文章编号:1001 - 9081(2011)03 - 0724 - 05
doi:10.3724/SP.J.1087.2011.00724

基于 Q-relief 的图像特征选择算法

范文兵[1],王全全[1],雷天友[2],朱　辉[3]
(1.郑州大学 信息工程学院,郑州 450001; 2.郑州大学 科研处,郑州 450001; 3.郑州大学 电气工程学院,郑州 450001)
(wqqzhu1@126.com)

摘　要:针对特征选择算法——relief 在训练个别属性权值时的盲目性缺点,提出了一种基于自适应划分实例集的新算法——Q-relief,该算法改正了原算法属性选择时的盲目性缺点,选择出表达图像信息最优的特征子集来进行模式识别。将该算法应用于列车运行故障动态图像监测系统(TFDS)的故障识别,经实验验证,与其他算法相比,Q-relief算法明显提高了故障图像识别的准确率。

关键词:特征选择;relief 算法;纹理特征;模式识别
中图分类号: TP391.41　　**文献标志码**:A

Image feature selection algorithm based on Q-relief

FAN Wen-bing[1], WANG Quan-quan[1], LEI Tian-you[2], ZHU Hui[3]
(1. *School of Information Engineering, Zhengzhou University, Zhengzhou Henan* 450001, *China*;
2. *Department of Scientific Research, Zhenzhou University, Zhengzhou Henan* 450001, *China*;
3. *School of Electric Engineering, Zhengzhou University, Zhengzhou Henan* 450001, *China*)

Abstract: Image feature selection is the significant part in pattern recognition, image understanding and so on. The relief algorithm has a blind deficiency in training feature weight. Q-relief was a new algorithm which was based on dividing instance set in self-adapting. Q-relief was proposed to solve the blind selection problem in the original relief algorithm. The presented algorithm was applied in Trouble of Moving Freight Car Detection System (TFDS). The classification results show that the Q-relief algorithm can improve the accuracy of recognition compared with other algorithms.

Key words: feature selection; relief algorithm; texture feature; pattern recognition

0　引言

特征选择在模式识别、特征降维等方面有着广泛的应用[1-2],也是图像处理等领域关键难题之一。一个好的特征选择算法可以有效地剔除特征空间中的冗余,找到与模式类别相关性最强的一组特征子集,这样为接下来的匹配识别铺平了道路。特别是针对特征维数较高的问题,特征选择更能体现出其优越性,先选出最优特征子集再进行识别,可以有效地降低计算复杂度。进行特征选择的算法有很多,例如主成成分分析(Principal Component Analysis, PCA)和独立成分分析(Independent Component Analysis, ICA)等算法[3]。在现有的特征选择算法中,1992 年由 Kira 等人提出的 relief 算法[4],因其具有简单、直观、计算量小等特点是目前较好的特征选择算法,在很多领域得到广泛的应用[5-6];之后,Kononenko 等人针对 relief 算法不能处理多类别问题的缺点又提出了 reliefF 算法[7-8];随后又由 Sun 提出的I-relief算法[9]通过探索期望最大化的框架对 relief 算法做了进一步的改进。但是之前的研究都未发现 relief 算法在训练个别属性时存在盲目性选择的缺点。

本文对 relief 算法存在的盲目性进行了分析,并提出了一种新的改进算法——Q-relief 算法。该算法根据引起 relief 算法盲目性的这些特征的特点对训练实例进行自适应划分,体现出这些特征与类别信息真实的相关度,从而克服了原算法的缺点,为进一步识别判断提供了有效信息。

1　特征选择算法

1.1　relief 算法

relief 算法的思想是给属性集中每一维属性赋予一个权值,利用权值更新公式进行实例训练。最终使得与聚类相关性较强的属性或者说对聚类贡献较大的属性获得较大的权值,用权值大的属性组成最优特征子集来表达类别信息,relief 算法的具体实现见算法 1。

算法 1　relief 算法。

/* $S = [s_1, s_2, \cdots, s_N]$ 为实例空间,$s_i = [s_{i1}, s_{i2}, \cdots, s_{iM}]$ 为第 i 个实例的特征空间 */

1) 赋属性权值初值 $W_M = zeros(1, M)$;

