Journal of Computer Applications
计算机应用,2020,40(3):626 - 630
ISSN 1001-9081
CODEN JYIIDU
2020-03-10
http://www.joca.cn

文章编号:1001-9081(2020)03-0626-05
DOI:10.11772/j.issn.1001-9081.2019071200

面向二类区分能力的干扰熵特征选择方法

曾元鹏[1,2],王开军[1*],林　崧[2]
(1. 福建师范大学 数学与信息学院,福州 350007;　2. 福建师范大学 数字福建环境监测物联网实验室,福州 350007)
(*通信作者电子邮箱 wkjwang@qq.com)

摘　要:针对现有的特征选择方法对衡量不同类别数据重叠/分离能力的不足,提出了一种用于评价特征的二类区分能力的干扰熵方法(IET-CD)。对于包含两个类别(正类和负类)样本的特征,首先,计算正类数据范围内的负类样本的混合条件概率,以及负类样本归属于正类的概率;然后,由混合条件概率和归属概率计算混淆概率,再利用混淆概率计算正类干扰熵,同理,计算负类干扰熵;最后,将正、负类干扰熵之和作为该特征的二类干扰熵。干扰熵用于评价特征对二类样本的区分能力,该特征的干扰熵值小,表明该特征的二类区分能力强,反之则弱。在3个UCI数据集和1个模拟基因表达数据集上,每个方法挑选出5个最优特征,并对比了这些特征的二类区分能力,由此比较这些方法的性能。实验结果表明:所提方法与NEFS方法相比,二类区分能力相当或更好;与单索引近邻熵特征选择(SNEFS)方法、相关性最大冗余性最小特征选择(MRMR)算法、联合互信息(JMI)方法、Relief方法相比,绝大多数情况都是所提方法获胜。IET-CD方法能有效地选择二类区分能力更好的特征。

关键词:特征选择;二类区分能力;条件概率;干扰熵
中图分类号:TP391.4　　**文献标志码**:A

Interference entropy feature selection method for two-class distinguishing ability

ZENG Yuanpeng[1,2], WANG Kaijun[1*], LIN Song[2]
(1. *College of Mathematics and Informatics, Fujian Normal University, Fuzhou Fujian* 350007, *China*;
2. *Digit Fujian Internet-of-Things Laboratory of Environmental Monitoring, Fujian Normal University, Fuzhou Fujian* 350007, *China*)

Abstract: Aiming at the existing feature selection methods lacking the ability to measure the overlap/separation of different classes of data, an Interference Entropy of Two-Class Distinguishing (IET-CD) method was proposed to evaluate the two-class distinguishing ability of features. For the feature containing two classes (positive and negative), firstly, the mixed conditional probability of the negative class samples within the range of positive class data and the probability of the negative class samples belonging to the positive class were calculated; then, the confusion probability was calculated by the mixed conditional probability and attribution probability, and the confusion probability was used to calculate the positive interference entropy. In the similar way, the negative interference entropy was calculated. Finally, the sum of positive and negative interference entropies was taken as the two-class interference entropy of the feature. The interference entropy was used to evaluate the distinguishing ability of the feature to the two-class sample. The smaller the interference entropy value of the feature, the stronger the two-class distinguishing ability of the feature. On three UCI datasets and one simulated gene expression dataset, five optimal features were selected for each method, and the two-class distinguishing ability of the features were compared, so as to compare the performance of the methods. The experimental results show that the proposed method is equivalent or better than the NEFS (Neighborhood Entropy Feature Selection) method, and compared with the Single-indexed Neighborhood Entropy Feature Selection (SNEFS), feature selection based on Max-Relevance and Min-Redundancy (MRMR), Joint Mutual Information (JMI) and Relief method, the proposed method is better in most cases. The IET-CD method can effectively select features with better two-class distinguishing ability.

Key words: feature selection; two-class distinguishing ability; conditional probability; interference entropy

www.joca.cn

0　引言

特征选择是数据挖掘的预处理步骤,旨在通过删除不必要特征,以达到降低算法复杂度、提高算法性能的目的。特征子集的评价准则或启发式确定后,特征选择问题从某种意义上来说就是在特征空间中搜索最优特征子集问题[1]。

近年来,使用信息熵方法[2]来进行特征选择越来越成为广大研究者的研究热点。文献[1]提出将粗糙集相对分类信息熵作为适应度函数实施粒子群算法进行特征选择;文献[2]采用基于互信息的近邻熵进行近似的最近邻前向搜索来选择特征(Neighborhood Entropy Feature Selection, NEFS),并通过优化索引结构提出单索引近邻熵特征选择(Single-indexed

