第44卷第6期 南京航空航天大学学报 Vol. 44 No. 6
2012年12月 Journal of Nanjing University of Aeronautics & Astronautics Dec. 2012

模糊聚类在特征选取中的应用

刘全金[1,2] 赵志敏[1] 李颖新[3]

(1.南京航空航天大学理学院,南京,210016;2.安庆师范学院物理与电气工程学院,安庆,246011;
3.北京经纬纺机新技术有限公司机器视觉与智能研究所,北京,100176)

摘要:提出了一种基于模糊聚类算法的高维特征选取方法。首先,利用Bhattacharyya距离过滤样本类别无关的特征;然后,基于递归特征剔除过程,提出了基于模糊迭代自组织数据分析技术(Interactive self-organizing data analysis technique,ISODATA)聚类方法,以样本与聚类中心的加权距离作为可分性指标,产生候选特征子集;最后,以候选特征子集分类和聚类的接受者操作特征曲线下面积(Area under the receiver operating characteristic curve,AUC)值和正确率作为目标函数,确定最佳特征子集。将该方法用于选取5个基因表达谱数据集的特征基因,结果显示该方法所选特征具有较好的分类和聚类能力,说明了提出的特征选取方法的有效性。

关键词:特征选取;模糊迭代自组织数据分析技术;层次聚类;支持向量机;K近邻

中图分类号:TP391;Q812 **文献标识码**:A **文章编号**:1005-2615(2012)06-0881-07

Application of Fuzzy Clustering Algorithm on Feature Selection

Liu Quanjin[1,2], *Zhao Zhimin*[1], *Li Yingxin*[3]

(1. College of Science, Nanjing University of Aeronautics & Astronautics, Nanjing, 210016, China;
2. School of Physics and Electronic, AnQing Normal University, Anqing, 246011, China;
3. Institute of Machine Vision and Machine Intelligence,
Beijing Jingwei Textile Machinery New Technology Co., Ltd., Beijing, 100176, China)

Abstract: A new feature selection method based on clustering algorithm is proposed to selecte informative features. First, category-unrelated features are kicked out according to Bhattacharyya distance. Then, based on the process of recursive feature elimination, a weighted distance between sample and the cluster center generated by the fuzzy interactive self-organizing data algorithm (ISODATA) is used as the index of feature for separating different classes. Finally, the candidate feature subset with the maximum area under the receiver operating characteristic curve (AUC) value and accuracy rate both in classification and clustering tests is selected as the optimal feature subset. The proposed feature subset selection method is applied to five gene expression profile datasets and experiment results show that the selected features have good performance in terms of both classification and clustering measurements. Results demonstrate that the proposed method is effective for selecting informative features from high-dimensional dataset.

Key words: feature selection; fuzzy iteractive self-organizing data analysis technique (ISODATA); hierachical clustering; support vector machine (SVM); K-nearest neighbor

---

基金项目:国家自然科学基金(10172043,61173068)资助项目;教育部博士点基金(20093218110024)资助项目;江苏省国际合作(BZ2010060)资助项目;江苏省技术监督局重点(KJ122714)资助项目;安徽省教育厅自然科研重点(KJ2010A226)资助项目。

