doi:10.3772/j.issn.1002-0470.2014.11.015

考虑稳定性要求的特征选择方法①

季金胜② 郭艺友 霍 宏 方 涛③

(*上海交通大学自动化系 上海 200240)

(**系统控制与信息处理教育部重点实验室 上海 200240)

摘 要 为了提高特征选择的稳定性和降低因样本数据变化引起的选择结果波动,提出了一种考虑稳定性要求的过滤式特征选择方法。不同于集成特征选择等现有的增强稳定性方法,该方法将特征的稳定性与相关性、冗余性一起作为特征评价准则,通过产生多个数据集来减少样本数据扰动,不断将新产生的选择结果迭代计算稳定性因子,并同时提高其在准则中的比重以使迭代收敛。最终将融合多次迭代信息的特征排序作为最终结果输出。实验表明,该方法能够在保持相当分类精度的基础上,能够较大幅度地提高选择结果的稳定性,达到兼顾分类精度与稳定性的目的。

关键词 特征选择, 相关性, 冗余性, 稳定性, 高维数据

0 引 言

随着新技术的不断出现,地球科学[1]、生物学[2]等领域数据呈现"爆炸式"增长,产生了大量的高维、低样本数据。如何从这些高维数据中挖掘出有用信息则成为机器学习亟待解决的核心问题[3]。特征选择(feature selection)就是在这种情况下提出来的,通过特征选择,可剔除冗余或无关的特征,降低特征维数,避免维数灾难[4,5]。特征选择是指根据一定的准则从原始高维数据中寻找最优特征子集,在最大限度地保留信息的同时大大降低数据的维数[6]。按照和分类器结合方式,特征选择分为过滤式(Filter)、嵌入式(Embedded)和封装式(Wapper)[7,8]。其中,尤以独立于分类器的 Filter 最受关注。当前,比较经典的有 Relief 算法[9]及其变形[10]和最大相关与最小冗余(max-relevance and min-redundancy, mRMR)[11]算法。前者是利用同类与不同类样本间距离,通过聚类不断更新特征的权重,进而达到特征选择的目的。后者则是基于最大相关与最小冗余(mRMR)原则,利用互信息,保证所选特征包含最大的分类信息且特征间的冗余性最小,这是一种应用最广泛的特征选择方法。

现有的特征选择算法偏重于以分类精度评价特征选择的性能,忽视了特征选择的稳定性要求,有时样本数据的变化可能会致使选择结果具有很差的稳定性,容易产生误导性,大大降低了其实用性。比如,地球科学领域的典型地物特性分析中,特征选择的目的是为了寻找反映各种地物的关键特征,若选出的结果不稳定,就很难判定哪些是关键特征。为此,特征选择过程中容易被忽视且又很关键的稳定性问题进入学者们的研究视野。特征选择算法的稳定性可定义为它对同分布产生的不同训练数据集的鲁棒性[12]。现有的特征选择稳定性研究主要是通过提出的稳定性度量指标评价特征选择方法的稳定性以及通过集成选择结果来提高特征选择的稳定性[12,13],还没有将稳定性作为特征的评价准则贯穿特征选择过程。为此,本文力求兼顾稳定性和分类精度,提出了一种基于最大相关、最大稳定性与最小冗余性(max-relevance, max-stability and min-redundancy, mRSmR)准则的 Filter 特征选择方法。该 mRSmR 方法将稳定性与相关性、冗余性同时作为特征选择的内部评价准则,综合这三个因素对特征分析评价,进而得出选择结果。利用支持向量机

① 973 计划(2012CB719903),国家自然科学基金委创新研究群体(61221003),国家自然科学基金青年科学基金(41101386)和国家自然科学基金(41071256)资助项目。

② 男,1990 年生,硕士;研究方向:特征选择,数据挖掘,计算机视觉;E-mail: jinsheng0629@163.com

③ 通讯作者,E-mail: tfang@sjtu.edu.cn

(收稿日期:2014-09-22)