收稿日期:2019-07-10;**修回日期**:2019-09-23;**录用日期**:2019-09-24。　**基金项目**:国家自然科学基金资助项目(61672157, 61772134);福建省自然科学基金资助项目(2018J01778);中国博士后科学基金资助项目(2016M600494)。

作者简介:曾元鹏(1995—),男,福建漳州人,硕士研究生,主要研究方向:机器学习、数据挖掘;　王开军(1965—),男,福建福州人,副教授,博士,主要研究方向:机器学习、数据挖掘;　林崧(1979—),男,福建福州人,教授,博士,主要研究方向:量子密码、量子智能计算。

Neighborhood Entropy Feature Selection, SNEFS)方法来选择更好的分类特征。互信息在相关性分析上有计算简单、可解释性强特点,因此基于互信息的特征选择方法被广泛应用于特征选择[3]。相关性最大冗余性最小特征选择(feature selection based on Max-Relevance and Min-Redundancy, MRMR)算法[4]计算待选特征与类别的互信息、待选特征与已选特征间的平均互信息,以两者的差值是否为最大值来决定该特征可否加入最优特征子集中;但是该方法没有对特征数据在类别上的分布情况作出合适判断,无法体现出数据的混乱情况。与MRMR方法类似的还有文献[5]提出的特征排序方法,度量特征间的独立和冗余程度,考虑了特征子集中不同特征间的多变量关系。文献[6]运用了互信息将特征进行排序,将排序后的特征逐个加入特征子集进行分类,去掉每次可以正确分类的样本,从而避免信息的冗余。联合互信息(Joint Mutual Information, JMI)方法[7]选取信息增益最大的特征作为第一个特征;接着计算待选特征的信息增益;然后,计算在已选特征子集条件下,待选特征与类别的条件互信息均值;最后,将两者之和最大的待选特征加入已选特征子集;该方法在不同类别数据的区分能力上存在欠缺。与该方法部分类似的有文献[8]提出在特征和类别之间的联合最大信息熵方法度量特征子集并用粒子群优化方法来搜索最优特征子集。Relief方法[9]提出先随机选取一个样本点,并从同类中寻找最近的样本,再寻找最近的非同类样本计算其距离,如果同类距离较小,则认为该特征对区分不同类别样本更有能力,增加该特征的权重;该方法以近邻的方式去判断二类的区分性,未从数据分布的整体情况进行考虑。文献[10]提出以相对分类信息熵作为适应度函数度量特征子集的重要性,使用遗传算法搜索最优特征子集;文献[11]假设特征之间相互独立,使用特征与标签集合之间的信息增益来衡量特征与标签集合之间的重要程度,然后采用信息增益阈值选择特征。文献[12]在K-近邻互信息的基础上将其扩展至高维,并采用前向搜索和后向交叉相结合的形式来获得最优强相关特征子集。文献[13]利用模糊等价关系将互信息推广至模糊互信息,接着计算该模糊度量下的候选与已选特征的相关关系,最后采用最大最小准则与前向贪婪算法来选出最优特征子集。

现有方法主要测量备选特征和类别之间的信息熵来筛选最优特征,但没有评价某个特征在类别区分上的能力和贡献大小,这造成所选特征的缘由是模糊的、不清晰的。本文方法的好处是对某个特征区分不同类别数据的能力给出了一个明确的评价指标,使得挑选某个特征具有明确的依据和含义。本文提出了一种面向二类区分能力的干扰熵方法(Interference Entropy of Two-Class Distinguishing, IET-CD),对每个特征下两类数据的相互干扰程度采用干扰熵值进行量化,使用干扰熵指标作为不同类别数据区分能力的评价指标,当干扰熵值越小,说明该特征的二类区分能力强,反之则弱,为寻找二类数据区分明显的特征提供了方法。

1　干扰熵方法的相关定义

给定具有n个m维样本的二类数据集$\boldsymbol{D}=[\boldsymbol{d}_1,\boldsymbol{d}_2,\cdots,\boldsymbol{d}_m]$,含有正类(A类)样本$X$和负类(B类)样本$Y$。对于数据集$\boldsymbol{D}$中的$\boldsymbol{d}_i(i\leqslant m)$维(特征)上的数据,以样本均值$u_1$较小的一类为正类,样本均值$u_2\geqslant u_1$的一类为负类,$k$个正类样本记为$X=\{X_1,X_2,\cdots,X_k\}$,$n-k$个负类样本记为$Y=\{Y_1,Y_2,\cdots,Y_{n-k}\}$,正、负类样本的取值范围为$[X_{\min},X_{\max}]$与$[Y_{\min},Y_{\max}]$,在正类范围内出现的负类样本为$Y^+=\{X_{\min}\leqslant Y_1^+,\cdots,Y_j^+\leqslant X_{\max}\}$,$j\leqslant n-k$,$Y^+\subset Y$,在负类范围内的出现正类样本为$X^-=\{Y_{\min}\leqslant X_1^-,\cdots,X_l^-\leqslant Y_{\max}\}$,$l\leqslant k$,$X^-\subset X$。

