DOI:10.13718/j.cnki.xsxb.2022.03.003

决策依赖聚类的高维数据特征选择①

邓廷权, 辛丽颖

哈尔滨工程大学 数学科学学院, 哈尔滨 150001

摘要: 针对启发式特征选择和特征聚类驱动特征选择方法的不足, 研究了决策依赖的特征冗余性问题, 提出了一种基于邻域粗糙集的决策依赖特征聚类的高维数据特征选择方法(RDCFS). 首先, 依据邻域粗糙集模型, 设计了一种特征联合依赖度增益度量, 刻画数据特征在分类和辨识层面上的冗余性和关联性. 其次, 构建了一种最优特征簇结构的评估准则和特征冗余图的最优图割划分. 再次, 给出了一种基于簇信息的特征中心度和特征依存度度量, 指导实现高维数据的特征选择. 在 UCI 数据库中选取 8 组真实数据集作对比实验, 实验结果表明本文所提特征选择方法能够获得更紧凑的特征子集, 且在分类性能上优于多种现有最新方法.

关 键 词: 高维数据; 特征选择; 决策依赖性; 聚类

中图分类号: TP301.6 **文献标志码**: A **文章编号**: 1000-5471(2022)03-0016-10

Decision Dependence Clustering Based Feature Selection for High Dimensional Data

DENG Tingquan, XIN Liying

College of Mathematical Sciences, Harbin Engineering University, Harbin 150001, China

Abstract: In order to improve the deficiency of heuristic feature selection and feature clustering-driven feature selection, the issue on feature redundancy was investigated and a feature selection method based on neighborhood rough set model and decision dependence feature clustering (RDCFS) was proposed for high dimensional data. Firstly, the gain of feature joint dependence based on the neighborhood rough set model was designed to describe redundancy and correlation between data features at the classification and identification levels. Secondly, an evaluation criterion of optimal feature cluster structure and optimal graph partition of feature redundant graph were constructed. Thirdly, the degrees of feature centrality and feature dependence based on cluster information were presented to guide feature selection for high-dimensional data. Eight real data sets from UCI Repository were selected for comparative experiments, and experimental results show that the proposed feature selection method brought out more compact feature subsets and achieved better classification performance than lots of recently existing methods.

Key words: high-dimensional data; feature selection; decision dependence; clustering

① 收稿日期: 2021-12-21
基金项目: 国家自然科学基金项目(12171115).
作者简介: 邓廷权, 教授, 博士, 主要从事不确定数学理论与方法、粒计算与知识发现、数据挖掘与机器学习等方面的研究.

随着信息时代的高速发展，大数据成为这个时代最具代表性的标志之一. 如何在规模庞大、结构复杂的高维数据中挖掘出最有价值的信息是研究人员和技术工作的热点关注问题. 特征选择是一种有效的高维数据预处理方法，旨在去除数据中的冗余和不相关信息，降低数据维度和学习算法的计算复杂度，是人工智能基础理论中的热点研究主题.

特征选择[1]可分为嵌入式、封装式和过滤式等诸多方法. 嵌入式特征选择[2]将特征选择与学习器融合，在学习器训练过程中自动地进行特征选择. 该方法缺乏特征选择的可解释性. 封装式特征选择[3]通过从初始特征集合中不断地选择特征子集和训练学习器，根据学习器的性能来对子集进行评价，直到选择出最佳的子集. 多次训练学习器导致封装式特征选择方法的计算复杂度很高. 过滤式方法[4]通过设计特征评估度量或相关统计量，通过阈值法或启发式搜索方法选出具有代表性的特征. 常用的特征评估度量有熵、互信息、信息增益、基尼系数等等. 基于粗糙集[5]、邻域粗糙集[6]和模糊邻域粗糙集[7]的特征选择方法是一类重要的启发式特征选择方法，通过设计具有良好性质的特征评价函数，确保选出的特征子集保持甚至超过原始特征的分类能力，因而得以广泛研究.

启发式特征选择方法需要多次遍历所有候选特征，一些与分类无关或对分类不产生影响的冗余特征可能会被反复计算，在数据维度很高的情况下消耗大量的计算资源. 鉴于此，基于特征聚类的特征选择方法应运而生. 该类方法的主旨思想是通过构建描述数据特征间相似性和关联性的度量，采用某种聚类方法对特征聚类，并在各类簇中选择有代表性的特征作为特征选择子集. 文献[8]利用余弦函数刻画特征向量之间的相似性，采用 K -均值对特征聚类，选择每一簇的中心特征构成特征子集. 该方法在特征聚类时仅考虑到特征间的相似性，缺乏特征与决策间的依赖关系，不能确保选出的特征子集具有独立性和最强的辨识能力. 文献[9]基于熵和信息增益概念构建集合间不确定性度量，描述特征间的相关性和特征与决策间的关联性，删除不确定度小于某一阈值的特征，构建以剩余特征为顶点、以特征间的不确定度为边权的最小生成树，通过阈值法获得特征的簇结构，并在每个簇中选择一个与决策关联性最大的特征作为特征选择的特征子集. 该方法既考虑了特征间的相关性，也分析了特征与决策间的关联性，但缺少特征集与决策间的依赖性以及特征集中元素的冗余性等方面的探索.

针对上述现有特征聚类驱动的特征选择存在的缺陷，本文提出一种决策依赖聚类的高维数据特征选择方法. 首先，在邻域粗糙集模型基础上分析了决策关于特征对的依赖度与决策关于单一特征的依赖度间的关系，构建了特征冗余度度量和特征冗余图，依据图割理论获得特征划分子集. 为了获得最优的特征分割，提出了一种簇内特征冗余度最大、簇间特征冗余度最小的聚类簇数评估方法. 通过分析聚类簇中特征关于决策的互信息及特征依存度度量，提出一种中心度和依存度联合的特征子集确定策略，实现高维数据的特征选择.