(SVM)分类与常见稳定性度量指标进行的评价表明,该方法能够在达到相当的分类精度的同时较大地提高选择结果的稳定性。

1 相关工作

1.1 Filer 特征选择算法:mRMR 算法和 Relief 算法

现有 Filter 算法模型有不同的评价准则,包含距离标准、一致性标准、依赖性标准和信息标准。Relief 算法及其变种 ReliefF 算法和 IRelief 算法等就是采用样本距离来度量特征的重要性程度。Kira[9]最早提出的 Relief 算法仅限于两类问题,为此,Kononeil[10]对其进行了扩展,得到了适用于多类问题的 ReliefF 算法,是一种特征权重算法。该算法将特征对近距离样本的区分能力视为特征对类别的依赖性,依据依赖性赋予特征不同的权重。从训练样本集中随机选择一个样本 s,再从同类 C_s 的样本集中找到 s 的 k 个最近邻;另外,从与 s 不同类的每个样本集中均找到 k 个近邻样本,最后更新每个特征的权重;通过 m 次重复抽样,不断更新特征的权重,最终得到相应的特征排序。特征 f 的迭代权重增量 $\Delta W(f)$ 如下式所示:

$$\Delta W(f) = \sum_{i=1}^{m}\left\{-\sum_{j=1}^{k}\frac{diff(f, s, Ns_j)}{k} + \sum_{C\neq C_s}\left[\frac{p(C)}{1-p(C_s)}\sum_{j=1}^{k}diff(f,s,Nns_j(C))/k\right]\right\}/m \tag{1}$$

式中样本同类别的最近邻求和为负,不同类别的最近邻求和为正,通过迭代求和评价特征 f 的分类性能。Ns_j 表示与 s 同类别的第 j 个最近邻样本,$Nns_j(C)$ 表示其它类别 C 中与样本 s 的第 j 个最近邻样本,$p(C_s)$ 为属于类别 C_s 样本所占总样本的比例,$p(C)$ 为其它类别 C 样本所占总样本比例。样本 s_1 和样本 s_2 在特征 f 上的差定义为

$$diff(f,s_1,s_2) = \begin{cases}\dfrac{|fs_1 - fs_2|}{f_{\max} - f_{\min}}, & \text{特征 } f \text{ 为连续}\\ 1, & \text{特征 } f \text{ 为离散且 } fs_1 \neq fs_2\end{cases} \tag{2}$$

式中 fs_1 与 fs_2 分别为样本 s_1 与 s_2 的特征值,$f_{\max}$ 与 $f_{\min}$ 为对应所有样本的最大与最小特征值。

Relief 及其变种都是选择包含分类信息最多的特征组合作为最优特征子集,但是,这些信息量最多的 m 个特征组合在一起,并不一定是最好的 m 特征组。Peng 等人[11]认为仅仅考虑特征子集所包含的类别信息多少是不够的,为确保特征选择结果是最好的 m 特征组,需要进一步考虑该子集特征间的冗余性,于是提出了 mRMR 算法。考虑到算法的复杂度,将寻找最优子集进一步转化为逐步从候选特征中寻找当前子集下新增的最优特征。首先,选择与类别 C 互信息最大的特征 f 作为第一个特征子集 F;其次,计算剩余特征 f 与类别 C 的互信息 $D(f, C)$ 以及与已选择的特征子集 F 之间冗余度 $R(f, F)$;依据 $D(f, C) - R(f, F)$ 或者 $D(f, C)/R(f, F)$ 最大,选择满足该准则的特征作为新增的最优特征;最后重复进行直到所有特征选择完毕。

1.2 特征选择的稳定性

特征选择的稳定性是指特征选择结果对训练样本变化的不敏感,也就是说,如果一个特征选择算法不具有稳定性,则当加入或者减少一些样本后它的结果是不可重复的,甚至训练集没有发生变动结果也可能不同。