定义 1　混合条件概率。在正类数据取值范围内,出现正类样本X概率为$P(X)=k/(k+j)$,出现负类样本Y^+的概率为$P(Y^+)=j/(k+j)$,则正负类样本出现的联合概率为$P(X,Y^+)$,出现负类样本Y^+的混合条件概率为$P(Y^+|X)$,公式如下:

$$P(X,Y^+)=\frac{C_k^1*C_j^1}{C_{k+j}^1*C_{k+j-1}^1}=\frac{k*j}{(k+j)*(k+j-1)} \tag{1}$$

$$P(Y^+|X)=\frac{P(X,Y^+)}{P(X)}=\frac{j}{k+j-1} \tag{2}$$

定义 2　样本归属正类的概率。设在正类数据范围出现负类样本点$Y_i^+(i=1,2,\cdots,j)$,混淆区域为$[Y_i^+,X_{\max}]$;负类数据范围出现正类样本点$X_i^-(i=1,2,\cdots,l)$,混淆区域为$[Y_{\min},X_i^-]$。统计在正类混淆区域内的r个正类样本和t个负类样本,正、负类样本出现在混淆区域的概率分别记为$P(A^*)=r/(r+t)$、$P(B^*)=t/(r+t)$。负类样本Y_i^+在正、负类数据的高斯概率密度下属于正、负类的概率记为$P(Y_i^+|A^*)$、$P(Y_i^+|B^*)$,则样本归属正类概率记为$P(A^*|Y_i^+)$,类似地计算出样本归属负类概率$P(B^*|X_i^-)$,公式定义如下:

$$\begin{cases}P(Y_i^+|A^*)=\int_{Y_i^+}^{+\infty}f(x)\mathrm{d}x\\P(Y_i^+|B^*)=\int_{-\infty}^{Y_i^+}g(x)\mathrm{d}x\end{cases} \tag{3}$$

$$P(A^*|Y_i^+)=\frac{P(A^*)P(Y_i^+|A^*)}{P(A^*)P(Y_i^+|A^*)+P(B^*)P(Y_i^+|B^*)} \tag{4}$$

其中:$f(x)$是具有均值u_1和标准差σ_1的正类数据高斯函数,$g(x)$是具有均值u_2和标准差σ_2的负类数据高斯函数。

在上述混合条件概率$P(Y^+|X)$和样本归属正类的概率$P(A^*|Y_i^+)$的基础上,设计混淆概率。

定义 3　混淆概率。在正类范围内,负类样本被误认为是正类样本的混淆概率$P(Y_i^+)$为混合条件概率和样本归属正类的概率的乘积,即:

$$P(Y_i^+)=P(Y^+|X)P(A^*|Y_i^+);\ i=1,2,\cdots,j \tag{5}$$

定义 4　类间干扰熵。在正类范围内,正类干扰熵$H(Y^+)$为每个负类样本点的干扰熵之和;类似地,在负类范围内,负类干扰熵$H(X^-)$为每个正类样本点的干扰熵之和。类间干扰熵$H(Y^+,X^-)$定义如下:

$$H(Y^+)=-\sum_{i=1}^{j}P(Y_i^+)\mathrm{lb}\,P(Y_i^+) \tag{6}$$

$$H(X^-)=-\sum_{i=1}^{l}P(X_i^-)\mathrm{lb}\,P(X_i^-) \tag{7}$$

$$H(Y^+,X^-)=H(Y^+)+H(X^-) \tag{8}$$

对于二值的离散型特征,设计如下的定义用于计算混淆概率。

定义 5　离散特征混合条件概率。对于取值为0、1的离散型特征,设该特征上正类样本集A中值为0和1的样本个数分别为a_0和a_1,且$a_0\geqslant a_1$,负类样本集B中值为0和1的样本个数分别为b_0和b_1。则值为0的样本x属于正、负类的概率分别为$P_A(x=0)=a_0/(a_0+a_1)$、$P_B(x=0)=b_0/(b_0+b_1)$,则值为1的样本x属于正、负类的概率分别为$P_A(x=1)=a_1/(a_0+a_1)$、$P_B(x=1)=b_1/(b_0+b_1)$。在A中出现值为1和0样本的混合条件概率分别为$P(x=1|A)=a_1/(a_0+a_1)$、$P(x=0|A)=a_0/(a_0+a_1)$;同理有$P(x=1|B)=b_1/(b_0+b_1)$和$P(x=0|B)=b_0/(b_0+b_1)$。