2) for $j = 1:M$

　/* 针对实例的每维属性进行训练 */

　for $i = 1:N$

　/* 对实例集 S 中的实例进行训练 */

3) 找到实例 s_i 的同类最近邻 s_{hit} 和非同类最近邻 s_{miss};

4) 针对上述实例的第 j 维属性进行如下权值更新:

$$w_j^i = w_j^{i-1} + \frac{1}{N}\left[\frac{|s_{i,j} - s_{miss,j}|}{\max(s_{*j}) - \min(s_{*j})} - \frac{|s_{i,j} - s_{hit,j}|}{\max(s_{*j}) - \min(s_{*j})}\right]$$

收稿日期:2010 - 08 - 16。　**基金项目**:国家自然科学基金资助项目(60574098);河南省教育厅自然科学基金资助项目(2010A510014);郑州市科技攻关项目(0910SGYG25229-6)。

作者简介:范文兵(1969 -),男,河南郑州人,副教授,博士,主要研究方向:图像处理、图像通信; 王全全(1985 -),男,河南开封人,硕士研究生,主要研究方向:图像处理、模式识别; 雷天友(1962 -),男,河南郑州人,副教授,主要研究方向:智能信息处理; 朱辉(1986 -),女,河南周口人,硕士研究生,主要研究方向:图像处理、模式识别。

/* $|s_{i,j}-s_{*j}|$为实例 s_i 与 s_* 第 j 个属性间的欧几里得距离;$\max(s_{*j})$ 与 $\min(s_{*j})$ 分别是实例空间中第 j 个属性的最大和最小值 */

5) end;

对于分类抉择具有较大影响的属性,同类实例间的距离较近,而不同类实例间的距离较远。利用更新公式对属性权值训练后,与类别相关性强的属性将获得较大的权值,反之与分类相关性弱或者无关的属性,最终训练出的权值较小。

从 relief 算法的原理可以发现,它只适用于问题中存在两种类别的情况;而对于问题中存在多个类别需要被识别时,已不再适用。reliefF 算法则可以实现多类别的特征选择,并且一定程度上抑制了噪声信号对分析结果的干扰。

1.2　reliefF 算法

reliefF 算法[10]与 relief 最大的不同之处在于前者的权值更新公式没有使用一个最近邻而是使用 k 个最近邻,并且能够处理多类别的问题,reliefF 算法实现如算法 2 所示。

算法 2　reliefF 算法。

/* $S=[s_1,s_2,\cdots,s_N]$ 为实例空间;$s_i=[s_{i1},s_{i2},\cdots,s_{iM}]$ 为第 i 个实例的特征空间;$CLASS=[class_1,class_2,\cdots,class_p]$ 为实例待识别的类别空间 */

1) 赋属性权值 $W_M=zeros(1,M)$;

2) for $j=1:M$

/* 针对实例的每维属性进行训练 */

for $i=1:N$

/* 对实例集 S 中的每个实例进行训练 */

3) 找到实例 s_i 的 k 个同类最近邻$[s_{\text{hit},1},s_{\text{hit},2},\cdots,s_{\text{hit},k}]$和 k 个非同类最近邻$[s_{\text{miss},1},s_{\text{miss},2}\cdots,s_{\text{miss},k}]$;

4) 针对上述实例的第 j 维属性进行如下权值更新:

$$w_j^i=w_j^{i-1}+\frac{1}{k*N}[dis\_miss(s_i,k)-dis\_hit(s_i,k)] \tag{1}$$

/* $dis_hit(s_i,k)=\sum_{m=1}^{k}\frac{|s_{i,j}-s_{\text{hit},m,j}|}{\max(s_{*j})-\min(s_{*j})}$代表 s_i 与同类实例间的区分度; $dis_miss(s_i,k)=\sum_{c\neq class(s_i)}\left[\frac{P(c)}{1-P(class(s_i))}\cdot\sum_{m=1}^{k}\frac{|s_{i,j}-s_{\text{miss},m,j}|}{\max(s_{*j})-\min(s_{*j})}\right]$代表 s_i 与所有非同类实例间的区分度;其中:$\max(s_{*j})$ 与 $\min(s_{*j})$ 分别是实例空间中第 j 个属性的最大和最小值;$P(c)$ 为第 c 类出现的概率,可以用第 c 类实例数比上实例集中的实例总数求出 */