通常造成特征选择不稳定的原因主要来自于两个方面:(1)算法设计本身就没有考虑到稳定性问题;(2)高维小样本问题[14]。在实际应用中,更多的是由于选择算法本身并没有考虑数据样本的变化会致使选择结果的波动[15,16]。若选出的特征子集不具有可重复性,将影响我们通过特征选择对数据的分析。因此,针对特征选择的稳定性问题,目前国内外很多学者主要研究了稳定性评价和通过集成特征选择方法获得稳定特征子集。在度量特征选择的稳定性方面,学者们提出了不少度量准则。由于特征选择结果的输出形式有特征子集、特征排序向量和特征权重向量三种形式[12],相应的稳定性度量准则也对应这三种形式,具体可见文献[17]。本文关注的是具备一定分类精度的特征子集的稳定性,在此主要讨论关于特征子集的稳定性度量准则。

目前大多数稳定性度量都是利用两个特征子集来定义的[15],虽然不同准则侧重点略有不同,但都是通过比较子集间的共有特征来表征选择结果的稳定性。表 1 列出的是一些常用的稳定性准则,其中 $|F \cup F'|$ 为特征子集 F 与 F' 并集的大小,X 为已有特征子集的并集,Fre_f 为特征 f 在所有特征子集的频数,ω 为特征子集个数,$S(F_i, F_j)$ 为两个特征子集的相似性度量。

前面仅仅是用提出的各种不同的稳定性指标对特征选择算法进行稳定性评价,此外,目前针对特征

选择的稳定性问题,学者们已经提出了很多解决方法,主要有集成的特征选择方法(ensemble feature selection, EnsembleFS)[22]、先验特征相关性[23]、Group 特征选择[24]和样本注入[25]等方法。在这几种方法当中,研究最多的是集成的特征选择(EnsembleFS),它考虑样本变化引起的不稳定性问题,通过集成不同数据集下选择结果来提高特征选择的稳定性,但是并没有关注如何从算法本身来提高其选择结果的稳定性。

表 1 特征子集的稳定性度量准则

种类	公式
相对 Hamming 距离[18]	$1-\frac{\lvert F\backslash F'\rvert+\lvert F'\backslash F\rvert}{\lvert F\cup F'\rvert}$
加权一致性[19]	$\sum_{f\in X}\frac{Fre_f}{\lvert X\rvert}\cdot\frac{Fre_f-1}{\omega-1}$
平均 Tanimoto 指标[20,21]	$\frac{2}{\omega(\omega-1)}\sum_{i=1}^{\omega-1}\sum_{j=i+1}^{\omega}S(F_i,F_j)$

总之,以 mRMR 为代表的 Filer 特征选择方法缺乏稳定性评价准则。对一个特征选择算法来说,仅有高分类精度往往是不够的,同时,抛开分类精度只谈稳定性也是不可取的,只有将稳定性和分类精度一起考虑才有意义。因此,本文提出了一种与上面的集成特征选择方法不同的 mRSmR 方法,它在 mRMR 方法的基础上,将稳定性评价指标作为特征选择准则,通过迭代过程,在稳定性和分类精度之间折中,同时考虑相关性、冗余性和稳定性,因而能够获取更好地反映数据特性的特征选择结果。

2 增加稳定准则的特征选择

现有的如 mRMR、Relief 等经典特征选择方法主要是将子集的分类精度作为特征选择的目标,并没有考虑选择结果的稳定性问题。集成特征选择方法(EnsembleFS)虽能有效解决因训练数据集和选择方法变化而带来的稳定性问题,但其仍未把稳定性作为一个准则纳入到特征选择方法中。本文提出基于最大相关性、最大稳定性与最小冗余性(mRSmR)准则的特征选择方法,把稳定性、相关性和冗余性一起作为评价特征的准则,利用随机生成的样本集来减少数据扰动,并相应地把稳定性因子纳入特征选择过程,兼顾特征子集的可预测性和稳定性。

2.1 最大相关性、稳定性和最小冗余度

与以往解决稳定性问题的方法不同,mRSmR 方法将稳定性与相关性、冗余性同时作为特征选择的内部评价准则,综合相关性、冗余性和稳定性三个因素对特征分析评价,进而得出选择结果。