定义 6　离散特征的混淆概率和干扰熵。设一个特征上正类样本子集A有$a_0\geqslant a_1$,则值为1的样本的混合条件概率与样本归属于正类的概率的乘积称为混淆概率$P(A_1^-)=P(x=1|A)P_A(x=1)$,同理有混淆概率$P(B_0^+)=P(x=0|B)P_B(x=0)$。则干扰熵定义为$H(A^-,B^+)=H(B^+)+H(A^-)$,其中$H(B^+)=$

www.joca.cn

$-a_0P(B_0^+)\mathrm{lb}\,P(B_0^+)$、$H(A^-)=-b_1P(A_1^-)\mathrm{lb}\,P(A_1^-)$。

2 面向二类区分能力的干扰熵方法

面向二类区分能力的干扰熵方法以类间干扰熵来衡量该特征所包含的两类数据混淆情况。对于二类数据集中的某一维度/特征，若类间干扰熵大，说明两个类别的数据混合程度大，二类区分性弱；若小，说明两个类别的数据混合程度小，二类区分性强。

当负类样本出现在正类数据范围内时，正负样本的数据范围出现重叠，使二类数据的区分度减弱。正类数据范围内的负类样本数量对特征二类区分能力的影响采用定义1或5的混合条件概率进行量化计算。

当负类样本出现在正类数据范围内不同位置时，干扰程度不同，越远离负类样本聚集的一侧，被误认为是正类的可能性越大，导致正负类数据不易区分。采用定义2或5的样本归属正类的概率来衡量在正类数据范围内负类样本归属于正类的程度，当概率值越大，说明该负类样本被误认为正类的概率越大，对数据区分的影响越大。

混合条件概率衡量了在正类数据范围内混入负类样本的数量情况，样本归属正类的概率考虑到混入的负类样本归属于正类的程度，将上述概率的乘积作为两类数据的混淆情况的量化，通过定义3或6的公式得到相应的混淆概率。

评价一个特征的二类区分能力的干扰熵方法步骤如下：

依据上文的定义，对于连续型数据有：

1）选择数据集D中的一个特征，分别计算正、负类样本的取值范围，统计正类范围内的负类样本。

2）计算正类数据范围内正、负类出现的概率$P(X)$、$P(Y^+)$，联合概率$P(X,Y^+)$，以及混合条件概率$P(Y^+|X)$。

3）在正类数据范围内，计算样本归属正类的概率$P(A^*|Y_i^+)$。对称地，在负类数据范围内，计算样本归属负类的概率$P(B^*|X_i^-)$。

4）计算正类数据范围内的负类样本被误认为是正类样本的混淆概率$P(Y_i^+)$。对称地，在负类数据范围内，计算正类样本被误认为是负类样本的混淆概率$P(X_i^-)$。

5）依据混淆概率计算类间干扰熵$H(Y^+,X^-)$。

对于离散型数据，类似地依据上文的定义计算混淆概率和类间干扰熵。

3 实验分析

为了检验本文IET-CD方法的有效性，利用IET-CD方法计算每个特征的类间干扰熵，按熵值从小到大排序，选取熵值小（二类区分程度大）的特征组成特征子集，将该特征子集用于评价一个特征二类区分能力的实验。类似地，5个对比方法——NEFS、SNEFS、MRMR、JMI、Relief特征选择算法也给出特征排序结果，取排序位置最前的几个特征组成特征子集，用于二类区分能力的分析。

实验选取3个UCI（University of California Irvine）[14]数据集Sonar、WDBC（Wisconsin Diagnostic Breast Cancer）、Ionosphere和4k20lap模拟基因表达数据集[15]进行测试，其中4k20lap1_2是4k20lap数据集中的第1和2类组成。Sonar、WDBC、Ionosphere、4k20lap1_2数据集的对应的样本数分别是208、569、351、134，特征数分别为60、30、34、20，类别数均为2。本实验的特征选择算法（IET-CD）的实现代码由作者使用Python语言编写。对比方法相关性最大冗余性最小方法（MRMR）、联合互信息方法（JMI）、Relief程序来源于Matlab的FEAST（FEAture Selection Toolbox）工具箱，这3个方法参数均采用工具箱默认参数。NEFS与SNEFS算法代码来源于文献[2]，并设定特征近邻参数为4。最后选取各方法所得前5个最优特征进行实验。