1　决策依赖特征聚类方法

基于聚类思想的特征选择方法是通过构建数据特征间相似性度量，采用 K -均值[10]或其他聚类方法将特征划分为不相交的子簇. 现有大部分方法在构建特征相似图时只考虑特征间的相似性，忽略数据的标签信息，导致数据特征与其对应决策类之间缺乏必要的关联性.

本节将基于邻域粗糙集理论构建一种决策依赖的特征聚类方法. 在回顾邻域粗糙集及相关概念基础上，提出一种基于决策依赖度的邻域依赖度增益的构造方法，并探讨了其性质. 基于特征间邻域依赖度度量构建特征相似图，采用谱聚类方法获得特征的类簇结构. 为了获得特征的最优聚类簇数，我们结合簇内冗余度和簇间冗余度给出一种最优特征簇的选择方法.

1.1　邻域粗糙集基本概念

设 $DIS=\langle U, A, V, D\rangle$ 为一个决策信息系统，其中 $U=\{x_1, x_2, \cdots, x_n\}$ 为非空论域，$A=\{a_1, a_2, \cdots, a_m\}$ 为条件属性集，也称特征集，V 是对象关于属性的(实数) 值集，D 为决策属性. 对 $\forall x \in U$，$B \subseteq A$，x 由 B 确定的 δ 邻域信息粒[6] 为

$$R_B^{\delta}(x)=\{y \in U \mid d_B(x, y) \leqslant \delta\}$$

其中：$\delta(\delta>0)$ 为邻域参数，$d_B(x, y)=\min\limits_{a_i \in B} \mid v(x, a_i)-v(y, a_i) \mid$ 为对象 $x, y \in U$ 关于属性集 B 的

距离函数，$v(x, a_i)$ 表示 x 在属性 a_i 下的取值.

对任意 $X \subseteq U$，$B \subseteq A$ 和 $\delta > 0$，X 的关于 B 的 δ 下近似和 δ 上近似[6] 分别定义为

$$\underline{R}_B^{\delta}(X) = \{x \in U \mid R_B^{\delta}(x) \subseteq X\}$$

$$\overline{R}_B^{\delta}(X) = \{x \in U \mid R_B^{\delta}(x) \cap X \neq \Phi\}$$

下近似可以理解为由邻域信息粒完全包含在 X 中的对象构成，上近似包含邻域信息粒可能属于 X 的对象全体. 在知识发现中，上近似和下近似被用于对集合 X 做逼近描述.

在决策信息系统 $DIS = \langle U, A, V, D\rangle$ 中，记 U 关于决策属性集 D 的划分为 $U/D = \{D_1, D_2, \cdots, D_M\}$，则对任意 $B \subseteq A$ 和 $\delta > 0$，决策属性集 D 的关于 B 的 δ 上近似、下近似和边界[11] 可分别表示为

$$\overline{R}_B^{\delta}(D) = \bigcup_{k=1}^{M} \overline{R}_B^{\delta}(D_k)$$

$$\underline{R}_B^{\delta}(D) = \bigcup_{k=1}^{M} \underline{R}_B^{\delta}(D_k)$$

$$BN_B^{\delta}(D) = \overline{R}_B^{\delta}(D) - \underline{R}_B^{\delta}(D)$$

决策属性集 D 关于条件属性子集 B 的 δ 正域和依赖度[11] 分别定义为

$$POS_B^{\delta}(D) = \bigcup_{k=1}^{M} \underline{R}_B^{\delta}(D_k)$$

$$\gamma_B^{\delta}(D) = \frac{|POS_B^{\delta}(D)|}{|U|}$$

其中：$|P|$ 表示集合 P 的基数. 显然，D 关于 B 的依赖度 $\gamma_B^{\delta}(D)$ 是一个不大于 1 的非负实数. 如果 $\gamma_B^{\delta}(D) = 1$，则 D 完全依赖于知识 B，也就是说，U 关于属性集 B 的 δ 邻域覆盖是 U 关于决策 D 的划分的子覆盖；如果 $\gamma_B^{\delta}(D) = 0$，则 D 完全独立于 B.

1.2 决策依赖特征关联性度量

依赖度反映的是决策属性与条件属性间的依赖关系. 为了探讨属性对间的关系及其对决策产生的影响，本文构建决策依赖属性(特征) 间的相似度度量. 该度量既描述决策属性与特征对的依赖性，又刻画特征间的关联性和相似性.

定义 1 给定一个决策信息系统 $DIS = \langle U, A, V, D\rangle$，$\delta > 0$，对任意 $a_i, a_j \in A$，简记 $\gamma_{\{a_j\}}^{\delta}(D)$ 为 $\gamma_{a_j}^{\delta}(D)$，称

$$\Gamma_{(a_i, a_j)}^{\delta}(D) = \frac{2\gamma_{(a_i, a_j)}^{\delta}(D) - \gamma_{a_i}^{\delta}(D) - \gamma_{a_j}^{\delta}(D)}{2} \tag{1}$$

为特征 a_i 和 a_j 的平均邻域依赖度增益.