特征与类别相关性度量[11]如下式所示:

$$D(f,C)=mi(f,C) \tag{3}$$

式中 $mi(f,C)$ 为候选特征 f 与类别 C 之间互信息。

特征之间冗余度度量[11]如下式所示:

$$R(f,F)=\frac{1}{\lvert F\rvert}\sum_{f_i\in F}mi(f,f_i) \tag{4}$$

式中 f_i 是属于已选特征子集 F 的特征,$\lvert F\rvert$ 表示已选特征子集 F 的大小。

对于已选的多个特征子集,当前主要通过稳定性度量平均 Tanimoto 指数(average Tanimoto index, ATI)评价所选择的特征子集的稳定性[20]。为了便于在特征选择过程中引入稳定性评价指标,我们重新定义了在当前特征子集中新增特征 f 后的特征稳定性度量:

$$S(f,F)=\frac{1}{\omega}\sum_{i=1}^{\omega}S(F_f,F_i) \tag{5}$$

$$S(F_f,F_i)=\frac{\lvert F_f\cap F_i\rvert}{\lvert F_f\cup F_i\rvert} \tag{6}$$

式中 ω 为已选特征子集数,F_f 为当前子集 F 新增特征 f 后的特征子集,F_i 为维数相同的已选特征子集,$S(F_f,F_i)$ 为两个特征子集的相似性度量[21],$F_f\cap F_i$ 表示两个特征子集交集的大小,$F_f\cup F_i$ 为其并集的大小。从式(5)可以看出,当前构造的特征子集 F_f 与已选子集相似性愈大,所选的特征愈稳定。

在特征选择过程中,总是希望能够选出这样的排序结果:能够满足特征对类别的相关性最大,而特征间的相关性(冗余度)最小。mRMR 方法就依据这样的准则进行特征选择。而我们希望在此基础上增加稳定性准则,能够评价在特定维数下不同特征对稳定性的影响。因此,根据我们期望所选的特征满足最大相关性、最大稳定性与最小冗余度原则,可以构造如下的特征选择准则:

$$\Phi(D,R,S)=\max(D-R+k\times S) \tag{7}$$

D 为特征与类别相关性度量,R 为特征之间的冗余性度量,S 为特征的稳定性度量,k 为稳定性度量比例系数。

2.2 mRSmR 算法流程

在实际应用中,若用穷举法来获得满足最大相关性、稳定性和最小冗余度的特征子集,则背离了特

征选择研究的初衷。根据 mRSmR 算法求解最大相关、最大稳定与最小冗余特征组的方法,即假设已有子集为当前维数下的最优子集,将求解最优子集转换为求解新增的最优特征,使新增特征满足对类别最大相关、与已有子集最小冗余以及最大稳定的特性。在用 mRSmR 进行选择的过程中,将样本按照一定的比例生成随机训练集,并通过迭代的方式逐步加入特征的稳定性 S, 以此减少样本数据变化对选择结果的扰动。在每一次迭代过程中,根据式(5)计算特征的稳定性 S 后,根据 $\max(D - R + k \times S)$ 的原则选择当前特征子集 F 下新增的最优特征。最后,根据式(5)和(6)计算当前迭代结果与已选特征子集的相似性 S_i, 如果 $|S_i - S_{i-1}| \leqslant \varepsilon$, 则迭代收敛,并输出最新迭代结果。训练集的重复率对选择算法的性能有一定的影响[26],实验表明,当训练集比例为0.8左右时选择算法性能最优。表 2 为 mRSmR 算法的详细算法流程。

表 2 mRSmR 算法

	mRSmR 算法
1	$i = 0,\ S_0 = 0,\ S_1 = 0$
2	**do**
3	$i = i + 1$
4	按比例生成随机训练集 set_i.
5	**for** $j = 1, \cdots, \dim_f$ **do**
6	$D(f, C) = mi(f, C)$
7	$R(f, F) = \frac{1}{\lvert F \rvert} \sum_{f_i \in F} mi(f, f_i)$
8	**if** $i = 1$
9	$S(f, F) = 0$
10	**else**
11	$S(f, F) = \frac{1}{i-1} \sum_{l=1}^{i-1} S(F_f, F_l)$
12	$S(F_f, F_l) = \frac{\lvert F_f \cap F_l \rvert}{\lvert F_f \cup F_i \rvert}$
13	**end if**
14	$f_j \leftarrow \max(D - R + k * S)$
15	**end for**
16	**if** $i \geqslant 2$
17	$S_i(F) = \frac{1}{i-1} \sum_{l=1}^{i-1} S(F, F_l)$
18	**end if**
19	**while** $\lvert S_i - S_{i-1} \rvert > \varepsilon$ **or** $i = 1$
20	输出最新迭代结果