本实验通过计算特征的干扰熵来衡量该特征中两类数据的分离/重叠程度，干扰熵值越小，特征区分能力（分离程度）越好。使用特征的数据散点图来验证以干扰熵值作为评判标准的特征区分能力的正确性，横坐标表示样本编号，纵坐标表示样本值。

在不同的数据集上对5个不同的特征选择算法所选特征的进行干扰熵的计算，实验结果包含对应的数据集、特征选择算法、所选特征中前5个特征的编号，所选特征对应的熵值，用于对比的配对特征、本文方法获胜情况，见表1。图1和图2列出Sonar、WDBC对比特征散点图用于验证本文方法的正确性。

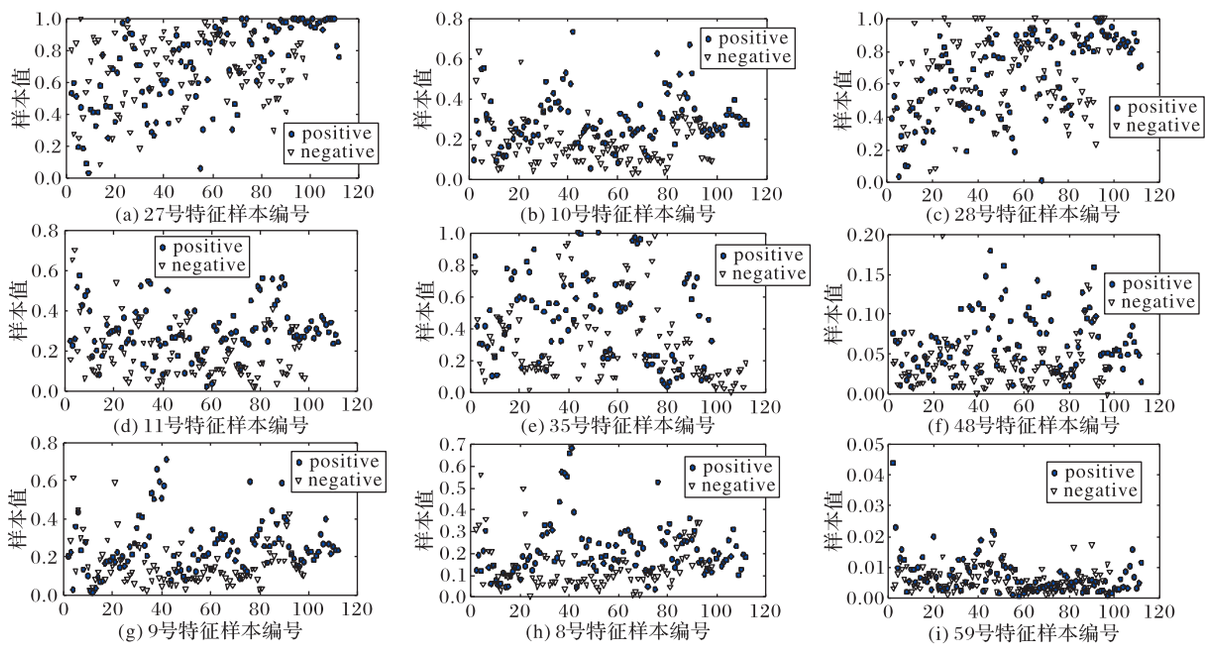


图1 Sonar数据集对比特征的散点图

Fig. 1 Scatter plot of contrast features comparison on dataset Sonar

www.joca.cn

表 1　4个数据集上各方法所选特征的二类区分能力对比

Tab. 1　Two-class distinguishing ability comparison of selected features of each method on four datasets