5) end

求出各维属性相应的权值后,权值由大到小代表该属性与分类相关性从强到弱,这样就可以剔除小于某个阈值的无关和相关性较弱的属性,从而组成新的特征子集进行接下来的模式识别,实现了特征降维的目的。

2　relief 相关算法的不足及改进

2.1　relief 相关算法的缺点

Kononenko 等人[7-8]根据统计理论对 reliefF 进行了分析,得出当训练实例集充分大时,属性 i 的权值 w_i 逼近于:

$$w_i=P(\text{miss}_i)-P(\text{hit}_i)$$

其中:$P(\text{miss}_i)$ 为非同类实例间属性 i 取值不同的概率,$P(\text{hit}_i)$ 为同类实例间属性 i 取值不同的概率。

虽然 reliefF 算法通过权值能够很简单地体现出属性与类别之间的相关性。但是 reliefF 算法在实际应用时仍存在着不足之处。例如针对实际问题所涉及到的模式类别都已确定,但是实例中的某类属性仍含有问题中没有涉及到的模式信息,如图 1 所示,图 1 给出了实例空间中属性 η_i 的分布。其中 $class_A$、$class_B$ 是问题中所涉及的模式,而 $class_C$ 和 $class_D$ 是属性 η_i 体现出的隐含模式,是此问题中不需要被识别的。

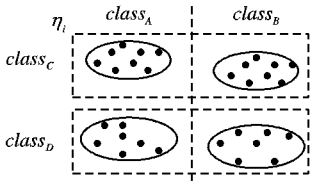


图 1　属性 η_i 的多模式表现实例

经验证发现,该类属性经过 reliefF 算法训练后,无论其与分类是否相关,都会得到一个较大的权值,这显然是不恰当的。为了便于进一步说明此问题,作出如下定义。

定义 1　若实例属性 η_i 包含已确定的类别空间$[class_1,class_2,\cdots,class_M]$中所没有的类别信息——$class_x$,则称 η_i 为伪属性,所包含的类别种类数称为伪属性的伪级,记做 τ。

定义 2　实际问题中已确定的类别称为显类别,而由伪属性所体现出来的类别称为隐类别。

定义 3　实例集中属性 η_i 的空间分布图称为 η_i 的属性图。

伪属性具有以下特点:不管该属性与类别是否具有相关性,经过 reliefF 训练后都获得了较大的权值,体现出其与分类有着强相关性。如果用伪属性所构成的特征子集进行下一步的模式识别,就极可能产生错误的判断,影响识别的效果。下面利用基于欧氏最小距离分类器的模式识别方法证明。

证明　实例空间为 S,实际问题中所确定的模式类别数为 M,伪属性 η_i 的伪级 $\tau=N$,不妨设 $M=N=2$,则 η_i 的属性如图 2 所示。

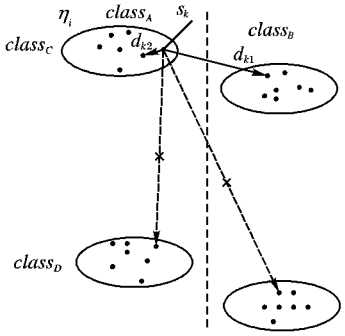


图 2　伪属性 η_i 的属性

通过 reliefF 算法对伪属性 η_i 进行学习,可得 η_i 的训练权值 w_i。

此问题中需识别的模式为显类别 $class_A$ 和 $class_B$,而 $class_C$ 和 $class_D$ 为隐类别。由图 2 可知,实例 s_k 同类最近邻 $s_{\text{hit}}\in class_A\cap class_C$,非同类最近邻 $s_{\text{miss}}\in class_B\cap class_C$,由于 $M=2$,所以$\frac{p(c)}{1-p(class(s_k))}=1$,因此$dis_miss(s_k,1)-dis_hit(s_k,1)=\frac{d_{k1}-d_{k2}}{\max(s_{*i})-\min(s_{*i})}$。$s_k$ 与其非同类和同类最近邻的距离分别为 d_{k1} 和 d_{k2},为图 2 中实箭头长度所示,显然 $d_{k1}-d_{k2}>0$,所以 $dis\_miss(s_k,1)-dis\_hit(s_k,1)>0$,同理对 $\forall s_k\in S$,有 $dis\_miss(s_k,n)-dis\_hit(s_k,n)>0$。