根据定义 1，$\gamma_{(a_i, a_j)}^{\delta}(D)$ 表示论域 U 被 a_i, a_j 粒化时，δ 邻域信息粒被决策 D 完全正确认识的比例；$\gamma_{a_i}^{\delta}(D)$ 和 $\gamma_{a_j}^{\delta}(D)$ 分别表示用 a_i 和 a_j 单独粒化 U 时，对象能够被正确辨识的比例；$\Gamma_{(a_i, a_j)}^{\delta}(D)$ 表示特征 a_i, a_j 同时被用于决策时，相对二者单独被用于决策时，被正确辨识的样本比例差值.

性质 1 $\Gamma_{(a_i, a_j)}^{\delta}(D) = \Gamma_{(a_j, a_i)}^{\delta}(D)$.

证明：显然.

性质 2 $0 \leqslant \Gamma_{(a_i, a_j)}^{\delta}(D) \leqslant 1$.

证明：对任意 $\delta > 0$，$a_i, a_j \in A$，对象集 U 关于 $\{a_i, a_j\}$ 的邻域粒化都是 U 分别关于 a_i 和 a_j 邻域粒化的子覆盖. 这样，$POS_{\{a_i\}}^{\delta}(D) \subseteq POS_{\{a_i, a_j\}}^{\delta}(D)$，$POS_{\{a_j\}}^{\delta}(D) \subseteq POS_{\{a_i, a_j\}}^{\delta}(D)$. 于是，$0 \leqslant \gamma_{a_i}^{\delta}(D) \leqslant \gamma_{(a_i, a_j)}^{\delta}(D) \leqslant 1$，$0 \leqslant \gamma_{a_j}^{\delta}(D) \leqslant \gamma_{(a_i, a_j)}^{\delta}(D) \leqslant 1$，所以 $0 \leqslant 2\gamma_{(a_i, a_j)}^{\delta}(D) - \gamma_{a_i}^{\delta}(D) - \gamma_{a_j}^{\delta}(D) \leqslant 2$，因此，$0 \leqslant \Gamma_{(a_i, a_j)}^{\delta}(D) \leqslant 1$.

性质 3 $\Gamma_{(a_i, a_j)}^{\delta}(D) = 0$ 当且仅当 $\gamma_{(a_i, a_j)}^{\delta}(D) = \gamma_{a_i}^{\delta}(D) = \gamma_{a_j}^{\delta}(D)$.

证明：由定义 1 可知，$\Gamma_{(a_i, a_j)}^{\delta}(D) = 0$ 当且仅当 $2\gamma_{(a_i, a_j)}^{\delta}(D) = \gamma_{a_i}^{\delta}(D) + \gamma_{a_j}^{\delta}(D)$. 由于 $\gamma_{a_i}^{\delta}(D) \leqslant \gamma_{(a_i, a_j)}^{\delta}(D)$ 且 $\gamma_{a_j}^{\delta}(D) \leqslant \gamma_{(a_i, a_j)}^{\delta}(D)$，所以，$\gamma_{(a_i, a_j)}^{\delta}(D) = \gamma_{a_i}^{\delta}(D) = \gamma_{a_j}^{\delta}(D)$.

性质 3 表明：特征 a_i, a_j 以及二者的联合 $\{a_i, a_j\}$ 产生的邻域信息粒关于决策的认识是相同的(下近似

相同). 在这种情况下，两个特征存在冗余性，最多仅需一个特征即可刻画对象的决策特性.

显然，$\Gamma^{\delta}_{(a_i,a_i)}(D)=0$. 该结论表明，同一特征多次联合与独立用于刻画对象的决策是相同的. 从性质 2 的证明可知，平均联合依赖度增益关于特征数量是单调递增的. 因此，下面性质成立.

性质 4　$\Gamma^{\delta}_{(a_i,a_j)}(D)=1$ 当且仅当 $\gamma^{\delta}_{(a_i,a_j)}(D)=1$ 且 $\gamma^{\delta}_{a_i}(D)=\gamma^{\delta}_{a_j}(D)=0$.

根据性质 4，如果 $\Gamma^{\delta}_{(a_i,a_j)}(D)=1$，则决策 D 独立于特征 a_i 和 a_j 中的任何一个，即论域 U 中的每个对象都不能被特征 a_i 或 a_j 有效辨识. 然而，决策 D 完全依赖这两个特征的联合. 也就是说，在论域被特征集 $\{a_i, a_j\}$ 做邻域粒化后，每个对象关于决策 D 都可以被辨识. 这一事实表明：特征 a_i 和 a_j 是独立的，它们对辨识对象均是必要的.

1.3　基于冗余度的特征聚类

基于 1.2 节的分析，如果两个特征的联合依赖度增益低，该特征对具有高的冗余度. 反之，如果其联合依赖度增益越高，则特征对决策分类越是必要的，且这两个特征的冗余性越低. 鉴于此，本文构建特征冗余图，并借助图割理论对特征做划分，形成特征冗余簇.

定义 2　给定一个决策信息系统 $DIS=\langle U, A, V, D\rangle$，$\delta>0$，记 $\boldsymbol{W}=(w_{ij})_{m\times m}$，其中

$$w_{ij}=1-\Gamma^{\delta}_{(a_i,a_j)}(D) \tag{2}$$

称 $\boldsymbol{W}$ 为特征冗余度矩阵.

显然，$\boldsymbol{W}$ 是一个相似矩阵，也可修正 $\boldsymbol{W}$，将其对角线元素全赋值为 0，表明特征和自身不存在冗余性问题. 在后续的特征冗余图图割划分中，特征冗余度矩阵修正与否对后续结果没有实质影响. 将特征集 A 中的元素作为顶点，将任意两个特征间的冗余度 w_{ij} 作为对应顶点间的边权值，形成一个特征冗余图 G. 该图是一个加权无向图.