为了加快迭代的收敛,引入了一个与迭代次数 $iter$ 相关的系数 $k \propto iter$, 使得其随迭代次数的增大而增大,进而稳定性因子在式(7)中的比例随迭代而提高。当达到一定的迭代次数时,在式(7)中,随迭代而增大的 k 使得 $k \times S$ 远远大于 $D - R$, 等同于按照 $\max(k \times S)$ 选择特征,此时的 $|S_i - S_{i-1}| \to 0$, 迭代将会收敛。

3 实验分析

为了分析评价本文提出的基于 mRSmR 的特征选择方法的性能,以 21 类遥感数据集[27]提取的不同纹理与颜色特征集为例进行了实验分析与讨论,该数据集是 Yang 等[27]从美国地质调查局国家卫星图像中截取的分辨率为 1 英尺的标注数据,共 21 种类别的遥感卫星数据,每个类别的图像数据为 100 幅。实验提取的纹理特征集主要有 GLCM、Fourier 纹理、Tamura 纹理、GMRF 纹理、LBP 和 Gabor 等,颜色特征集主要有 RGB 直方图、TRGB 直方图和颜色矩等,特征维数总计达 1488 维。将 21 类数据按照 0.5 比例随机分成训练和测试两部分,分别用 4 种特征选择方法对训练样本集进行特征选择,重复 10 次得出稳定性曲线,并用 RBF-SVM[28]进行分类实验,利用 5 折交叉检验确定 SVM 各参数值,得出随维数变化的分类精度曲线。首先分析评价 mRSmR 算法与 mRMR 算法[11]、ReliefF 算法[10]、集成特征选择[22](EnsembleFS)算法等不同特征选择算法的分类精度。然后再选择不同算法下满足一定分类精度的特征子集,并计算其加权一致性(CW)[18]、平均 Tanimoro 指数(ATI)[19,20]和相对汉明距离[17]三个稳定性度量指标,对不同选择算法的稳定性进行评价。

其中,ReliefF 算法抽样次数为 500,最近邻值为 30。集成特征选择算法为基于数据扰动进行选择结果的集成,具体实现为采用多个 mRMR 算法特征选择器对生成的样本数据集特征选择并对多个结果进行集成,最终输出一个集成的选择结果。

3.1 分类精度对比

图 1 给出了 mRSmR、mRMR、ReliefF 和 EnsembleFS 等四种特征选择算法的分类精度曲线。由于特征维数较高,仅取前 500 维分类精度变化曲线。由图可知,本文提出的 mRSmR(最大相关、最大稳定性及最小冗余度)算法能够保持较好的分类性能,与 mRMR(最大相关、最小冗余度)算法、EnsembleFS(集成特征选择)算法相当。