由以上分析可知对 $\forall s_k\in S$,实例 s_k 的同类最近邻都会选择与其属于同一隐类别的同类的最近实例,它的非同类最近邻都会选择与其属于同一隐类别而不同类的最近实例,于是,与同类最近邻的距离都小于其与非同类最近邻的距离。从

而通过权值更新式(1),无论伪属性 η_i 与分类是否相关,η_i 都将获得一个较大的权值 w_i。这表明 reliefF 算法在处理此类问题时存在很大的不足,具有“盲目性”的缺点,这是不符合 reliefF 算法初衷的。

再利用欧氏最小距离分类器[11]进行模式识别时,由实例确定的显类别 $class_A$ 与 $class_B$ 的决策面如图 3 虚线框所示。利用待测实例 s_k(如图 2 所示)的属性 η_i 与决策面的距离远近来判断其属于哪一类。但由图 3 可得,当待测实例 s_k 属于隐类别 $class_C$ 时,其伪属性 η_i 将在隐类别 $class_C$ 的区域内分布,可是无论事实上 s_k 是属于 $class_A$ 还是 $class_B$ 都会被认为属于 $class_A$,因为这时它与 $class_A$ 决策面(上方框区域)的距离最近。同理当 $s_k \in class_D \cap class_A$ 时,由于 s_k 与 $class_B$ 的决策面距离(下面方框区域)最近,所以经过分类器判断后会被确定 $s_k \in class_B$,这显然是不正确的。这样就可以证明当伪属性参与到模式表达中时,其包含的隐类别信息将会影响到对模式的正确判断。

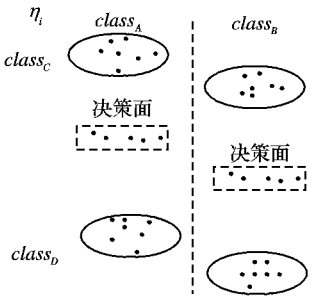


图 3 由 η_i 确定的决策面

出现这个问题的根本原因在于伪属性 η_i 经过 reliefF 算法训练后总能得到较大的权值,从而将其纳入特征空间中,最终导致错误的结论。这说明 reliefF 训练伪属性时,得到的权值不能真实反映其与类别的相关性。之前与 relief 相关的算法都未指出此问题并提出改进方案,所以针对此问题提出有效解决方案是必要的。

2.2 Q-relief 算法

实际应用中伪属性无处不在,例如对基因进行分类识别时,某些基因突变是导致肌无力和剪刀脚等症状的原因,同时这些基因突变还伴随着智力低下等现象。但是当将分辨导致智力低下的基因作为研究的课题时,那么肌无力等病症就属于隐类别,所以包含肌无力等症状信息的基因就是伪属性。这种例子还有很多,可见并不能忽视伪属性的特性,不然将会影响到最终判断的正确性。

为了解决 reliefF 算法在对伪属性权值判断时存在的不合理问题,本文提出了一种改进的 reliefF 算法——Q-relief 算法。

Q-relief 算法思想 对属性权值训练前,先对属性进行查询,判断是否存在伪属性 η_i。判断一个属性是否是伪属性,即判断其是否含有隐类别的信息。这需要从先验知识或专家知识中得到一个阈值向量 $\boldsymbol{T}$,$\boldsymbol{T}$ 的作用是根据该属性将实例划分为不同的组份,如果组份内实例之间的最大距离要远远小于组份间实例之间的最小距离(如图 2 所描绘),经分析这个差距通常是 3.5 倍以上,即可判断该属性是伪属性。如果确定为伪属性,则通过阈值 $\boldsymbol{T}$ 将训练实例集 S 按照伪属性中的隐类别信息 $class_1, class_2, \cdots, class_M$ 划分成不同子集 $S_1, S_2, \cdots, S_M$。当训练到伪属性 η_i 的权值时,实例分别按照子集 $S_1, S_2, \cdots, S_M$ 单独来训练,得到权值 $w_1, \cdots, w_M$,用这些权值分别与其他属性训练出的权值进行比较,来得出最终的最优特征空间;反之,如果不是伪属性则直接训练。具体实现步骤如下。

