在含有错误率的一致特征子集中,我们的目的就是使获得的特征子集尽可能好地区分开正、反例子集,而不一定是完全彻底地区分开正、反例,从而使得到的特征子集较小,所选特征具有代表性. 因而由此产生的规则质量较高,克服了噪音数据的影响. 根据含有错误率的一致特征子集的定义,不能保证得到的特征子集能够完全区分开训练例子集,却有可能获得较小的、有代表性的特征子集,

提高归纳学习的质址,从而对新的例子有较好的分类效果,因为它基本上是反映训练例子集中大多数例子的特征之间的本质关系.

实际应用领域中的噪音问题,主要是由数据采集或特征提取过程中不可避免的一些错误造成的.噪音数据的判断是一个较为复杂的问题.[11]这里,我们只简单讨论两种主要情况:① 在正反例集中存在很少出现的异常例子;② 正例集和反例集不一致,存在相同的例子.下面给出两个理想情况下的有噪音的连接扩张矩阵例子,比较一下在连接扩张矩阵中根据一致特征子集定义和含有错误率的一致特征子集定义,是如何寻找较优的特征子集的.为简单起见,我们对连接扩张矩阵进行了压缩:将正例集中的每个正例在反例集背景下的扩张矩阵压缩为最有代表性的一行,然后连接到一起,构成正例集在反例集背景下的连接扩张矩阵.

	X_1	X_2	X_3	X_4
1	*	3	0	1
2	3	*	*	1
…	…	…	…	…
$n-1$	*	3	*	1
n	0	1	*	*

例1　$EM1(PE,NE)$

	X_1	X_2	X_3	X_4
1	*	1	0	*
2	*	1	0	*
…	…	…	…	…
$n-1$	*	1	0	*
n	*	*	*	*

例2　$EM2(PE,NE)$

其中,例 1 是在正例集中出现一个异常例子的情况,正例集的所有例子在反例集背景下生成的连接扩张矩阵 $EM1(PE,NE)$,只有最后一行的属性 X_4 与其他行的 X_4 取值不同;例 2 是在正例集中存在与反例 e 相同的例子,正例集的所有例子在反例集背景下生成的连接扩张矩阵 $EM2(PE,NE)$,最后一行全部由死元素组成.

如果根据一致特征子集的定义来选取特征,在例 1 中,$EM1(PE,NE)$有两条较优的路:$l_{14}\to l_{24}\to\cdots\to l_{n-14}\to l_{n1}$ 和 $l_{14}\to l_{24}\to\cdots\to l_{n-14}\to l_{n2}$,对应的特征子集为$\{X_4,X_1\}$或$\{X_4,X_2\}$;例 2 中 $EM2(PE,NE)$没有路,无法得到特征子集,只能通过删除正反例集中的相同例子,重新形成连接扩张矩阵,这样会损失一些有用信息.

如果根据含有错误率的一致特征子集定义,那么在例 1 中,$EM1(PE,NE)$将有一条比较优的路:$l_{14}\to l_{24}\to\cdots\to l_{n4}$,对应的特征子集是$\{X_4\}$;例 2 中 $EM2(PE,NE)$将有两条比较优的路:$l_{12}\to l_{22}\to\cdots\to l_{n2}$和 $l_{13}\to l_{23}\to\cdots\to l_{n3}$,对应的特征子集是$\{X_2\}$或$\{X_3\}$.显然,与一致特征子集相比,利用含有错误率的一致特征子集得到的特征子集较为简单,虽然没有将少址的正例(主要是噪音数据)与反例集区分开,但却能够生成较优的特征子集,可以有效克服噪音问题.相应的启发式算法将在第 2 节给出.

2　基于信息熵的特征子集启发式算法

2.1　评价函数

根据定理 1.2.1,给定正反例集,寻找较优的特征子集,就是正例集在反例集背景下建立的连接扩张矩阵 $EM(PE,NE)$中寻找一条较优的路(可能包含死元素)的过程.选择的路所对应的特征子集,应该能够尽可能地区分开正、反例集.其主要思路是:首先根据一定的评价标准,选择一列(特征),然后在 $EM(PE,NE)$中删去该列的非死元素所对应的行,再在余下部分的矩阵中选择新的列(特征),…,递归进行,直到矩阵满足含有错误率的一致特征子集的条件.