数据集	方法	前5个特征编号	特征对应熵值	对比的配对特征	IET-CD获胜情况
Sonar	IET-CD	10,11,48,9,8	49.6,53.6,59.1,59.4,62.5	(10,11),(11,48),(48,9),(9,8)	胜,胜,相同,胜
	NEFS	11,10,8,48,9	53.6,49.6,62.5,59.1,59.4	(10,11),(11,10),(48,8),(9,48),(8,9)	胜,败,胜,相同,败
	SNEFS	8,3,10,1,28	62.5,71.1,49.6,75.9,85.2	(10,8),(11,3),(48,10),(9,1),(8,28)	胜,胜,败,胜,胜
	MRMR	27,28,35,21,14	77.1,85.2,65.3,71.6,83.2	(10,27),(11,28),(48,35),(9,21),(8,14)	胜,胜,胜,胜,胜
	JMI	27,28,35,21,14	77.1,85.2,65.3,71.6,83.2	(10,27),(11,28),(48,35),(9,21),(8,14)	胜,胜,胜,胜,胜
	Relief	11,59,9,8,10	53.6,89.6,59.4,62.5,49.6	(10,11),(11,59),(48,9),(9,8),(8,10)	胜,胜,相同,胜,败
WDBC	IET-CD	27,22,20,7,23	30.0,30.7,33.7,34.5,37.8	(27,22),(22,20),(20,7),(7,23)	相同,胜,胜,胜
	NEFS	22,20,27,7,23	30.7,33.7,30.0,34.5,37.8	(27,22),(22,20),(20,27),(7,7),(23,23)	相同,胜,败,相同,相同
	SNEFS	22,27,20,0,23	30.7,30.0,33.7,54.4,37.8	(27,22),(22,27),(20,20),(7,0),(23,23)	相同,相同,相同,胜,相同,
	MRMR	23,4,5,6,7	37.8,152.5,88.9,47.1,34.5	(27,23),(22,4),(20,5),(7,6),(23,7)	胜,胜,胜,胜,胜
	JMI	23,3,21,22,13	37.8,56.5,123.6,30.7,69.5	(27,23),(22,3),(20,21),(7,22),(23,13)	胜,胜,胜,胜,胜
	Relief	25,27,24,26,5	88.6,30.0,139.7,54.4,88.9	(27,25),(22,27),(20,24),(7,26),(23,5)	胜,相同,胜,胜,胜
Ionosphere	IET-CD	2,6,4,14,0	94.9,100.7,101.1,113.6,113.7	(2,6),(6,4),(4,14),(14,0)	胜,相同,胜,相同
	NEFS	5,4,3,33,0	147.7,101.1,145.3,142.8,113.7	(2,5),(6,4),(4,3),(14,33),(0,0)	胜,相同,胜,胜,相同
	SNEFS	5,4,15,13,7	147.7,101.1,142.9,138.8,135.8	(2,5),(6,4),(4,15),(14,13),(0,7)	胜,相同,胜,胜,胜
	MRMR	26,33,0,4,3	124.4,142.8,113.7,101.1,145.3	(2,26),(6,33),(4,0),(14,4),(0,3)	胜,胜,胜,败,胜
	JMI	26,4,0,2,6	124.4,101.1,113.7,94.9,100.7	(2,26),(6,4),(4,0),(14,2),(0,6)	胜,相同,胜,败,败
	Relief	1,4,23,9,14	170.9,101.1,150.5,148,8,113.6	(2,1),(6,4),(4,23),(14,9),(0,14)	胜,相同,胜,胜,相同
4k20lap 1_2	IET-CD	6,3,16,17,1	16.9,19.1,22.2,26.5,27.3	(6,3),(3,16),(16,17),(17,1)	胜,胜,胜,相同
	NEFS	6,1,16,3,11	16.9,27.3,22.2,19.1,31.7	(6,6),(3,1),(16,16),(17,3),(1,11)	相同,胜,相同,败,胜
	SNEFS	6,3,7,8,18	16.9,19.1,35.3,32.5,30.8	(6,6),(3,3),(16,7),(17,8),(1,18)	相同,相同,胜,胜,胜
	MRMR	7,15,16,17,6	35.3,35.1,22.2,26.5,16.9	(6,7),(3,15),(16,16),(17,17),(1,6)	胜,胜,相同,相同,败
	JMI	7,2,15,12,17	35.3,43.7,35.1,38.7,26.5	(6,7),(3,2),(16,15),(17,12),(1,17)	胜,胜,胜,胜,相同
	Relief	17,3,8,6,1	26.5,19.1,32.5,16.9,27.3	(6,17),(3,3),(16,8),(17,6),(1,1)	胜,相同,胜,败,相同