图2给出上述四种算法的局部分类精度曲线。

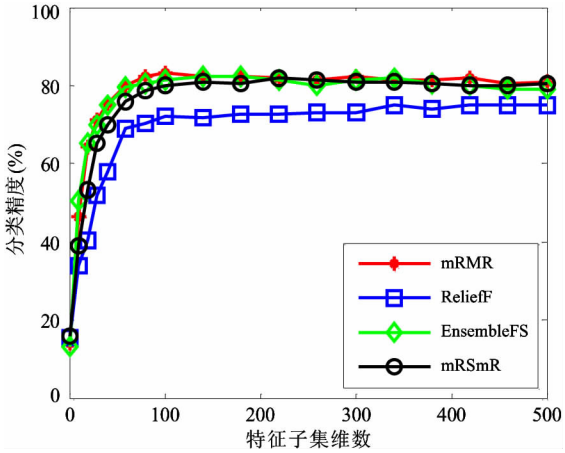

图 1　四种特征选择分类精度曲线

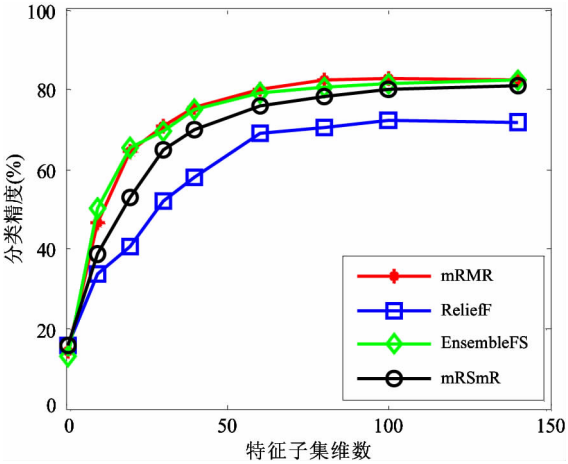

图 2　四种选择方法的分类精度曲线(局部)

从中可以很清晰地看到,分类精度曲线在特征子集为 100 维时已达到最高并趋于稳定。在对特征最优子集的分类精度进行定量分析后,mRSmR 算法的选择结果在分类精度上高于 ReliefF 算法,低于 mRMR 算法与 EnsembleFS 算法 3% 左右。mRSmR 方法虽在选择过程中加入了稳定性因子与相关性、冗余度一起参与特征的选择,且稳定因子在选择准则中的比例随迭代次数的增多不断提高,但它仍能够保持较高的分类性能,并且能够在维数较少时达到较高的分类精度,说明该方法在选择最优特征子集方面是有效的。

3.2　稳定性对比

通过上面的分类实验可知,mRSmR 算法虽然具有较好的分类性能,但如果单纯地按照分类精度来评判特征选择方法的性能,并不能体现 mRSmR 算法的优势。正如 1.2 节所提到的,较高的分类精度固然体现出选择算法对特征信息的提取效果,但保持选择结果的较高稳定性也是很有必要的。为验证 mRSmR 算法是否在维持较高分类精度的的同时具有较高的稳定性,我们利用加权一致性(CW)、平均 Tanimoto 指数(ATI)和汉明距离三个特征子集稳定性度量准则来分析评价上述四种特征选择方法的稳定性。为了能够更好地对比选择算法的分类性能与稳定性能,稳定性对比实验所用到的选择结果与分类实验所用的相同。特征组的最优特征子集由排名靠前的小于或等于 100 维特征组成,具体可根据实际对分类精度与特征维数的不同需要来选择。因为特征选择的目的是选择精度达到一定要求的低维特征子集以达到降维的目的,而此处一定的分类精度往往是低于最高分类精度的。相应地,关注选择结果的稳定性也转为关注选择最优子集的稳定性上。因此我们选取排序结果的 1 ~ 100 维来观察其稳定性曲线的变化。