收稿日期:2012-03-13;**修订日期**:2012-05-16

通讯作者:赵志敏,女,教授,博士生导师,E-mail:nuaazhzhm@126.com。

从高维数据中选出与类别相关的特征是机器学习和模式分类的重要一步，特征选取方法的优劣将影响到分类和聚类结果[1-2]，选取的特征应该同时具有较强的分类和聚类能力。Filter 方法和 Wrapper 方法是两种常用的高维数据集的特征选取手段[3]。Filter 方法利用可分性指标评定特征的重要性，选取有效的类别特征，但这种可分性指标仅从单个特征出发，没有考虑特征间的相互关系，所以选取的特征并非最优[4-5]。Wrapper 方法则围绕学习算法，根据该算法执行情况选取相关的特征，这种方法能够取得比 Filter 方法更优的结果[6-8]，但 Wrapper 方法的计算量大，需要较大的时间开销。兼顾Filter 方法的计算复杂度小和 Wrapper 方法的准确度高的优点，结合两种方法选取样本数据的特征，已成为高维数据集特征选取研究的重点[1,9-10]。Cai 等人基于有条件的互信息方法分析了 3 个高维数据集，提出了一种特征选取标准并以此进行特征选取[11]。Golub 等人用“信噪比”衡量特征对样本分类贡献大小，结合支持向量机(Support vector machine，SVM)分类模型，分析急性白血病基因表达谱数据，从 7 129 个基因中选出了 50 个急性髓性白血病(Acute myeloid leukemia，AML)和急性淋巴胞白血病(Acute lymphoblastic leukemia，ALL)亚型特征基因[12]。他们的实验结果表明 Filter 和 Wrapper 方法结合的有效性。

还有不少文献研究如何融合分类算法和聚类算法的优势，利用二者的互补性，进行特征选取和样本分类研究。Alon 等人用层次聚类等方法对高维数据进行了分析研究，并取得了良好的分类效果[13]；Guyon 等人对 Alon 选取的特征做了进一步研究，在递归特征剔除过程中基于线性支持向量机(Feature selection based on support vector machine during procedure of recursive feature eleminate，SVM-RFE)进行了特征选取实验[14]。这些研究表明聚类有助于发现数据集的结构特点，有利于提高分类模型学习的性能。文献[15]对聚类和分类的目标函数进行合并和补充，以聚类时类内数据的紧密程度为参数，完成最佳的分类和聚类。

特征选取方法多数是基于分类模型进行的，选取特征的分类能力较强，但聚类能力相对较弱。鉴于此，本文提出了基于模糊迭代自组织数据分析技术(Interactive self-organizing data analysis technique，ISODATA)的特征选取方法(ISODATA-based feature selection，ISODATA-FS)。首先，将数据集随机分成训练集、校验集和独立测试集，在训练集中，过滤Bhattacharyya 距离较小的特征；然后，在特征递归剔除过程中，进行模糊ISODATA 聚类，分析样本特征与各个聚类中心的距离关系，定义隶属度加权的距离关系作为特征可分性指标，以特征可分性指标为标准，生成一组嵌套的候选特征子集；最后，在校验集中测试候选特征子集的分类和聚类能力，用分类和聚类的接受者操作特征曲线下面积(Area under receiver operating characteristic curve，AUC)[16]值和正确率构建目标函数，从候选特征子集中选取最佳候选特征子集，并用独立测试集验证最佳候选特征子集的分类和聚类性能。

为考察本文提出的特征选取方法 ISODATA-FS 的可靠性，在 5 个基因表达谱数据集中分别进行 20 次特征选取实验，每次将原始数据集样本按 4∶1∶1随机分配至训练集、校验集和独立测试集，用训练集和校验集选取特征，用独立测试集测试被选取特征性能。同时，其他 3 种特征选取方法也被用于 5 个数据集的特征选取以比较特征选取方法的性能。实验结果表明ISODATA-FS 所选特征具有较好的分类和聚类能力，较其他方法所选特征包含更多的样本类别信息。这说明利用聚类模型选取的特征兼有较强的分类和聚类能力，也表明 ISODATA-FS 方法能较好地完成高维数据的特征选取任务。