本文将信息熵和拉普拉斯错误估计作为特征子集选择的评价标准.

(1) 信息熵:给定例子集 E,包括正例集 PE 和反例集 NE,分别含有 P 个正例和 N 个反例,选择特征 X_i 后,X_i 的值域 D_i 具有 V 个值$\{V_1,V_2,\ldots,V_v\}$,它将例子集 E 分为 V 个子集$\{E_1,E_2,\ldots,E_v\}$,假设 E_i 中含有 P_i 个正例和 N_i 个反例.我们的目的是选择一个好的特征,使它能够最大限度地区分开正反例子集.设子集 E_i 所需的期望信息是 $I(P_i,N_i)$,以特征 X_i 来区分例子集 E 所需的期望信息是

$$Entropy(E,X_i)=\sum_{i=1}^{v}\frac{P_i+N_i}{P+N}I(P_i,N_i),$$

其中

$$I(P_i,N_i)=-\frac{P_i}{P_i+N_i}\log_2\frac{P_i}{P_i+N_i}-\frac{N_i}{P_i+N_i}\log_2\frac{N_i}{P_i+N_i}.$$

本文选择使 $Entropy(E,X_i)$最小的特征 X_i 作为候选特征.因为 $Entropy(E,X_i)$的值越小,越能最大限度地区分正反例集.

在子集$\{E_1,E_2,\ldots,E_c\}$中,我们将能够完全确定类别的集合称为确定子集,否则称为非确定子集;对于非确定子集序列$\{E_m,\ldots,E_r\}(1\leqslant m\leqslant r\leqslant v)$,我们从整体上来考虑信息熵的作用,选择新的特征$X_j$,计算所有非确定子集的信息熵之和$\sum_{i=m}^{r}Entropy(E_i,X_j)$,选择使信息熵之和最小的特征$X_j$作为新的特征子集成员,然后根据特征$X_j$的取值将非确定子集序列$\{E_m,\ldots,E_r\}$继续向下划分,形成更多新的子集序列$\{E_{m1},\ldots,E_{mu},\ldots,E_{r1},\ldots,E_{rv}\}$,再对其中的非确定子集序列递归地按照上述原则继续进行,直到所有的子集都是确定的.

然而,由于在训练例子中存在噪音数据,要使所有的子集都是确定的,可能要选择较多的只对噪音数据有用的特征来划分,从而对特征子集的优化产生较大影响.因此,在信息熵标准的基础上,本文采用拉普拉斯错误估计来排除噪音带来的影响.

(2) 拉普拉斯错误估计:设当前非确定子集E_i的拉普拉斯错误估计为$Laplace(E_i,\varnothing)=\frac{S-S'+K-1}{S+K}$,其中$S$是$E_i$的例子总数$(S=p+n)$,$S'$是属于$E_i$中某个最大类别的例子数$(p$或$n)$,$K$代表整个训练例子集的所有类别数.当选择一个特征$X_i$后,所形成的拉普拉斯错误估计是$Laplace(E_i,X_i)=\sum_{i=1}^{v}\frac{S'_i}{S}(\frac{S_i-S'_i+K-1}{S_i+K})$,其中$S_i$是$E_i$的第$i$个子集的例子数$(S_i=p_i+n_i)$,$S'_i$是$E_i$的第$i$个子集中某个最大类别的例子数$(p_i$或$n_i)$.通俗地讲,拉普拉斯错误估计反映了在特征选择的过程中所能够区分的例子集的期望错误率.一般来说,在利用信息熵选择特征X_i的过程中,$Laplace(E_i,X_i)$是小于$Laplace(E_i,\varnothing)$的.然而,在实际应用中,当含有噪音数据时,$Laplace(E_i,X_i)$有可能大于$Laplace(E_i,\varnothing)$.因此,为处理噪音带来的问题,我们将拉普拉斯错误估计的减少作为特征子集选择的另一个标准.即:当$Laplace(E_i,X_i)>Laplace(E_i,\varnothing)$时,我们将非确定子集$E_i$标记为确定的,不再用特征$X_i$继续划分$E_i$,并记录$E_i$中的非最大类别的例子个数作为识别错误的例子数.