通过图 3 ~ 图 5 可以看出,本文提出的 mRSmR

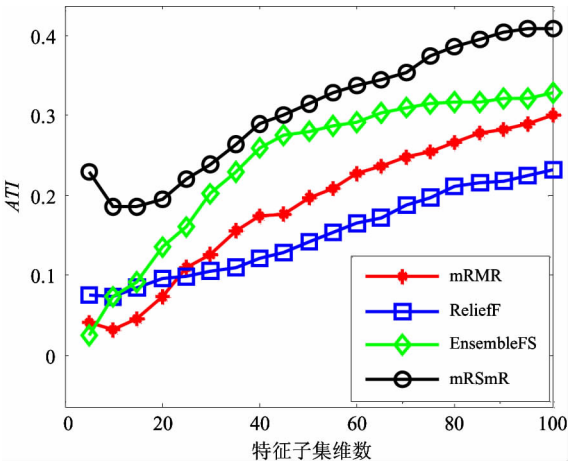

图 3　四种选择方法的 ATI 曲线

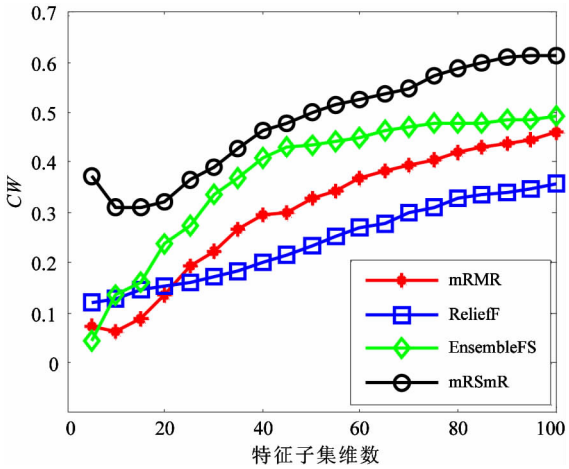

图 4　四种选择方法的 CW 曲线

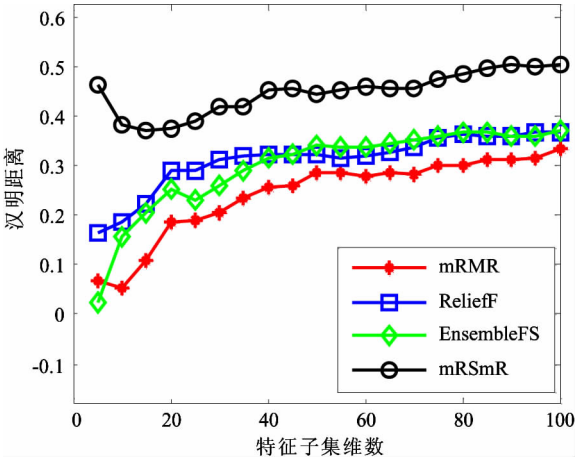

图 5 四种选择方法的汉明距离曲线

算法能够比较明显地提升选择结果的稳定性,相比较其他三种选择方法,其在 CW、ATI 和汉明距离上都有明显提升,并且相比较同为 Filter 方法的 mRMR 与 ReliefF 来说,更有近 20% 的提升。说明 mRSmR 算法能够在牺牲较少分类性能的基础上,大幅提高选择结果的稳定性,能够很好地实现兼顾分类精度与稳定性的目标。

4 结 论

本文从特征选择过程出发,通过产生随机样本集来减少现实中的样本数据扰动,把特征的稳定性、相关性与冗余性一起作为特征选择的评价准则,提出了基于最大相关性、最大稳定性和最小冗余性的过滤式特征选择方法——mRSmR 算法。实验显示,本文提出的 mRSmR 算法能够在保持较高分类性能的基础上,较大地提高选择结果的稳定,实现了兼顾可预测性与稳定性的目标。

参考文献

[1] Wright D J, Wang S W. The emergence of spatial cyberinfrastructure. *Proceedings of the National Academy of Sciences of the United States of America*, 2011, 108(14): 5488-5491

[2] Hafler D A, Pyne S, Hu X L, et al. Automated high-dimensional flow cytometric data analysis. *Proceedings of the National Academy of Sciences of the United States of America*, 2009, 106(21): 8519-8524

[3] Mjolsnes E, DeCoste D. Machine learning for science: State of the art and future prospects. *Science*, 2001, 293(5537): 2051-2055

[4] Mjolsness E, DeCoste D. Machine learning for science: state of the art and future prospects. *Science*, 2001, 293(5537): 2051-2055

[5] 李超,李文法,段洣毅. 用于网络入侵检测的 VFSA-C4.5 特征选择算法. 高技术通讯, 2011, 21(12): 1240- 1245

[6] John G H, Kohavi R, Pfleger K. Irrelevant Features and the Subset Selection Problem. *ICML*, 1994, 94: 121-129

[7] Liu H, Yu L. Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on Knowledge and Data Engineering*, 2005, 17(4): 491-502

[8] Blum A L, Langley P. Selection of relevant features and examples in machine learning. *Artificial intelligence*, 1997, 97(1): 245-271