1 基于模糊ISODATA 的特征可分性分析

模糊 ISODATA 算法在聚类过程中不断利用已有的聚类经验，是一种在迭代过程中寻求最优模糊隶属度矩阵的模糊聚类方法。模糊ISODATA 算法结构简单，计算速度快[17-18]。设数据集有 n 个样本，样本 $\boldsymbol{X}_k$ 有 m 个特征 $\{x_{k1},\cdots,x_{kj},\cdots,x_{km}\}$，其中 x_{kj} 为样本第 j 个特征分量，若数据集样本来自 s 个类别，则有 s 个聚类中心，第 i 个聚类中心 $\boldsymbol{V}_i=\{v_{i1},\cdots,v_{ij},\cdots,v_{im}\}$，其中 v_{ij} 为其第 j 个特征分量。各个样本隶属于 s 个类别的隶属度矩阵为：$\boldsymbol{U}_{s\times n}=(u_{ik})_{s\times n}$，$u_{ik}$ 为第 k 个样本属于第 i 类的隶属度，隶属度值在0 至1 之间。样本愈接近某聚类中心，此样本隶属于该类的隶属度愈大。隶属度矩阵 $\boldsymbol{U}_{s\times n}$ 同时满足以下约束条件：$\sum_{i=1}^{s}u_{ik}>0,\forall\ k$；$\sum_{k=1}^{n}u_{ik}>0,\forall\ i$。

基于样本与聚类中心的距离及约束条件，样本 X_k 隶属于第 i 类的隶属度[17-18]为

$$u_{ik}=\frac{1}{\sum_{t=1}^{s}\left(\frac{\|\boldsymbol{X}_k-\boldsymbol{V}_i\|}{\|\boldsymbol{X}_k-\boldsymbol{V}_t\|}\right)^{\frac{1}{r-1}}} \tag{1}$$

第 i 个聚类中心的第 j 个特征分量 v_{ij} 为

$$v_{ij}=\frac{\sum_{k=1}^{n}(u_{ik}^{2}x_{kj})}{\sum_{k=1}^{n}(u_{ik})^{2}} \tag{2}$$

样本的隶属度矩阵和 s 个聚类中心矩阵可以通过用迭代法求解[18]。具体的模糊 ISODATA 算法如下：

初始化隶属度矩阵 $\boldsymbol{U}_{s\times n}^{0}$；

Do

计算聚类中心矩阵 $\boldsymbol{V}_{s\times m}^{*}$；

刷新隶属度矩阵 $\boldsymbol{U}_{s\times n}^{*}$；

$\Delta\boldsymbol{U}=|\boldsymbol{U}_{s\times n}^{0}-\boldsymbol{U}_{s\times n}^{*}|_{\max}$；

$\boldsymbol{U}_{s\times n}^{0}=\boldsymbol{U}_{s\times n}^{*}$；

Until $\Delta\boldsymbol{U}<\varepsilon$；

阈值 ε 为给定的某任意小数，当隶属度值 u_{ik} 与前一次 u_{ik} 的差值绝对值的最大值小于阈值 ε 时，迭代终止。隶属度矩阵 $\boldsymbol{U}_{s\times n}^{0}$ 元素初始值愈接近样本在不同类别中的实际隶属度，迭代收敛愈快，聚类也愈精确[18]。

从样本在特征空间中的位置看，同类样本聚集越紧密，样本离本类聚类中心越近，距异类聚类中心越远，样本越易被识别；反之，样本分布就比较分散，不易被识别。从这个角度出发，特征选取就是为了选取令样本离本类聚类中心距离近，距异类聚类中心距离远的特征。可知，样本特征与本类聚类中心对应的特征分量差值越小，与异类聚类中心对应的特征差值越大，该特征对样本可分性贡献就越大。

由式(1)知，隶属度 u_{ik} 以样本离聚类中心的欧氏距离作为主要参数，隶属度隐含了样本特征与样本类别的关系，样本特征也影响着样本属于某类别的隶属度值。鉴于此，本文在考量样本特征在类内和类间距离关系的同时，以样本的隶属度为权重量化特征对样本可分性的贡献。式(3)定义了训练集样本第 p 个特征对样本可分性贡献值

$$\mathrm{Index}(p)=\sum_{X_k\in\text{Training set}}\left(\prod_{j\neq i}\left(\frac{\|x_{kp}-v_{jp}\|}{\|x_{kp}-v_{ip}\|}\cdot\frac{u_{ik}}{u_{jk}}\right)\right) \tag{3}$$