假设将特征冗余图 G 划分成 k 个连通子图 $A_1, A_2, \cdots, A_k$，构建图割损失函数[12]

$$Loss(A_1, \cdots, A_k)=\sum_{i=1}^{k}\frac{cut(A_i, \overline{A_i})}{|A_i|} \tag{3}$$

其中 $cut(A, B)=\sum_{i\in A, j\in B} w_{ij}$. 使损失函数(3) 达到最小的图割就是最优特征聚类. 设

$$h_{i,j}=\begin{cases}\dfrac{1}{\sqrt{|A_j|}}, & a_i\in A_j\\ 0, & a_i\notin A_j\end{cases}$$

称 $\boldsymbol{H}=(h_{ij})_{m\times k}$ 为指标矩阵，其满足 $\boldsymbol{H}'\boldsymbol{H}=\boldsymbol{I}_{k\times k}$. 这样，(3) 式可转化为如下的矩阵形式

$$Loss(A_1, \cdots, A_k)=\mathrm{tr}(\boldsymbol{H}'\boldsymbol{L}\boldsymbol{H}) \tag{4}$$

其中 $\boldsymbol{L}=\boldsymbol{\Lambda}-\boldsymbol{W}$，称之为图 G 的拉普拉斯矩阵，$\boldsymbol{\Lambda}$ 为对角矩阵，其对角线元素分别为冗余度矩阵 $\boldsymbol{W}$ 对应的行元素之和.

求解(3) 式或(4) 式的极小值问题是一个 NP 难问题. 将损失函数(4) 连续化，根据 Rayleigh-Ritz 定理[13]，目标函数(4) 的最优解 $\boldsymbol{H}$ 为 $\boldsymbol{L}$ 的前 k 个最小非零特征值对应的特征向量按列组成的矩阵.

矩阵 $\boldsymbol{H}$ 的每个列向量实质上就是图 G 顶点被分割后的簇标. 为了获取明确的簇标特征，采用常用的 K-均值聚类方法对矩阵 $\boldsymbol{H}$ 的行向量聚类. K-均值聚类结果代表了特征的类簇结构.

1.4　最优特征簇结构

由于 K-均值聚类方法需要事先给定类簇数，但数据特征的类簇数并不明确. 为了获得特征的最优聚类结果，引入一种特征聚类簇数的评估度量方法指导特征类簇数的选取.

假设特征集 A 的聚类结果为 $FC=\{A_1, A_2, \cdots, A_k\}$，其中 $A_i=\{a_{i1}, a_{i2}, \cdots, a_{i|A_i|}\}$ 为 A 的第 i 簇特征集，称

$$\overline{\lambda}(A_i)=\frac{1}{|A_i|^2}\sum_{p,q=1}^{|A_i|} w(a_{ip}, a_{iq}) \tag{5}$$

为 A_i 的簇内平均冗余度；称

$$\overline{\lambda}(A_i, A_j)=\frac{1}{|A_i|\cdot|A_j|}\sum_{q=1}^{|A_j|}\sum_{p=1}^{|A_i|} w(a_{ip}, a_{jq}) \tag{6}$$

为两个特征簇 $A_i=\{a_{i1}, a_{i2}, \cdots, a_{i|A_i|}\}$ 和 $A_j=\{a_{j1}, a_{j2}, \cdots, a_{j|A_j|}\}$ 的簇间平均冗余度.

最优聚类结果应具有簇内特征冗余度最大而簇间特征最不相关的特点，也即簇内平均冗余度最大而簇间平均冗余度最小. 因此，构造如下最优类簇结构评估指标：

$$C_{\text{index}}^{(k)}=\frac{2}{(k-1)}\sum_{i,\,j=1,\,i\neq j}^{k}\bar{\lambda}(A_i, A_j)-\sum_{i=1}^{k}\bar{\lambda}(A_i) \tag{7}$$

最优类簇结构评估指标依赖于聚类方法的不同，也取决于类簇数的选取. 评估指标值 $C_{\text{index}}^{(k)}$ 越小，说明簇内的冗余性越大，簇间的冗余性越小，聚类效果越好. 因此，使式(7)达到最小的正整数 k' 就是特征的最优划分簇数，对应的划分簇是 $FC=\{A_1, A_2, \cdots, A_{k'}\}$.

2 簇内特征选择机制

在特征聚类基础上，本节将探讨簇内特征选择策略. 互信息是度量两个随机变量间相关性的重要指标[14]；本文提出一种基于互信息的簇内特征选择方法.

假设特征集 A 的聚类结果为 $FC=\{A_1, A_2, \cdots, A_k\}$，对任意特征 $a_i\in A$，存在唯一类簇 A_j，使得 $a_i\in A_j$，称

$$\chi(a_i)=\frac{1}{|A_j|-1}\sum_{a_p\in A_j}w(a_i, a_p) \tag{8}$$

为特征 a_i 的依存度.

特征的依存度表示该特征与其簇内其他特征的平均冗余度. 特别地，如果某一特征簇中仅有一个元素，则该元素的依存度为0.

设对象集关于决策属性 D 的划分为 $U/D=\{D_1, D_2, \cdots, D_M\}$，对任意特征 $a_i\in A$，存在唯一类簇 A_j，使得 $a_i\in A_j$，a_i 关于决策 D 的互信息可表示为

$$I(a_i; D)=H(a_i)+H(D)-H(a_i; D)=$$
$$\sum_{D_t\in U/D}\sum_{x\in U}P(R_{\{a_i\}}^{\delta}(x), D_t)\log\frac{P(R_{\{a_i\}}^{\delta}(x), D_t)}{P(R_{\{a_i\}}^{\delta}(x))P(D_t)} \tag{9}$$

其中：$P(R_{\{a_i\}}^{\delta}(x))=\dfrac{|R_{\{a_i\}}^{\delta}(x)|}{|U|}$，$P(D_t)=\dfrac{|D_t|}{|U|}$，$P(R_{\{a_i\}}^{\delta}(x), D_t)=\dfrac{|R_{\{a_i\}}^{\delta}(x)\cap D_t|}{|U|}$.