2.2 启发式算法 EFS

基于信息熵的特征子集选择启发式算法 EFS 描述如下.

正例集 PE,反例集 NE,特征子集 FSS;

步骤 1. $FSS=\varnothing$,$E=PE\cup NE$,$S=|E|$,识别错误的例子数 $Err\_num=0$;E 为当前的子集序列;

步骤 2. 若 E 中的所有特征都被选择过或当前的子集序列$\{E_1,E_2,\ldots,E_c\}$都是确定的,则转步骤 4;

否则,对于当前的子集序列中的非确定子集序列$\{E_m,\ldots,E_r\}$;

FOR(E 中没有选择过的特征 X_j)

计算$\sum_{i=m}^{r}Entropy(E_i,X_j)$;(见 2.1 节的评价函数)

保留使$\sum_{i=m}^{r}Entropy(E_i,X_j)$最小的特征 X_k;

END FOR

步骤 3. FOR($\{E_m,\ldots,E_r\}$的每个子集)

计算 $Laplace(E_i,\varnothing)$和 $Laplace(E_i,X_k)$;(见 2.1 节的评价函数)

若 $Laplace(E_i,\varnothing)<Laplace(E_i,X_k)$,则

将 E_i 标记为确定的,不再进行子集划分,并将 E_i 中非最大类别的例子(错误例子)数累加到 Err_num;

否则

根据特征 X_k 的取值,将 E_i 划分为新的子集序列;

END FOR

若所有新划分的子集序列都为空集,则转步骤 4;

否则,标记 X_k 为选择过,$FSS=FSS\cup\{X_k\}$,将新划分的子集序列作为当前的子集序列,转步骤 2;

步骤 4. 求出错误率 $\varepsilon=Err\_num/S$,输出含有错误率的一致特征子集(FSS,ε),结束.

EFS 的主要思想是:在特征子集的选择过程中,首先递归地选择使当前的非确定子集序列的信息熵的总和最小的特征,然后,在递归过程中,对每个非确定子集利用拉普拉斯错误估计是否减少作为是否继续划分该子集的依据;如果所有新划分的子集序列不全为空集,则将所选的特征作为新的特征子集成员,然后继续递归执行,最后输出含有错误率的一致特征子集(FSS,ε).当 $\varepsilon=0$ 时,EFS 就成为一种精确学习算法.

2.3 一个有噪音的例子

下面,我们通过构造一个有噪音的例子来看看 EFS 是如何工作的.为简单起见,我们构造一个只有两类的训练例

子集：给定正例集 PE，反例集 NE，各含有例子数 100 个；正例集 PE 在反例集 NE 背景下的连接扩张矩阵 $EM(PE, NE)$，如图 1 所示.

	X_1	X_2
1	0	0
2	1	0
…	…	…
99	1	0
100	0	0

正例集 PE

	X_1	X_2
1	1	1
2	0	1
…	…	…
99	1	1
100	2	0

反例集 NE

	X_1	X_2
1	1	1
2	*	1
…	…	…
9999	1	1
10000	2	*

$EM(PE, NE)$

图1　一个有噪音的例子

在 PE 中，所有例子的属性 X_2 的取值都是 0，其中有 70 个例子的属性 X_1 的取值为 1，其余取值为 0；在 NE 中，除最后一个例子为(2,0)外，其余例子的属性 X_2 的取值都是 1，X_1 的取值都是 0 或 1；初步判断反例集 NE 的最后一个例子可能为噪音.