式中：j 表示不同于 i 的聚类号。易知，样本的本类隶属度 μ_{ik} 越大，异类隶属度 μ_{jk} 越小，权重就越大，反之权重就越小。因为隶属度体现了样本与各聚类中心的相对关系，所以用隶属度加权能较为科学地评价模糊 ISODATA 聚类中样本的特征对样本可分性的影响。

通过隶属度加权样本特征在类内和类间的距离关系以衡量特征对样本可分性的贡献，该贡献值亦可被视为特征影响模糊 ISODATA 聚类能力的重要性指标。因为模糊ISODATA 算法的“聚类”是在无监督情况下得到的，能较客观地反映样本数据的内在结构，所以特征的贡献值越高，其所含的样本类别信息会越丰富。

2　候选特征子集生成及选取

样本的类别是样本主要特征相互作用的结果，并非各样本特征单个作用结果叠加[4-5]。基于模糊 ISODATA 算法研究特征对样本类别可分性的贡献，其过程就是先从特征集合角度考察特征对样本聚类影响，依据特征的“贡献值”度量特征在样本分类和聚类中发挥的作用大小。将特征对样本聚类的影响看成其对样本类别可分性的贡献指标，从原有特征集合中去除那些对样本类别影响小的特征，保留对类别贡献大的特征，形成维数较低的候选特征集合。但仅凭一次模糊聚类得到的特征对样本类别可分性的贡献来选取特征仍属于 Filter 方法的范畴，无法满足特征选取的需要[14]。

本文采用反向递归特征剔除(Recursive feature elimination, RFE)方法产生候选特征子集[14]，并将候选特征子集在训练集中的分类和聚类实验的AUC 值与正确率作为递归特征剔除过程的目标函数。分类实验选用SVM 和 K 近邻分类器，聚类实验选用层次聚类算法。目标函数反映了各候选特征子集的分类和聚类能力，最高目标函数对应的候选特征子集即为最佳候选特征子集，其特征将作为数据集样本的类别特征。首先，在训练集进行模糊 ISODATA 聚类，根据式(3)计算特征对样本可分性的贡献并排序，保留贡献值较高的90% 特征组成候选特征子集；然后，基于该候选特征子集，重新进行模糊 ISODATA 聚类，并刷新特征对样本类别可分性的贡献值，形成维数更低的新的候选特征子集；如此往复，得到一组嵌套的候选特征子集。

数据集各类别样本数量不均匀使得仅以分类和聚类正确率评价被选特征子集的分类和聚类能力不够充分，为可靠估计候选特征子集所含样本类别信息，本文还采用分类和聚类的 AUC 值[16]评价特征子集的分类和聚类能力。也就是在递归特征剔除过程中同时用分类和聚类实验的AUC 值和正确率构建目标函数，以确定分类和聚类能力最强的最

佳特征子集。最佳特征子集的分类和聚类能力在独立测试集中得到进一步验证，AUC 值和正确率越高，最佳特征子集的分类能力就越强。

递归特征剔除过程中，用训练集样本的模糊 ISODATA 聚类结果计算特征可分性指标，以校验集样本的分类和聚类结果构建目标函数选取最佳特征子集，再用独立测试集样本测试最佳特征子集的分类和聚类能力。目标函数避开特征分析模型模糊 ISODATA 算法有利于避免特征选取方法过度拟合于模糊 ISODATA 算法；独立测试集中最佳特征子集的分类和聚类结果可反映被选特征对训练集和校验集样本的依赖程度，体现特征选取方法的鲁棒性。

3 特征选取实验

为不失一般性，本文研究了两类样本数据的特征选取问题，在 5 个基因表达谱数据集中进行特征选取实验。基因表达谱数据集结构如表 1 所示，为避免特征选取时样本分配偏置，分别对 5 个基因表达谱数据集中进行 20 次特征选取实验，每次将数据集样本按 4∶1∶1 随机分配组成训练集、校验集和独立测试集，用训练集和校验集选取特征，用独立测试集中测试被选取特征性能。