从式(9)可以看出，如果特征 a_i 诱导的邻域信息粒 $R_{\{a_i\}}^{\delta}(x)$ 中绝大部分与相关决策等价类 D_t 都有较大交集，则 a_i 关于决策 D 的互信息就大，此时，a_i 与决策更具相关性，a_i 的分类能力就很强. 特别地，如果对于任意 $x\in U$ 都有 $R_{\{a_i\}}^{\delta}(x)=U$ 则 $I(a_i; D)=0$，即 a_i 关于决策 D 的互信息取得最小值0. 此时该特征没有任何分类能力. 由此可见，互信息 $I(a_i; D)$ 可以反映特征与决策之间的相关程度，互信息越大，特征与决策越相关，反之亦然.

对于任意特征簇 A_i，设 $a_{ij}\in A_i$ 是满足 $I(a_{ij}; D)=\max\limits_{k=1, 2, \cdots, |A_i|}I(a_{ik}; D)$ 的特征，则 a_{ij} 诱导的邻域信息粒包含在决策等价类中的整体程度最大. 因为簇内特征具有强冗余性，当簇内其余特征与 a_{ij} 共同粒化论域时，只能进一步细化邻域信息粒，对于邻域信息粒包含在决策等价类的整体程度越大的特征，细化能力越弱，该特征在簇内的整体依存度越大. 在这种情况下，特征 a_{ij} 在簇 A_i 中起着中心点的作用，称特征 a_{ij} 为对应簇 A_i 的中心特征，称互信息 $I(a_{ij}; D)$ 为簇 A_i 的中心度.

中心特征是与决策具有最大互信息的特征，是该簇中最具代表性特征. 根据聚类思想，中心特征与该簇中其余特征间具有最大的冗余性. 另一方面，不同簇中的中心特征具有最小的冗余性和最大的独立性，而且，由于特征类簇是按照特征簇内冗余度最大、簇间冗余度最小确定的最优特征分割模式，将每个簇的中心特征作为特征选择子集是合理的，也是足够的. 简记这种特征选择方法为 RDCFS.

下面我们以一个具体实例说明中心特征与其依存度和中心度间的关联性. 从 UCI 数据库中选取 glass 数据集，并从中随机抽取6个样本，如下表1.

表 1　glass 数据集中部分数据

	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	D
x_1	35	1.517 83	12.69	3.54	1.34	72.95	0.57	8.75	0	0	1
x_2	31	1.517 68	12.65	3.56	1.30	73.08	0.61	8.69	0	0.14	1
x_3	120	1.516 52	13.56	3.57	1.47	72.45	0.64	7.96	0	0	2
x_4	124	1.517 07	13.48	3.48	1.71	72.52	0.62	7.99	0	0	2
x_5	159	1.517 76	13.53	3.41	1.52	72.04	0.58	8.79	0	0	3
x_6	194	1.517 19	14.75	0	2	73.02	0	8.53	1.59	0.08	4

记其特征为 $A=\{a_1, a_2, \cdots, a_{10}\}$，将数据归一化后，按照第 1 节方法得到特征聚类结果：$A_1=\{a_1, a_6\}$，$A_2=\{a_2, a_4, a_7, a_9, a_{10}\}$，$A_3=\{a_3, a_5, a_8\}$. 根据式(8)和(9)分别计算 10 个特征的依存度和中心度，见图 1. 在该图中，同一个类簇中的特征用同一种符号和相同颜色的点表示，不同簇中特征用不同符号和不同颜色表示.

从图 1 可以看出，在同一个簇内，中心度越大的特征，在该簇内的依存度也越大，中心特征的依存度最大. 易见，a_6，a_2 和 a_3 的中心度和依存度在对应簇中最大，它们分别就是簇 A_1，A_2 和 A_3 的中心特征. 因此，$\{a_2, a_3, a_6\}$ 是 A 的最优特征选择子集.

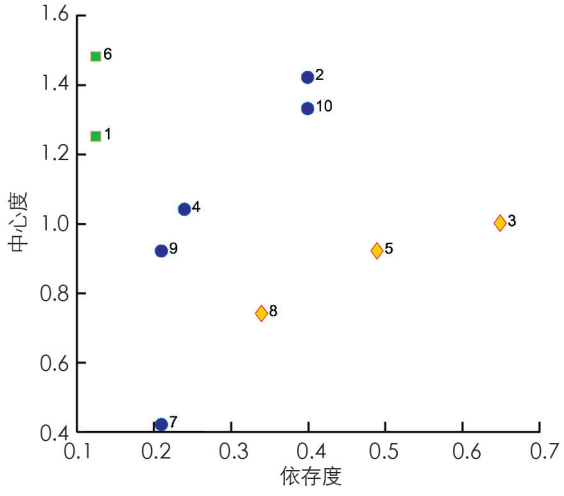

图 1　特征依存度和中心度分布

综合上述分析，下面给出决策系统的基于决策依赖聚类的特征选择算法.

算法 1　基于决策依赖聚类的特征选择算法(RDCFS)

输入：训练数据集 IS=〈U，A，V，D〉，m 为特征数，邻域参数 δ，指定聚类个数范围 Ξ；

输出：特征子集 S.