算法 3 Q-relief 算法。

```
/* S 为实例集,由先验知识得到阈值向量 T */
/* W = [w_1, w_2, ..., w_M] 为各属性的权值 */
/* num 的第 l 行 num(l,:) 存放实例集 S 中属性 η_l 大于阈值 T(l) 的实例编号 */
/* class_rec 得到实例集按隐类别分类结果;S_j^Δ 是由伪属性的隐类别所确定的实例集 */
/* w_j^Δ 是伪属性 η_i 由实例集 S_j^Δ 经过 reliefF 训练得到的权值向量 */
/* w_j 是属性 η_j 由实例集 S 经过 reliefF 训练得到的权值 */
1) for l = 1:M
     num(l,:) = {S(η_l) > T(l)};
     由 num(l,:) 将 S 划分为不同组份 C = {C_l1, C_l2, ..., C_lr};
   end
2) for j = 1:M
     if min_dis(C_j*, C_j#) > 3.5 × max_dis(C_j)
       /* 如果组份间实例之间的最小距离大于组份内实例之间的最大距离的 3.5 倍,则属性 η_j 为伪属性;C_j*, C_j# 代表由属性 η_j 划分的不同组份 */
       S_j^Δ = [s_1^Δ, s_2^Δ, ..., s_k^Δ] = class_rec(S, η_j, C);
       /* 根据 η_j 和 C 将原实例集 S 按伪属性所含有的隐类别重新划分,使得得到的每个 s_k^Δ 只含有一个隐类别的信息 */
3) for i = 1:k
       w_j^Δ(i) = reliefF(η_i, S_j^Δ(i));
       /* 将重新划分的实例集 S_j^Δ(i) 和 η_i 代入到算法 2——reliefF 算法中,进行权值训练 */
     end
4) else
     w_j = reliefF(η_j, S);
     /* 如果 η_j 不是伪属性,实例集则不重新划分,直接代入算法 2 中进行属性权值训练 */
   end
5) end
```

Q-relief 算法在训练伪属性权值时先按隐类别将实例集进行自适应划分再训练。此算法优点在于:将实例按隐类别划分,其实也就是使得隐类别信息能够在训练中表现出来,这样最终得到的伪属性权值 w 是由同类隐类别实例训练得到的,能够真实体现出此属性在该隐类别内对分辨显类别所起到的作用。

最后将训练结果综合比较,如果属性经过 Q-relief 训练得到的权值小于设定的阈值(对属性取舍的标准),就将其舍去;反之,应将其纳入表征模式信息的特征子集中。将这些被选出的属性组成相应的模板进行最终的模式识别,这样就从根本上克服了 reliefF 算法在处理具有伪属性的特征选择时的盲目性缺点。

3 Q-relief 算法的应用

近年来,列车运行故障动态监测系统(trouble of moving freight car detection system, TFDS)已在各大交通枢纽的列检所投入使用。TFDS 是一款通过照相机拍摄高速行驶中列车的部件,实时监控列车运行状态,以保障列车正常运行的监测系统。但在列车部件图像(如图 4 所示)故障判断上,仍是依

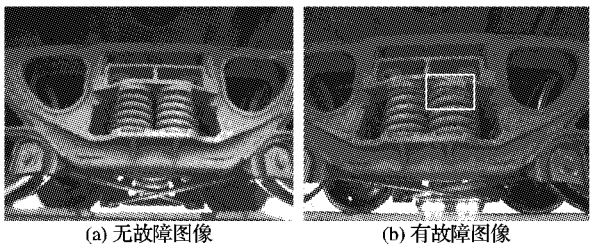
靠人眼去判断,这不免会受限于操作人的精神状态、注意力等主观因素。本文将利用 Q-relief 实现对故障图像的自动识别,并验证 Q-relief 的实际应用效果。