表 1 基因表达谱数据集数据结构

数据集	基因数	样本数（正类/负类）	训练集/校验集/独立测试集	特征选取范围	参考文献
Acute Leukemia	7 129	72(47/25)	48/12/12	100	[12]
Multiple Myeloma	7 129	105(74/31)	70/18/17	100	[19]
Colon	2 000	62(40/22)	42/10/10	500	[13]
DLBCL	7 129	77(58/19)	52/13/12	1 000	[20]
Prostate	12 600	102(52/50)	68/17/17	1 000	[21]

其他3 种特征选取方法也被用于这5 个数据集进行特征选取。T-test 方法基于异类样本特征的 T-test 值衡量特征的重要性[22]，Relief 方法利用异类样本特征间的相关性来评价特征重要性[23]；SVM-RFE 方法在递归特征选取过程中基于线性支持向量机模型分析特征重要性[14]并进行特征选取。前两种方法属于 Filter 特征选取方法，SVM-RFE 和本文提出的ISODATA-FS 方法属于 Wrapper 特征选取方法。

3.1 噪声基因滤除

5 个数据集的特征维数远高于样本数，数据集中只有小部分特征与样本类别有关，其余特征数据为与样本类别无关的“噪声”数据，所以有必要先降低特征维数再进行特征选取。ISODATA-FS 算法利用 Bhattacharyya 距离过滤“噪声”特征。Bhattacharyya 距离体现了特征在异类样本中的均值的差异，也体现了异类样本间特征方差的不同[4,24]。Bhattacharyya 距离公式如下

$$B(f)=\frac{(\mu_+(f)-\mu_-(f))^2}{4(\sigma_+^2(f)+\sigma_-^2(f))}+\frac{1}{2}\ln\left(\frac{\sigma_+^2(f)+\sigma_-^2(f)}{2\sigma_+(f)\sigma_-(f)}\right) \tag{4}$$

式中：μ_+、μ_-分别为特征 f 在两类不同样本中的均值，σ_+、σ_-为相应的标准差。特征的 Bhattacharyya 距离越大，其在异类样本中的差异就越大，包含的样本类别信息就越丰富。

利用 Bhattacharyya 距离过滤 5 个数据集的噪声特征。通过测试，保留表 1 特征选取范围所示的特征维数进行下一步特征选取。每次数据集样本随机分配后，都根据训练集样本异类特征间的 Bhattacharyya 距离重新过滤噪声特征。

3.2 特征选取

特征选取实验中，设置模糊 ISODATA 算法的聚类中心个数 s 为 2、算法迭代终止阈值 ε 为 0.000 1。分类模型 SVM 选用线性核函数，惩罚参数 C 设置为 400；K 近邻分类算法取 K 为 5，即 5 近邻分类。层次聚类算法中样本间的距离以欧氏距离度量，样本与小类、小类与小类之间的亲疏程度用“平均值”来度量。

ISODATA-FS 方法分别对过滤后的 Acute Leukemia 等 5 个数据集进行特征选取实验。在递归特征选取过程中，基于模糊 ISODATA 聚类根据式(3)计算特征对样本类别可分性的贡献，并以据该指标依次剔除 10%的特征生成一组候选特征子集；然后基于候选特征子集用训练集样本训练 SVM 和 K 近邻分类器，识别校验集样本类别，基于候选特征子集对校验集样本做层次聚类实验；再根据分类和聚类实验的AUC 值和正确率确定最佳特征子集；最后用独立测试集测试被选最佳特征子集的分类和聚类能力。

表 2 记录了 ISODATA-FS 方法在 5 个数据集中的 20 次特征选取结果。“特征子集维数”为 20 次选取的最佳特征子集维数的平均值(mean)；AUC