BEGIN

1. 初始化被选特征子集 S=Φ，全部特征集合为 F；
2. for each k ∈ Ξ；　　// Ξ：指定聚类个数范围
3. 基于公式(4)实现特征聚类；
4. 计算公式(5),(6),(7)；
5. end for
6. $k'=\arg\min_{k} C_{\text{index}}^{(k)}$；　　// 选出最优簇类数
7. FC=$\{A_1, A_2, \cdots, A_{k'}\}$；
8. for each $A_i \in$ FC
9. for each $a_i \in A_j$，计算中心度(9)；
10. $b_i=\arg\max_{a_i \in A_j} I(a_i; D)$；　　// 选出每个簇内中心度最大的特征
11. end for
12. end for
13. 特征子集 S=$\{b_1, b_2, \cdots, b_{k'}\}$.

END

3 模拟实验及结果分析

为了验证本文提出的特征选择方法的合理性和有效性，从UCI数据库中选取8个数值型数据集，详细信息见表2. 利用这8个数据集验证本文所提特征选择方法选出的特征子集在分类精度上的性能.

表2 数据描述及特征选择结果

数据集	样本数	特征数	决策类数
CT	221	36	2
wpbc	198	32	2
sonar	208	60	2
autovalve_B	414	72	5
PG	48	321	16
colon	62	1224	2
gene10	88	2308	5
gene3	88	4026	6

首先以表2中前4个数据集为例验证根据式(7)选取聚类数的合理性. 实验采取十折交叉验证的方法. 根据邻域粗糙集特征选择方法邻域参数的一般选取规则，本文中δ在区间$[0.01, 0.5]$内取值. 给定类簇数的取值范围，以一定步长通过遍历方法得到使式(7)达到最小的聚类数目. 在聚类基础上，在每个簇内选择中心度最高的特征形成特征子集. 采用如下所述邻域投票精度作为特征选择有效性的度量指标.

设特征集$A=\{a_1, a_2, \cdots, a_m\}$的特征选择结果为$S=\{a_{c1}, a_{c2}, \cdots, a_{ck}\}$，任意对象$x_i \in U$关于特征子集$S$的$\delta$邻域信息粒为$R_S^{\delta}(x_i)=\{y \in U \mid d_S(x_i, y) \leqslant \delta\}$，规定$x_i$的准类别标签$d_i^{\delta}$为$R_S^{\delta}(x_i)$中对象的真实标签采用多数投票策略获得的标签，则邻域投票精度为

$$\Pr(S)=\sum_{x_i \in U} \frac{|\, v(x_i, D)=d_i{}^{\delta} \,|}{|U|} \tag{10}$$

其中$|\, v(x_i, D)=d_i^{\delta} \,|=\begin{cases}1, & v(x_i, D)=d_i^{\delta} \\ 0, & \text{否则}\end{cases}$，$v(x_i, D)$为$x_i$的真实类别标签，$v(x_i, D)=d_i^{\delta}$表示$x_i$的真实类别标签与准类别标签相同.

图2绘出了表2中前4个数据集特征对的不同聚类数目对应的类簇结构评价指标和邻域投票精度的变化情况. 其中不同线形、不同颜色分别代表数据的类簇结构评价指标和邻域投票精度的变化情况.

由图2分析可知，给定一定的聚类数目范围：在CT和wpbc两个数据集上，类簇结构评价指标达到极小，即将特征分别聚成17类和14类时，邻域投票精度达到最高；在sonar和autovalve_B两个数据集的类簇结构评价指标达到极小时，所选特征的邻域投票精度达到次最高，且当聚类数目逐渐增加时，邻域投票精度呈现下降趋势. 这说明，我们构造的类簇结构评价指标能够驱动确定有效的聚类数目，从而促进特征选择的精度提高.

为了进一步验证本文提出方法的优势，分别与联合谱聚类与邻域互信息的特征选择算法(SCNMI)[15]、基于特征聚类的特征选择方法(FSFC)[16]、模糊粗糙测度特征选择(FRSE)[17]和模糊邻域粗糙集特征选择(FNRS)[7]4种特征选择方法在8个数据集上进行模拟实验比较. 表3给出了5种特征选择方法对8个数据集做出的特征选择结果，其中的数值代表最终特征选择子集中的特征数.

从表3可以看出，在8个数据集的实验中，本文所提方法(RDCDS)在7个数据集上都选出了最优(最小)特征子集，尤其对高维数据集具有更好的降维作用.