EFS 的工作情况如下. 从步骤 2 开始，由于 E 是非确定的子集，首先找到使信息熵之和最小的特征 X_2，$Entropy(E, X_2) = -\frac{99}{99}\log_2\frac{99}{99} - \frac{1}{101}\log_2\frac{1}{101} - \frac{100}{101}\log_2\frac{100}{101} = 0.024$，$Laplace(E, \varnothing) = \frac{100+2-1}{200+2} = 0.5$，$Laplace(E, X_2) = \frac{99}{200} * \frac{0+2-1}{99+2} + \frac{101}{200} * \frac{1+2-1}{101+2} = 0.015$，显然，$X_2$ 满足我们的启发式评价标准，被选入特征子集；然后根据特征 X_2 的取值{0,1}，将 E 划分成新的子集序列{E_0, E_1}，其中 E_1 是确定的子集{99 个反例}，E_0 是非确定的子集{100 个正例和 1 个反例}. 重新返回步骤 2，对非确定子集 E_0，第 2 次选择特征 X_1，其取值为{0,1,2}，$Entropy(E_0, X_1) = -\frac{1}{1}\log_2\frac{1}{1} - \frac{30}{30}\log_2\frac{30}{30} - \frac{70}{70}\log_2\frac{70}{70} = 0.0$，$Laplace(E_0, \varnothing) = \frac{1+2-1}{101+2} = 0.0194$；$Laplace(E_0, X_1) = \frac{1}{101} * \frac{0+2-1}{1+2} + \frac{30}{101} * \frac{0+2-1}{30+2} + \frac{70}{101} * \frac{0+2-1}{70+2} = 0.0222$，因为 $Laplace(E_0, X_1) > Laplace(E_0, \varnothing)$，所以将 E_0 标记为确定的，不再对 E_0 继续划分，并将 E_0 中非最大类别的例子数 1 累加到 Err_num，由于新的子集序列都是空集，因此 X_1 不被选入特征子集，递归结束. 最后，计算错误率 $\varepsilon = \frac{1}{200} = 0.005$，输出含有错误率的一致特征子集({$X_2$}, 0.0099).

在上述的算法执行过程中，我们看到，EFS 算法只选择了一个特征{X_2}，没有将连接扩张矩阵中的所有行都删除，存在不一致性；但是它却用很简单的特征构造了含有错误率的一致特征子集，以很小的错误率就将正反例集基本分开，有效地处理了噪音. 而如果根据传统的精确学习启发式算法 GFS[1]，得到的特征子集是{X_1, X_2}；这说明，EFS 算法选择的特征简单，具有代表性，能够较为有效地克服噪音问题带来的影响.

3　应用结果比较和结论

我们将 EFS 应用于两个实际问题，并与 GFS 启发式算法进行了比较.

(1) 自由手写数字识别：这里提供由不同人书写的 0～9 的 10 个数字作为训练例子，共 6 000 个例子，每例提取了包括孔数、孔中心的位置、边缘轮廓、四叉点数、几何和笔划密度在内等共 29 个特征.

(2) 手写汉字识别：提供了 1 000 个常用字，每个字有 30 个手写样本，每例提取了周边特征、结构划分特征、特征点、网格特征和笔划密度分布等 18 个特征.

在上述两个实际问题中，我们随机选取 70%个例子作为训练例子集用于特征子集选择，剩下的 30%作为测试例子集. 共进行 5 次实验，取平均值. 由于在实际应用中存在噪音现象，采用 GFS 算法时首先要排除训练例子集各类中含有相同例子的噪音情况，然后再进行学习. 而采用 EFS 算法则不必这样做，直接学习即可.

在手写数字识别中，利用 EFS 和 GFS 算法得到的特征子集个数平均分别是 16 和 19 个. 其中 EFS 的特征子集错误率 $\varepsilon = 0.023$；在将得到的相应特征子集应用 AQ15 对训练例子学习之后，进行测试的正确率平均分别为 98.8%和 98.2%. 在手写汉字识别中，利用 EFS 和 GFS 算法得到的特征子集个数平均分别是 14 和 17 个. 其中 EFS 的特征子集错误率 $\varepsilon = 0.038$；在将得到的相应特征子集应用 BP 神经网络对训练例子学习之后，进行测试的正确率平均分别为 85.6%和 83.7%.