表 2　最佳特征子集的分类和聚类结果比较

数据集	特征选取方法	特征子集维数	AUC			错误数		
			SVM	K 近邻	聚类	SVM	K 近邻	聚类
		mean	mean±std	mean±std	mean±std	mean±std	mean±std	mean±std
Acute Leukemia	T-test	8.40	0.89±0.09	0.87±0.10	0.76±0.15	1.90±1.41	2.30±1.53	3.85±1.63
	Relief	6.75	0.92±0.09	0.88±0.10	0.80±0.13	1.30±0.98	1.75±1.07	2.85±1.60
	SVM-RFE	10.00	0.96±0.04	0.95±0.07	0.78±0.12	0.80±0.89	1.45±0.86	2.95±1.73
	ISODATA-FS	6.20	0.94±0.07	0.96±0.04	0.82±0.13	1.05±0.89	1.30±0.98	2.70±1.69
Multiple Myeloma	T-test	10.85	0.97±0.04	0.97±0.05	0.92±0.13	0.85±0.99	0.75±0.91	1.65±1.79
	Relief	3.30	0.96±0.03	0.98±0.03	0.97±0.04	0.15±0.37	0.25±0.55	0.45±0.89
	SVM-RFE	6.00	0.96±0.09	0.98±0.03	0.80±0.16	1.00±1.69	0.80±1.01	1.75±2.79
	ISODATA-FS	3.10	0.99±0.02	0.99±0.02	0.98±0.04	0.15±0.49	0.20±0.41	0.45±0.86
Colon	T-test	15.25	0.80±0.12	0.76±0.07	0.81±0.13	2.25±1.52	2.85±1.07	2.50±1.55
	Relief	21.50	0.81±0.13	0.82±0.13	0.84±0.12	2.25±1.35	2.20±1.38	2.20±1.74
	SVM-RFE	21.70	0.83±0.12	0.83±0.14	0.74±0.12	2.05±1.28	2.05±1.57	3.70±1.13
	ISODATA-FS	22.60	0.84±0.13	0.88±0.12	0.89±0.12	2.10±1.39	1.75±1.27	2.00±1.29
DLBCL	T-test	23.65	0.89±0.11	0.86±0.13	0.69±0.11	2.10±1.26	2.10±1.37	3.90±1.36
	Relief	25.90	0.93±0.10	0.92±0.12	0.68±0.15	1.65±1.23	1.75±1.62	3.80±1.79
	SVM-RFE	25.60	0.91±0.09	0.94±0.07	0.72±0.12	1.40±1.14	1.4±1.35	3.05±1.29
	ISODATA-FS	26.45	0.93±0.11	0.96±0.08	0.76±0.14	1.10±1.17	1.25±1.12	3.95±1.04
Prostate	T-test	29.45	0.81±0.08	0.83±0.09	0.68±0.08	3.50±1.93	3.75±2.17	6.90±1.17
	Relief	23.75	0.89±0.09	0.84±0.09	0.73±0.11	3. 35±2.00	4.10±1.83	6.20±1.70
	SVM-RFE	18.60	0.87±0.09	0.87±0.09	0.70±0.11	3. 00±1.50	3.30±1.84	6.45±1.50
	ISODATA-FS	21.80	0.88±0.09	0.89±0.07	0.73±0.11	3.15±1.66	3.15±1.46	6.10±1.77

表示最佳特征子集在独立测试集中分类和聚类测试实验的AUC 值的平均值(mean)和标准差(std)；错误数表示最佳特征子集在独立测试集中测试实验的错分样本数或错聚样本数的平均值(mean)和标准差(std)。

表 3 列出了ISODATA-FS 方法在 5 个数据集选取特征所用的平均时间。特征选取时间与数据集的特征维数有关系，维数高则耗时就多。

表 3　候选特征子集产生时间比较　　s

数据集	T-test	Relief	SVM-RFE	ISODATA-FS
Acute Leukemia	1.053 1	13.302 5	4.786 7	2.257 8
Multiple Myeloma	1.084 3	20.052 4	8.450 1	3.293 9
Colon	0.282 8	3.225 8	5.239 8	2.356 2
DLBCL	1.056 2	14.503 2	8.982 8	4.827 3
Prostate	2.194 5	35.113 3	14.612 6	9.015 8