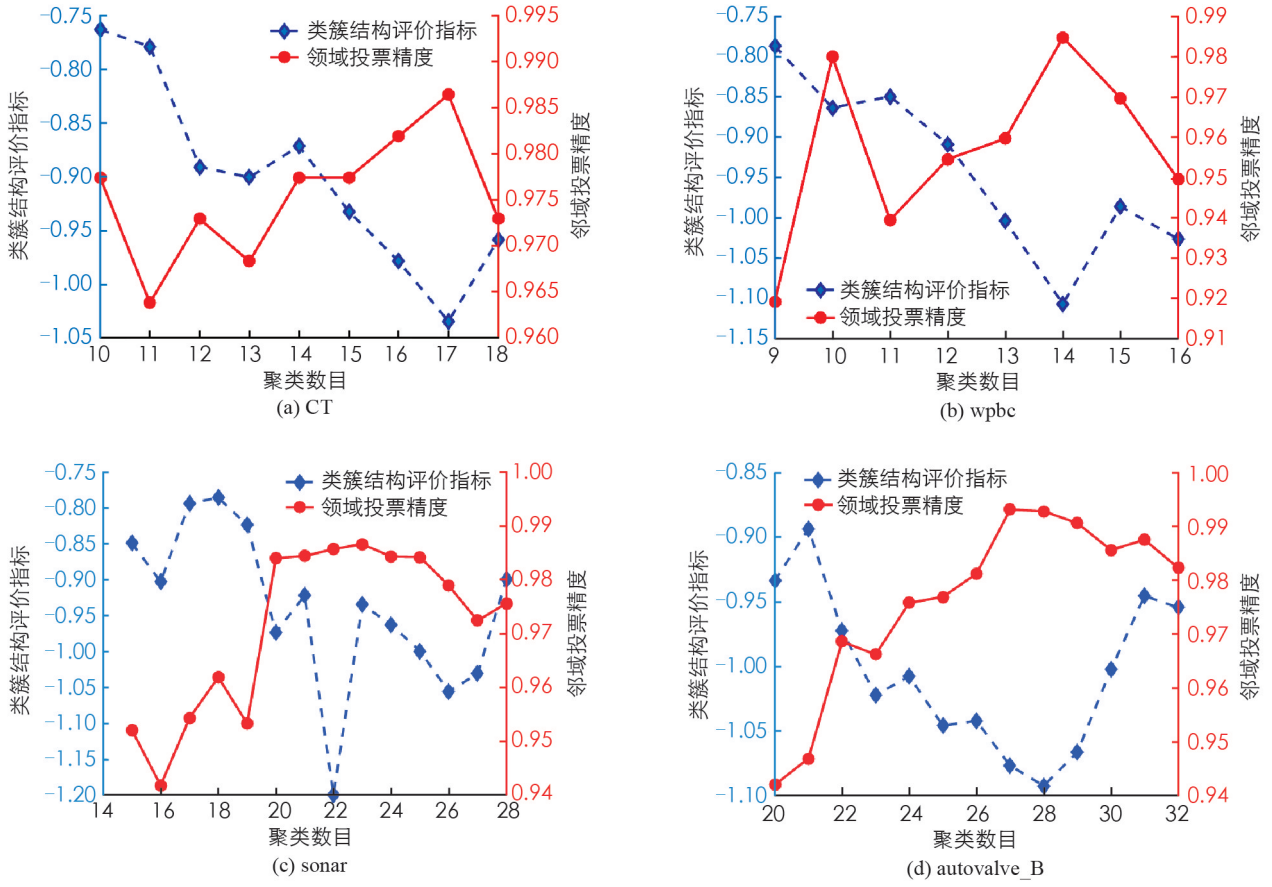

图 2　类簇结构评价指标和精度的变化

表 3　特征子集中的特征数

数据集	RDCDS	SCNMI	FSFC	FRSE	FNRS
CT	17	18	23	21	28
wpbc	14	18	19	14	15
sonar	22	22	25	35	42
autovalve_B	28	31	31	30	39
PG	51	56	48	43	61
colon	54	54	75	94	73
gene10	52	88	86	126	86
gene3	104	167	104	177	170
平均值	43	57	51	68	64

为了验证特征子集的分类能力，分别采用 KNN 分类器(本文取 $K=3$)和支持向量机分类器(SVM)对数据集在特征选择前后的分类效果进行了对比实验，各个方法在不同数据集上的分类精度如表 4 和表 5 所示.

从表 4 和表 5 可以看出，不论采用 KNN 分类器还是 SVM 分类器，在 8 个数据集的实验中本文所提方法(RDCDS)在 5 个数据集上取得了最高的分类精度. FNRS 在 2 个数据集上取得了最高 KNN 分类精度，FSFC 只在一个数据集上取得最高 KNN 分类精度，而 SCNMI 和 FRSE 表现最差. 同时，FSFC，FRSE 和 FNRS 分别仅在一个数据集上取得最高的 SVM 分类精度，SCNMI 依然表现最差.

尽管如此，特征选择后的数据分类精度都高于原始数据的分类精度. 这一事实说明原始数据包含冗余和不相关信息，扰乱了对数据的分类学习，进一步证实特征选择的必要性. 上述对比实验表明，从统计学意义上讲，本文提出的特征选择方法得到的特征子集不仅紧凑，数据分类精度还得以显著提高.

表4 KNN分类器($K=3$)下的分类精度

数据集	Raw	RDCDS	SCNMI	FSFC	FRSE	FNRS
CT	86.90±7.80	91.86±7.64	88.70±5.74	89.17±5.22	91.30±3.36	91.57±3.44
wpbc	74.28±3.37	77.74±4.26	75.28±9.76	77.89±5.47	76.33±6.34	75.96±3.24
sonar	87.94±7.60	92.10±3.60	89.37±7.81	88.42±7.59	88.9±8.19	92.15±2.94
autovalv_B	94.74±4.25	98.95±6.35	96.78±5.62	97.26±3.68	98.32±7.24	97.25±2.56
PG	2.00±6.32	27.33±16.87	15.33±14.76	19.33±12.35	26.67±16.61	30.67±23.82
colon	64.58±9.47	77.50±14.33	72.61±10.48	65.83±12.7	75.83±19.42	69.17±18.45
gene10	25.08±10.25	79.37±11.78	64.92±17.35	66.03±19.21	42.22±20.51	65.71±13.18
gene3	15.79±7.44	52.70±14.53	37.89±13.54	64.21±17.58	47.68±8.14	50.98±7.66
平均值	56.41±7.06	74.69±9.92	67.61±10.63	71.01±12.87	68.40±11.22	73.33±9.48