[9] Kira K, Rendell L A. A practical approach to feature selection. In: Proceedings of the 9th international workshop on Machine learning, Aberdeen, Scotland, UK, 1992

[10] Kononenko I. Estimating Attributes: Analysis and Extensions of RELIEF. Machine Learning: ECML-94. Springer Berlin Heidelberg, 1994

[11] Peng H, Long F, Ding C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2005, 27(8): 1226-1238

[12] Kalousis A, Prados J, Hilario M. Stability of feature selection algorithms: a study on high-dimensional spaces. *Knowledge and information systems*, 2007, 12(1): 95-116

[13] Abeel T, Helleputte T, Van de Peer Y, et al. Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. *Bioinformatics*, 2010, 26(3): 392-398

[14] 李云. 稳定的特征选择研究. 微型机与应用, 2012, 31(15): 1-2

[15] Kim S Y. Effects of sample size on robustness and prediction accuracy of a prognostic gene signature. *BMC bioinformatics*, 2009, 10(1): 147

[16] Ein-Dor L, Zuk O, Domany E. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proceedings of the National Academy of Sciences*, 2006, 103(15): 5923-5928

[17] He Z, Yu W. Stable feature selection for biomarker discovery. *Computational biology and chemistry*, 2010, 34(4): 215-225

[18] Dunne K, Cunningham P, Azuaje F. Solutions to instability problems with sequential wrapper-based approaches to feature selection. Department of Computer Science,

Trinity College, Dublin, Ireland, Technical Report TCD-CD-2002-28, 2002

[19] Kuncheva L I. A stability index for feature selection. In: Proceedings of the 25th International Multi-Conference on Artificial intelligence and applications, Anaheim, USA, 2007. 309-395

[20] Somol P, Novovicova J. Evaluating stability and comparing output of feature selectors that optimize feature subset cardinality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010, 32(11): 1921-1939

[21] Kalousis A, Prados J, Hilario M. Stability of feature selection algorithms. In: Proceedings of the 5th IEEE International Conference on Data Mining, Houston, Texas, USA, 2005. 218-225

[22] Saeys Y, Abeel T, Van de Peer Y. Robust feature selection using ensemble feature selection techniques. In: Proceedings of the Machine Learning and Knowledge Discovery in Databases, Heidelberg, Germany: Springer Berlin, 2008. 313-325

[23] Helleputte T, Dupont P. Partially supervised feature selection with regularized linear models. In: Proceedings of the 26th Annual International Conference on Machine Learning, Montreal, Canada, 2009. 409-416

[24] Loscalzo S, Yu L, Ding C. Consensus group stable feature selection. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France, 2009. 567-576

[25] Vapnik V N, Vapnik V. Statistical Learning Theory. New York: Wiley, 1998

[26] Haury A C, Gestraud P, Vert J P. The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. *PloS one*, 2011, 6(12): e28210

[27] Yang Y, Newsam S. Bag-of-visual-words and spatial extensions for land-use classification. In: Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, San Jose, USA, 2010. 270-279

[28] Chang C C, Lin C J. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent systems and Technology*, 2011,2(3):27:1-27:27

A feature selection algorithm taking account of stability

Ji Jinsheng, Guo Yiyou, Huo Hong, Fang Tao

(* Department of Automation, Shanghai Jiao Tong University, Shanghai 200240)

(** Key Laboratory of System Control and Information Processing, Ministry of Education, Shanghai 200240)

Abstract

To improve the stability of feature selection and reduce the fluctuations caused by the variation of sample data, a filter-type feature selection method considering the stability index is proposed. Unlike the integrated feature selection and other methods, the propose method takes feature's stability, together with the relevance and redundancy, as the evaluation criteria for feature selection, reduces the fluctuations of sample, data by producing multiple data sets, continuously puts new selection results into the iterative calculation of stability, and increases the proportion of the stability factor until the iteration is converged. At last, the achieved feature sequence fusing multi-iteration information is taken as the final result of feature selection. The experimental results show that the proposed method can improve the stability of feature selection obviously, and reach the satisfied classification accuracy meanwhile.

Key words: feature selection, relevance, redundancy, stability, high-dimension data