3.3　特征选取方法比较

表2 也列出了其他3 种特征选取方法在5 个数据集中的特征选取实验结果。噪声特征过滤阶段，T-test 和Relief 方法分别根据特征在异类样本间的 t 统计量和相关性过滤噪声特征，SVM-RFE 方法用 Bhattacharyya 距离过滤噪声特征，过滤后的特征选取范围与表1 所列一样。递归特征选取过程中 3 种特征选取方法根据各自的特征重要性指标生成候选特征子集，同样，分类和聚类能力最佳的候选特征子集被确定为最佳特征子集，其在独立测试集中的分类和聚类结果如表 2 所示。

由 表 2 知，ISODATA-FS 方法选取的特征在分类和聚类能力总体上优于其他 3 种方法选取的特征。其中ISODATA-FS 方法在 5 个数据集中选取的特征的 K 近邻分类结果优于其他 3 个方法选出的特征。SVM-RFE 方法在3 个数据集中选取的特征的SVM 分类正确率高于其他特征选取方法选取的特征；Relief 方法在 2 个数据集中选取的特征的SVM 分类 AUC 值高于其他特征选取方法选取的特征；SVM-RFE 方法在DLBCL 数据集中选取的特征的层次聚类正确率高于其他方法选取的特征。

实验中ISODATA-FS 方法选取的特征有较好的分类和聚类能力。这表明基于模糊聚类的特征可分性指标客观地反映了异类样本的数据结构，也说明用分类和聚类的 AUC 值和正确率构建目标函数，既能保证选取的特征有较强的分类和聚类能力，又可避免单一的分类正确率构成目标函数导致的特征选取偏置。20 次样本随机分配得到的特征

选取实验结果也验证了这种目标函构建方法可有效地回避特征选取的过学习现象的发生,也排除了选取特征对选取过程中所用模型(分类或聚类模型)的依赖。

从4种特征选取方法选取特征消耗的时间来看,T-test方法运行速度最快、算法最简单,但其选取特征的分类和聚类性能结果比其他3个方法差。ISODATA-FS方法选取特征用时比Relief和SVM-RFE方法短,得益于简单快速的模糊ISODATA算法。

4 结束语

本文结合Filter方法和Wrapper方法,提出基于模糊ISODATA算法考察特征对样本类别可分性的贡献,并在递归特征剔除过程中通过目标函数综合测试被选特征的分类和聚类能力。通过与其他3种特征选取方法的特征选取实验比较,ISODATA-FS方法所选特征有较强的分类和聚类能力,能较好地反映异类样本间的差异性,表明了本文提出的高维特征选取方法的可行性和有效性,具有一定的应用价值。

鉴于ISODATA-FS方法的特征可分性指标是基于简单快速的模糊ISODATA算法,特征选取耗时少,速度快。以模糊ISODATA聚类获取的特征可分指标体现了聚类样本的数据结构,同时代表候选特征子集的分类和聚类性能力的目标函数保证了被选特征拥有丰富的类别信息,分类和聚类算法的综合应用也使得特征选取方法具有较强的鲁棒性。

参考文献:

[1] Kudo M, Sklansky J. Comparison of algorithms that select features for pattern classifiers [J]. Pattern Recognition, 2000, 33 (1): 25-41.

[2] Liu H, Motoda H, Yu Lei. A selective sampling approach to active feature selection [J]. Artificial Intelligence, 2004, 159(1/2):49-74.

[3] Hua J P, Tembe W D, Dougherty E R. Performance of feature-selection methods in the classification of high-dimension data [J]. Pattern Recognition, 2009, 42(3):409-424.

[4] Duda R O, Hart P E, Stork D G. Pattern classification (Second ed.) [M]. New York: John Wiley & Sons, 2001.