表5 SVM分类器下的分类精度

数据集	Raw	RDCDS	SCNMI	FSFC	FRSE	FNRS
CT	88.68±5.38	93.95±6.43	90.47±5.87	93.36±3.28	89.65±2.86	92.58±3.67
wpbc	72.17±8.75	75.74±9.90	73.00±3.67	74.56±2.69	73.22±2.74	77.55±3.66
sonar	82.13±10.00	87.26±5.60	82.66±13.37	82.61±11.46	84.35±2.14	85.25±3.58
autovalv_B	95.17±3.45	98.65±3.57	97.25±3.62	95.99±7.16	97.57±4.85	99.10±3.16
PG	33.33±20.91	50.33±24.80	49.36±16.69	48.46±9.63	52.56±9.36	49.89±12.54
colon	74.58±16.49	79.17±15.34	80.83±20.05	83.75±17.62	77.50±15.74	76.25±16.44
gene10	68.57±15.11	87.14±11.84	85.56±14.86	83.33±13.09	79.52±19.38	79.84±15.06
gene3	52.11±8.02	96.67±3.37	61.05±6.18	60.00±8.66	93.89±3.51	78.86±7.64
平均值	70.84±11.01	83.61±10.23	76.27±10.48	77.75±9.20	81.03±7.56	79.91±8.21

4 结论

本文基于邻域粗糙集模型建立了一种特征聚类驱动的特征选择方法. 该方法不同于现有启发式特征选择和基于特征聚类的特征选择，从特征依赖度出发构建了特征冗余图，给出了最优特征冗余图聚类准则和聚类方法. 在此基础上，通过引入簇内特征依存度和中心度的概念，给出了特征依赖聚类的特征选择算法. 理论分析和多个数据集上的对比实验结果表明新提出的特征选择方法不仅可以选出特征数更小的特征子集，其对应的分类精度还得以提高.

本文中最优特征簇结构的评估指标通过遍历方式得到，探索一种特征类簇数自适应确定方法是很有意义的. 通过分析特征与特征间的因果关联，基于有向特征图图割的特征选择方法是一个值得深入研究的问题.

参考文献：

[1] VENKATESH B, ANURADHA J. A Review of Feature Selection and Its Methods [J]. Cybernetics and Information Technologies, 2019, 19(1): 3-26.

[2] PENG C, KANG Z, YANG M, et al. Feature Selection Embedded Subspace Clustering [C]//IEEE Signal Processing Letters. New York: IEEE Press, 2018: 1018-1022.

[3] MAFARJA M, MIRJALILI S. Whale Optimization Approaches for Wrapper Feature Selection [J]. Applied Soft Computing, 2018, 62: 441-453.

[4] ABU SHANAB A, KHOSHGOFTAAR T. Filter-Based Subset Selection for Easy, Moderate, and Hard Bioinformatics Data [C]//2018 IEEE International Conference on Information Reuse and Integration. New York: IEEE Press, 2018: 372-377.

[5] SWINIARSKI R W, SKOWRON A. Rough Set Methods in Feature Selection and Recognition [J]. Pattern Recognition Letters, 2003, 24(6): 833-849.

[6] HU Q H, YU D R, LIU J F, et al. Neighborhood Rough Set Based Heterogeneous Feature Subset Selection [J]. Information Sciences, 2008, 178(18): 3577-3594.

[7] WANG C Z, SHAO M W, HE Q, et al. Feature Subset Selection Based on Fuzzy Neighborhood Rough Sets [J]. Knowledge-Based Systems, 2016, 111: 173-179.

[8] TANG X, DONG M, BI S, et al. Feature Selection Algorithm Based on K-Means Clustering [C]//2017 IEEE 7th Annual International Conference on CYBER Technology in Automation, Control, and Intelligent Systems. New York: IEEE Press, 2017: 1522-1527.

[9] SONG Q B, NI J J, WANG G T. A Fast Clustering-Based Feature Subset Selection Algorithm for High-Dimensional Data [J]. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(1): 1-14.

[10] LIU Y, LI B F. Bayesian Hierarchical K-Means Clustering [J]. Intelligent Data Analysis, 2020, 24(5): 977-992.

[11] CHEN Y M, ZENG Z Q, LU J W. Neighborhood Rough Set Reduction with Fish Swarm Algorithm [J]. Soft Computing, 2017, 21(23): 6907-6918.

[12] WANG B J, ZHANG L, WU C L, et al. Spectral Clustering Based on Similarity and Dissimilarity Criterion [J]. Pattern Analysis and Applications, 2017, 20(2): 495-506.

[13] XIAO J Y, ZHOU H, ZHANG C Z, et al. Solving Large-Scale Finite Element Nonlinear Eigenvalue Problems by Resolvent Sampling Based Rayleigh-Ritz Method [J]. Computational Mechanics, 2017, 59(2): 317-334.

[14] YU S W, HUANG T Z. Exponential Weighted Entropy and Exponential Weighted Mutual Information [J]. Neurocomputing, 2017, 249: 86-94.

[15] 胡敏杰，郑荔平，唐莉，等. 联合谱聚类与邻域互信息的特征选择算法 [J]. 模式识别与人工智能，2017，30(12)：1121-1129.

[16] 王连喜，蒋盛益. 一种基于特征聚类的特征选择方法 [J]. 计算机应用研究，2015，32(5)：1305-1308.

[17] HU Q H, YU D, XIE Z X, et al. Fuzzy Probabilistic Approximation Spaces and Their Information Measures [J]. IEEE Transactions on Fuzzy Systems, 2006, 14(2): 191-201.

责任编辑　张枸