www.joca.cn

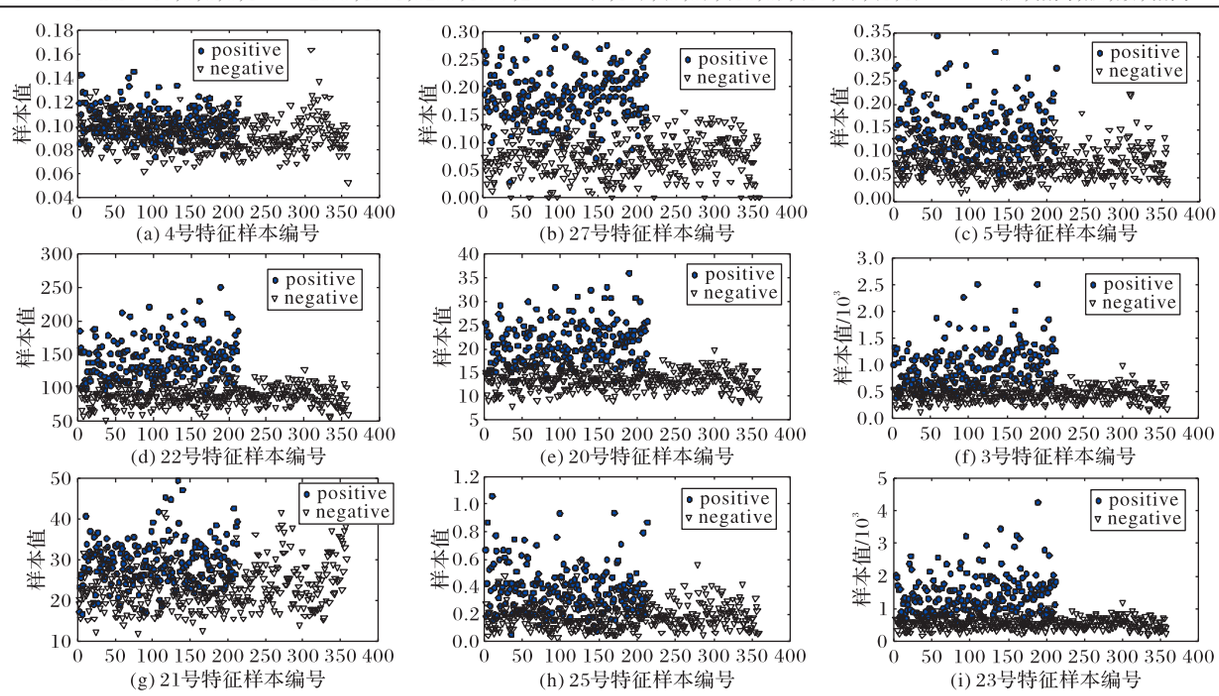


图2　WDBC数据集对比特征的散点图

Fig. 2　Scatter plot of contrast features comparison on dataset WDBC

从表1可知,在Sonar数据集上,10号特征比11号特征的干扰熵值小,反映出10号特征二类区分能力强,通过对比图1中的数据散点图,可看出10号特征比11号特征的数据重叠区域小,验证了干扰熵值小其二类区分能力强。对11号特征与59号特征进行对比,11号特征获胜,在图1数据散点图中也反映了11号特征数据重叠区域小,混淆程度小,说明本文IET-CD方法的性能优于对比方法。

WDBC数据集上的27号和22号特征,Ionosphere数据集上的6号和4号特征的情况与Sonar数据集上的48号和9号特征情况相似,两个特征的干扰熵值相差较小,说明两个特征的二类区分能力相近,通过相应的散点图反映的结果可以看出,特征的数据重叠情况相近,验证了干扰熵值相近的两个特征的二类区分能力相似,故对比结果表现相同。对表1中的WDBC数据集上的5号与20号特征对比,20号特征获胜,说明20号特征的二类区分能力强,从图2的数据散点图可看出20号特征的数据重叠区域小,逆向验证了20号特征干扰熵值小,5号特征干扰熵值大,以此证明干扰熵值能正确反映二类数据的区分程度。

表1中的Ionosphere数据集和4k20lap1_2模拟基因数据集仍以相同的方式进行对比,故不再一一列出对应的散点图。因为4k20lap1_2模拟基因数据集的部分特征二类样本混淆数量较少,所以在该数据集上反映出的特征干扰熵值也相对较小,符合干扰熵的定义。4k20lap1_2模拟基因数据集的3号特征比15号特征的干扰熵值小,表明3号特征的二类分区能力强。因3号特征的不同类别数据分离情况较好,15号特征的不同类别数据混合程度比3号特征混乱,故3号特征的区分能力胜过15号特征。由表1中的配对特征的对比结果可看出,3号特征的区分能力胜过15号特征,其特征对应的干扰熵值能正确反映二类数据的区分程度。其他特征类似对比,并通过散点图可验证其结论的正确性。

4 结语

本文通过计算特征的二类干扰熵来进行数据区分能力的判断从而进行特征选择,选出区分能力较好的特征。干扰熵值越小区分能力越强,干扰熵值越大区分能力越弱,该方法能有效筛选二类混合程度大、区分度不明显的特征,发现二类区分能力好的特征,为识别特征的二类区分能力提供了评价指标。

参考文献(References)

[1] 翟俊海,刘博,张素芳. 基于粗糙集相对分类信息熵和粒子群优化的特征选择方法[J]. 智能系统学报, 2017, 12(3):397-404. (ZHAI J H, LIU B, ZHANG S F. A feature selection approach based on rough set relative classification information entropy and particle swarm optimization[J]. CAAI Transactions on Intelligent Systems, 2017, 12(3):397-404.)