首先采用基于灰度共生矩阵的纹理特征提取方法提取图像的特征。根据对图像中搜索时选取的方向不同,可以有 0°、45°、90° 和 135° 四个方向上的灰度共生矩阵[12]。由共生矩阵再提取下列纹理特征[13]:对比度 $Con = \sum_{i,j}(i-j)^2 * p(i,j)$;熵 $Ent = \sum p(i,j) * \lg[p(i,j)]$;能量 $Ene = \sum_{i,j}(p(i,j))^2$;均等性 $Hom = \sum_{i,j} p(i,j)/[1+(i-j)^2]$;不相似性 $Dis = \sum_{i,j} |i-j| * p(i,j)$;相关性 $Cor = \sum_{i,j} i*j*p(i,j) - u_1u_2/\sigma_1^2\sigma_2^2$,其中:$u_1 = \sum i \sum p(i,j)$,$u_2 = \sum j \sum p(i,j)$,$\sigma_1^2 = \sum (i-u_1)^2 \sum p(i,j)$,$\sigma_2^2 = \sum (j-u_2)^2 \sum p(i,j)$。

由每个灰度共生矩阵都能得到以上 6 个属性,故由一幅图像共可以提取 24 个属性,用这 24 个属性来表征一幅图像的信息。为了方便说明,对属性标识做如下规定:如由 0° 方向的灰度共生矩阵得到的对比度——Contrast0;由 135° 方向的灰度共生矩阵得到的能量——Energy135 等。

4　实验结果

针对这 24 个属性分别用算法 1——relief、算法 2——reliefF 和本文提出的算法 3——Q-relief 算法训练属性权值选择与类别相关性较强的最优特征集来表征实例。在实现算法 3时,本例中由于需分辨的是图片中存在故障与否,涉及两个类别,阈值向量 $\boldsymbol{T}$ 可选择为各属性在实例集内的均值,代入到 Q-relief 算法中进行权值训练。



(a) 无故障图像　　(b) 有故障图像

图 4　枕簧实例图像

本文用三种算法分别对图片个数为 120、240、350 的三组实例集进行训练,表 1 给出了由 relief、reliefF 和 Q-relief 算法训练得到的权值对比,由表 1 可以看出实例经过 relief 和 reliefF 算法训练后,属性 Contrast0、Contrast45、Contrast90 和 Contrast135 都获得较大的权值,但是由 Q-relief 算法训练得到的权值较小。这表明 Q-relief 算法在 Contrast 这个特征选择上与 relief 和 reliefF 两者发生了分歧。由算法 3 可知,只有当碰到伪属性时,Q-relief 才与 reliefF 不同,所以属性 Contrast 被 Q-relief 判断为伪属性。为了验证三者到底谁正确,用三种算法所选择出的特征子集分别进行模式识别来比较。现抽取每种算法中训练出的最大的 6 个属性作为表达图像的特征子集(共三组),分别对图片个数为 45、80、150、200、300 的五组待测图像进行识别,该待测实例与训练实例都不相关,识别结果如图 5 所示。