从实验结果可以看出，EFS 生成的特征子集较为优化，测试的精度也高于 GFS，充分体现了处理噪音的能力. 我们还发现，在几次实验中，利用 EFS 所获得的特征子集都相同，而 GFS 在手写数字识别中有两个特征不相同，手写汉

字识别中有1个特征不相同.这说明EFS具有较好的稳定性.另外,由于EFS产生的特征子集较小,识别速度也有较大提高(因为在识别过程中用于特征抽取的时间较长).但是,由于EFS采用递归方法,在寻找特征子集的空间和时间消耗上要大于GFS.实验结果表明,EFS选择的特征子集更具有代表性,较为有效地解决了实际应用中的噪音影响问题.

参考文献

1 陈彬,洪家荣,王亚东.最优特征子集选择问题.计算机学报,1997,20(2):133～138
(Chen Bin, Hong Jia-rong, Wang Ya-dong. The optimal feature subset selection problem. Chinese Journal of Computers, 1997,20(2):133～138)

2 Almualim H, Dietterich T G. Learning with many irrelevant features. In: Mitchell T M ed. Proceedings of the 9th National Conference on Artificial Intelligence. Anaheim CA: AAAI Press, 1991. 547～552

3 John G H, Kohavi R, Pfloger K. Irrelevant features and the subset selection problem. In: Dietterich T ed. Proceedings on Machine Learning'94. Morgan Kauffmann Publishers, 1994. 121～129

4 Caruana R, Freitag D. Greedy attribute selection. In: Dietterich T ed. Proceedings on Machine Learning'94. Morgan Kauffmann Publishers, 1994. 28～36

5 Quinlan J R. Induction of decision trees. Machine Learning, 1986,1(1):81～106

6 Niblett T. Constructing decision trees in noisy domains. In: Mitchell T M ed. Proceedings of the 2nd European Working Session on Learning. UK: Sigma Press, 1987. 67～78

7 Michalski R S, Stepp R E. Automated construction of classifications: conceptual clustering versus numerical taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1983,5(4):396～410

8 Michalski R S *et al*. The multi-purpose incremental learning system AQ15 and its testing application to three medical domains. In: Proceedings of the 5th National Conference on Artificial Intelligence. Los Altos, California: Morgan Kaufmann Publishers, 1986. 1041～1045

9 洪家荣.示例学习的扩张矩阵理论.计算机学报,1991,14(6):401～410
Hong Jia-rong. The extension matrix theory of learning from examples. Chinese Journal of Computers, 1991,14(6):401～410

10 Wu X D. HCV: a heuristic covering algorithm for extension matrix approach. Technical Report. Edinburgh DAI 578, University of Edinburgh, 1992

11 Norton S W, Hirsh H. Classifier learning from noisy data as probabilistic evidence combination. In: Mitchell T M. Proceedings of the 10th National Conference on Artificial Intelligence. Anabeim, CA: AAAI Press/MIT Press, 1992. 141～146

Research on a Heuristic Algorithm of Feature Subset Selection Based on Entropy

QIAN Guo-liang SHU Wen-hao CHEN Bin QUAN Guang-ri

(Department of Computer Science and Engineering Harbin Institute of Technology Harbin 150001)

Abstract FSS(feature subset selection) is an important problem in the fields of machine learning and pattern recognition. Minimum FSS problem has been proved NP hard. However, existing heuristic algorithms are based on the consistency of positive and negative examples set, and a more optimal feature subset is hard to be produced under the noisy data in application to real-world domains. In this paper, from the degree of statistics, the effects of noisy data on FSS is analyzed firstly, and a concept of consistent feature subset which contains error rate is given. Then a heuristic algorithm —EFS (entropy based feature subset selection) based on information-theoretic entropy measure and Laplace error rate is presented. It is also applied to two real-world domains and is compared with GFS (greedy feature subset selection). The experimental results show that EFS can produce more representative feature subset, and can solve the noisy problem in the practical application effectively.

Key words Feature subset selection, machine learning, extension matrix, entropy, noise.