[2] MARIELLO A, BATTITI R. Feature selection based on the neighborhood entropy[J]. IEEE Transactions on Neural Networks and Learning Systems, 2018, 29(12): 6313-6322.

[3] 毛莺池,曹海,平萍,等. 基于最大联合条件互信息的特征选择[J]. 计算机应用, 2019, 39(3):734-741. (MAO Y C, CAO H, PING P, et al. Feature selection based on maximum conditional and joint mutual information[J]. Journal of Computer Applications, 2019, 39(3):734-741.)

[4] PENG H, LONG F, DING C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(8):1226-1238.

[5] 董红斌,滕旭阳,杨雪. 一种基于关联信息熵度量的特征选择方法[J]. 计算机研究与发展, 2016, 53(8):1684-1695. (DONG H B, TENG X Y, YANG X. Feature selection based on the measurement of correlation information entropy[J]. Journal of Computer Research and Development, 2016, 53(8):1684-1695.)

[6] 马忱,姜高霞,王文剑. 面向函数型数据的动态互信息特征选择方法[J]. 计算机科学与探索, 2019, 13(1):158-168. (MA C, JIANG G X, WANG W J. Dynamic mutual information feature selection for functional data[J]. Journal of Frontiers of Computer Science and Technology, 2019, 13(1): 158-168.)

[7] YANG H H, MOODY J. Data visualization and feature selection: new algorithms for nonGaussian data[C]// Proceedings of the 12th International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 1999:687-693.

[8] ZHENG K, WANG X. Feature selection method with joint maximal information entropy between features and class[J]. Pattern Recognition, 2018, 77: 20-29.

[9] KIRA K, RENDELL L A. A practical approach to feature selection[C]// Proceedings of the 9th International Workshop on Machine Learning. San Francisco: Morgan Kaufmann Publishers Inc., 1992:249-256.

[10] 翟俊海,刘博,张素芳. 基于相对分类信息熵的进化特征选择算法[J]. 模式识别与人工智能,2016, 29(8):682-690. (ZHAI J H, LIU B, ZHANG S F. Feature selection via evolutionary computation based on relative classification information entropy[J]. Pattern Recognition and Artificial Intelligence, 2016, 29(8): 682-690.)

[11] 张振海,李士宁,李志刚,等. 一类基于信息熵的多标签特征选择算法[J]. 计算机研究与发展, 2013, 50(6):1177-1184. (ZHANG Z H, LI S N, LI Z G, et al. Multi-label feature selection algorithm based on information entropy[J]. Journal of Computer Research and Development, 2013, 50(6):1177-1184.)

[12] 周红标,乔俊飞. 基于高维k-近邻互信息的特征选择方法[J]. 智能系统学报, 2017, 12(5):595-600. (ZHOU H B, QIAO J F. Feature selection method based on high dimensional k-nearest neighbors mutual information[J]. CAAI Transactions on Intelligent Systems, 2017, 12(5):595-600.)

[13] 董泽民,石强. 基于归一化模糊联合互信息最大的特征选择[J]. 计算机工程与应用, 2017, 53(22):105-110. (DONG Z M, SHI Q. Feature selection using normalized fuzzy joint mutual information maximum[J]. Computer Engineering and Applications, 2017, 53(22):105-110.)

[14] BLAKE C L, MERZ C J. UCI repository of machine learning databases[EB/OL]. [2019-05-13]. http://mlearn.ics.uci.edu/MLRepository.html.

[15] WANG K, ZHENG J, ZHANG J, et al. Estimating the number of clusters via system evolution for cluster analysis of gene expression data[J]. IEEE Transactions on Information Technology in Biomedicine, 2009, 13(5):848-853.

This work is partially supported by the National Natural Science Foundation of China (61672157, 61772134), the Natural Science Foundation of Fujian Province (2018J01778), the China Postdoctoral Science Foundation (2016M600494).

ZENG Yuanpeng, born in 1995, M. S. candidate. His research interests include machine learning, data mining.

WANG Kaijun, born in 1965, Ph. D., associate professor. His research interests include machine learning, data mining.

LIN Song, born in 1979, Ph. D., professor. His research interests include quantum cryptography, quantum intelligent computation.

www.joca.cn