表 1　relief、reliefF 和 Q-relief 训练得到的权值表

属性	relief 算法训练的图片数			reliefF 算法训练的图片数			Q-relief 算法训练的图片数		
	120	240	350	120	240	350	120	240	350
Contrast0	0.978	0.965	0.982	0.896	0.868	0.882	0.125	0.130	0.129
Correlation0	0.065	0.080	0.071	0.189	0.192	0.184	0.189	0.192	0.184
Energy0	0.004	0.004	0.004	0.009	0.009	0.008	0.009	0.009	0.008
Homogeneity0	0.198	0.194	0.202	0.252	0.260	0.262	0.252	0.260	0.262
Dissimilar0	0.604	0.597	0.591	0.435	0.457	0.441	0.435	0.457	0.441
Entropy0	0.260	0.251	0.258	0.057	0.061	0.052	0.057	0.061	0.052
Contrast45	0.904	0.915	0.912	0.825	0.831	0.826	0.033	0.039	0.036
Correlation45	0.010	0.011	0.009	0.094	0.090	0.097	0.094	0.090	0.097
Energy45	0.004	0.004	0.004	0.009	0.008	0.009	0.009	0.008	0.009
Homogeneity45	0.182	0.171	0.193	0.206	0.210	0.202	0.206	0.210	0.202
Dissimilar45	0.451	0.429	0.468	0.200	0.204	0.198	0.200	0.204	0.198
Entropy45	0.371	0.370	0.359	0.140	0.138	0.114	0.140	0.138	0.144
Contrast90	0.942	0.921	0.948	0.873	0.861	0.874	0.062	0.067	0.068
Correlation90	0.032	0.040	0.041	0.137	0.133	0.140	0.137	0.133	0.140
Energy90	0.004	0.004	0.004	0.008	0.009	0.008	0.008	0.009	0.008
Homogeneity90	0.112	0.108	0.127	0.198	0.200	0.197	0.198	0.200	0.197
Dissimilar90	0.520	0.496	0.509	0.235	0.230	0.241	0.235	0.230	0.241
Entropy90	0.265	0.270	0.262	0.104	0.102	0.108	0.104	0.102	0.108
Contrast135	0.928	0.920	0.919	0.842	0.836	0.848	0.039	0.041	0.034
Correlation135	0.012	0.019	0.016	0.112	0.114	0.118	0.112	0.114	0.118
Energy135	0.004	0.004	0.004	0.009	0.009	0.009	0.009	0.009	0.009
Homogeneity135	0.092	0.084	0.089	0.199	0.199	0.191	0.199	0.199	0.191
Dissimilar135	0.396	0.402	0.415	0.167	0.159	0.168	0.167	0.159	0.168
Entropy135	0.334	0.318	0.320	0.139	0.134	0.136	0.139	0.134	0.136

图 5 显示由三种算法选择出的特征集进行模式识别的结果,可以看出用 Q-relief 算法选择出的特征进行分类识别误判的数目要远远小于 relief 和 reliefF 算法。之所以有这样的差别是因为在这 24 个属性中确实存在伪属性“对比度”。

经过对训练实例进一步分析发现:存在伪属性对比度的原因是图像是在列车行驶时由高速照像机摄取的,拍摄环境

较复杂,会产生曝光过度的现象,这是无法避免的。于是曝光过度的图像对比度就较曝光适中的图片大得多,如图6(a)和6(b)所示。图7是部分图像的对比度值,从图中可以看出,无论图像是否存在故障,其Contrast0与Contrast135都呈两个区域分布,其中Contrast135则以值150为分界线区分得更明显。这符合伪属性的特点,由此可以断定"对比度"包含了问题中需判断的显类别(是否是故障图)未体现出的隐类别——曝光的差异,所以对比度是伪属性。并且在实验中由relief和reliefF算法识别错误的图片均是如图6(a)所示的曝光过度的图片。这说明隐类别"曝光差异"影响到了对显类别正确判断的决策面,从而由relief和reliefF训练出的Contrast的权值并不是该属性与模式相关性强弱的真实体现,这也是本文称此类属性为伪属性的另一层含义。