[5] Uncu Ö, Türksenb I B. A novel feature selection approach: Combining feature wrappers and filters [J]. Information Sciences, 2007,177(2):449-466.

[6] John G, Kohavi R, Pfleger K. Irrelevant features and the subset selection problem [C]// Machine Learning: Proceedings of the Eleventh International. San Francisco, CA: Morgan Kaufmenn Publisher, 1994:121-129.

[7] 刘全金,李颖新 阮晓钢. 基于BP网络灵敏度分析的肿瘤亚型分类特征基因选取[J]. 中国生物医学工程学报, 2008,27(5):710-715.
Liu Quanjin, Li Yinxin, Ruan Xiaogang. Informative gene selection for cancer subtype classification with BP neural networks [J]. Chinese Journal of Biomedical Engineering, 2008,27(5):710-715.

[8] Kohavi R, John G H. Wrappers for feature subset selection [J]. Artificial Intelligence, 1997, 97(1/2):273-324.

[9] Das S. Filters, wrappers and a boosting-based hybrid for feature selection [C]// Machine Learning: Proceedings of the Eighteenth International Conference. Willianstown, MA: [s.n.]. 2001:74-81.

[10] Hong Y, Kwong S, Chang Yuchou, et al. Unsupervised feature selection using clustering ensembles and population based incremental learning algorithm [J]. Pattern Recognition, 2008, 41(9):2742-2756.

[11] Cai R, Hao Z, Yang X, et al. An efficient gene selection algorithm based on mutual information [J]. Neurocomputing, 2008, 72(4/6):991-999.

[12] Golub T R, Slonim D K, Tamayo P, et al. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring [J]. Science, 1999, 286(5439):531-537.

[13] Alon U, Barkai N, Notterman D A, et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays [J]. Proceedings of the National Academy of Sciences of the United States of American, 1999, 96(12):6745-6750.

[14] Guyon I, Weston J, Barnhill S, et al. Gene selection for cancer classification using support vector machines[J]. Machine Learning, 2002, 46(1/3):389-422.

[15] Cai W L, Chen S C, Zhang D Q. A simultaneous

learning framework for clustering and classification [J]. Pattern Recognition, 2009, 42(7):1248-1259.

[16] Provost F, Fawcett T. Analysis and visualization of classifier performance: comparison under imprecise class and cost distributions[C]// Knowledge Discovery and Data Mining: Proceeding of the Third International Conference. Menlo Park, CA: AAAI Press, 1997:43-48.

[17] Bezdek J C. Physical Interpretation of Fuzzy ISODATA[J]. IEEE Trans, Systems Man, Cybern, 1986, SME-6(2):32-37.

[18] 朱剑英. 智能系统非经典数学方法[M]. 武汉：华中科技大学出版社, 2001:141-145.
Zhu Jianyin. Non-classical mathematics method for Intelligent systems [M]. Wuhan: Huazhong University of Science and Technology Publishing House, 2001: 141-145.

[19] Zhan F, Hardin J, Kordsmeier B, et al. Global gene expression profiling of multiple myeloma, monoclonal gammopathy of undetermined significance, and normal bone marrow plasma cells[J]. Blood, 2002, 99(5):1745-1757.

[20] Shipp M A, Ross K N, Tamayo P, et al. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning [J]. Nat Med, 2002,8(1):68-74.

[21] Singh D, Febbo PG, Ross K, et al. Gene expression correlates of clinical prostate cancer behavior[J]. Cancer Cell, 2002,1(2):203-209.

[22] Theodoridis S, Koutroumbas K. Pattern Recognition[M]. New York: Academic Press, 1999.

[23] Kononenko I. Estimating attributes: analysis and extensions of RELIEF[C]// Proceedings of the 7th European Conference on Machine Learning. Catania, Italy:[s.n.], 1994:171-182.

[24] Gunala S, Edizkan R. Subspace based feature selection for pattern recognition [J]. Information Sciences, 2008, 178 (19): 3716-3726.