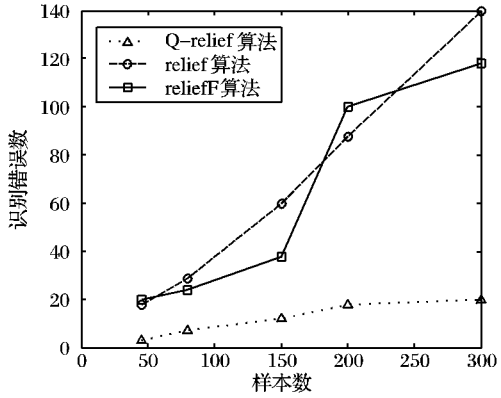


图5 三种算法识别错误数对比

(a)曝光过度 (b)曝光适中

图6 曝光不同的图像

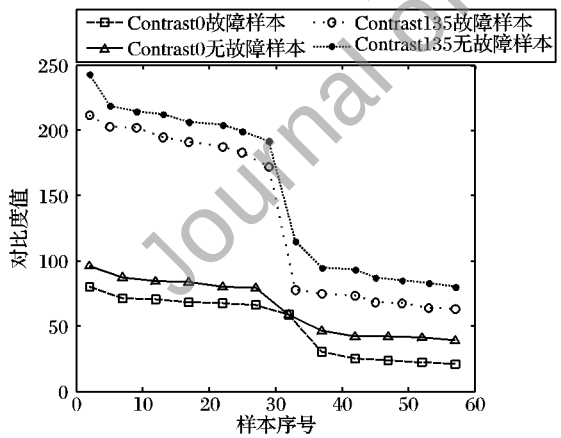


图7 部分实例对比度值

Q-reliefF算法弥补了relief和reliefF算法的不足,伪属性Contrast经过Q-relief训练得到较小的权值,证明其对分类识别的贡献不大,属于无关属性,在属性选择时将其剔除。所以在模式识别时,减少了对显类别决策面的干扰,从而具有较高的识别率。经验证,正确率可达到93.3%。

5 结语

本文提出了一种改进的特征选择算法——Q-relief算法。此算法针对原relief算法在训练伪属性权值时存在的盲目性,提出了划分实例再训练的方法,使得伪属性能够获得真实的权值,并对原因作了详细分析。并成功地将该算法运用到列车运行故障动态监测系统中的故障图像识别中。经实验验证,用Q-relief选择出的特征集进行模式识别时,正确识别率远高于relief和reliefF算法,可达到93.3%,说明Q-relief算法有较好的实用性。但该算法仍存在不足之处:当伪属性维数很大时,会存在多个匹配模板,计算复杂度会大大增加,识别速率会有所下降,这有待进一步研究解决。

参考文献:

[1] SCHERWANI K, ALI N, LOTIA N. A computational economy based job scheduling system for clusters [J]. Software Practice and Experience, 2004, 34(6): 581-598.

[2] KANG O-H, KANG S S. A Web-based toolkit for scheduling simulation using GridSim [C]// GCC'06: Proceedings of the Fifth International Conference on Grid and Cooperative Computing. Changsha, China: IEEE, 2006: 256-271.

[3] SWINIARSKI R W, SKOWRON A. Rough set methods in feature selection and recognition [J]. Pattern Recognition Letters, 2003, 24(6): 833-849.

[4] KIRA K, RENDELL L A. A practical approach to feature selection [C]// Proceedings of the 9th International Workshop on Machine Learning. Washington, DC: [s.n.], 1992: 249-256.

[5] ZHANG JIANJIE, LIN HAO, ZHAO MINGGUO. A fast algorithm for hand gesture recognition using relief [C]// ICNC 2009: Proceedings of the Sixth International Conference on Fuzzy Systems and Knowledge Discovery. Washington, DC: IEEE Computer Society, 2009: 8-13.

[6] 吴浩苗,漪中行,孙富春. Relief算法在笔迹识别中的应用[J]. 计算机应用,2006,26(1):174-176.

[7] KONONENKO I. Estimating attributes: Analysis and extensions of RELIEF [C]// ECML-94: Proceedings of the 1994 European Conference on Machine Learning, LNCS 784. Berlin: Springer, 1994: 171-182.

[8] ROBNIK-ŠIKONJA M, KONONENKO I. Theoretical and empirical analysis of ReliefF and RReliefF [J]. Machine Learning, 2003, 53(1/2): 23-69.

[9] SUN YI-JUN. Iterative relief for feature weighting algorithms, theories, and applications [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(6): 1035-1051.

[10] 李颖新,李建更,阮晓钢. 肿瘤基因表达谱分类特征基因选取问题及分析方法研究[J]. 计算机学报,2006,23(2):324-330.

[11] TOTH D, AACH T. Improved minimum distance classification with Gaussian outlier detection for industrial inspection [C]// Proceedings of the 11th International Conference on Image Analysis and Processing. Washington, DC: IEEE Computer Society, 2001: 584-588.

[12] CARNEIRO G, JEPSON A D. Flexible spatial configuration of local image features [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(12): 2089-2104.

[13] HIRUKAWA T, KOMADA S, HIRAI J. Image feature based navigation of nonholonomic mobile robots with active camera [C]// SICE 2007: Proceedings of the 2007 International Conference on Instrumentation, Control and Information Technology. Washington, DC: IEEE Computer Society, 2007: 2502-2506.