中国科学 信息科学
SCIENTIA SINICA Informationis

基于样本稳定性的聚类方法

李飞江, 钱宇华, 王婕婷, 梁吉业 and 王文剑

Citation: 中国科学: 信息科学 **50**, 1239 (2020); doi: 10.1360/SSI-2019-0110

View online: http://engine.scichina.com/doi/10.1360/SSI-2019-0110

View Table of Contents: http://engine.scichina.com/publisher/scp/journal/SSI/50/8

Published by the 《中国科学》杂志社

中国科学: 信息科学 2020 年 第 50 卷 第 8 期: 1239–1254
SCIENTIA SINICA Informationis
论文

基于样本稳定性的聚类方法

李飞江[1], 钱宇华[1,2*], 王婕婷[1], 梁吉业[2], 王文剑[2]

1. 山西大学大数据科学与产业研究院, 太原 030006
2. 山西大学计算智能与中文信息处理教育部重点实验室, 太原 030006
* 通信作者. E-mail: jinchengqyh@126.com

收稿日期: 2019–05–30; 修回日期: 2019–08–23; 接受日期: 2019–11–01; 网络出版日期: 2020–08–06

国家重点研发计划 (批准号: 2018YFB1004300)、山西省重点研发计划 (批准号: 201903D421003)、国家自然科学基金 (批准号: 61672332, U1805263, 61872226, 61802238, 61673249) 和山西省海外归国人员研究项目 (批准号: 2017023, 2016004) 资助

摘要 数据类型和分布的复杂化导致样本间关系的不确定性增强, 给有效挖掘数据的潜在类簇结构带来挑战. 为降低样本关系不确定性对数据聚类带来的影响, 本文将聚类集成中样本稳定性概念扩展至聚类分析中. 本文从理论上分析样本稳定的合理性, 并提出基于信息熵的样本稳定性度量方法. 此外, 本文提出一个基于样本稳定性的聚类方法, 该方法先将数据分为稳定样本集和不稳定样本集, 然后挖掘稳定样本的团簇结构, 并将不稳定样本划分至该团簇结构中. 最后, 本文通过二维人造数据和图像分割场景可视化显示样本稳定性的合理性, 并在基准数据集上验证本文所提聚类算法的有效性.

关键词 机器学习, 无监督学习, 聚类分析, 样本稳定性, 稳定性理论

1 引言

随着数据采集、存储、传输方式多样化、泛在化发展, 各类数据智能分析场景面临大量无标记数据. 机器学习中, 对这类数据的学习称为无监督学习. 图灵奖获得者 Geoffrey Hinton, Yann LeCun, 和 Yoshua Bengio [1] 在 *Nature* 上合作发表的文章 "Deep learning" 中指出: "长远看, 我们预计无监督学习将变得更加重要." 聚类分析 [2] 是无监督学习中的重要研究内容, 其目的在于学习数据中潜在的团簇结构, 即相似的样本划分在同一类簇中. 目前, 聚类分析已经广泛应用于实际数据分析任务中, 如目标检索 [3]、自然语言处理 [4]、生物信息学 [5] 等. 此外, 数据聚类也作为重要预处理策略应用在很多数据分析技术中, 如特征选择 [6]、深度学习 [7] 等.

现阶段, 基于数据类型、数据分布假设、应用场景等, 研究者已经提出大量聚类算法, 通常可粗略归类为原型聚类、密度聚类、层次聚类. 由于缺乏监督信息, 样本间关系的度量在聚类算法中起重要作用, 比如原型聚类将各样本点与最相似或最近的代表点划分为同类; 密度聚类中样本密度通常由相似

引用格式: 李飞江, 钱宇华, 王婕婷, 等. 基于样本稳定性的聚类方法. 中国科学: 信息科学, 2020, 50: 1239–1254, doi: 10.1360/SSI-2019-0110
Li F J, Qian Y H, Wang J T, et al. Clustering method based on sample's stability (in Chinese). Sci Sin Inform, 2020, 50: 1239–1254, doi: 10.1360/SSI-2019-0110

或相近的样本数估计; 层次聚类则逐层合并最相似的类簇. 此外, 很多聚类算法则直接基于样本的关系矩阵, 如谱聚类算法 (spectral clustering)、Frey 等 [8] 在 *Science* 上提出的仿射传播聚类算法 (affinity propagation, AP)、Rodriguez 等 [9] 在 *Science* 上提出的密度峰值聚类算法 (density peak, DP)、Otto 等 [10] 提出的处理大规模人脸数据的聚类方法等. 大数据时代, 数据类型呈现多样化, 如符号型数据、图像数据、文本数据、基因数据以及多种数据类型的混合等 [5, 11~14]; 数据分布呈现复杂化, 如流型数据、混合分布数据、不平衡数据、多视图数据等 [15~19]. 这些数据特点导致样本间关系具有更强的不确定性, 给有效挖掘数据潜在团簇结构带来挑战.

文献 [20] 在聚类集成场景下研究了样本关系的不确定性, 提出了样本稳定性的概念来发现具有稳定关系的样本集, 降低了样本关系不确定性对聚类的影响, 并建立了基于样本稳定性的聚类集成算法. 实验显示其集成性能优于传统聚类集成算法. 文献 [20] 的研究为应对样本间关系的不确定性提供了新思路, 但其研究背景仅在聚类集成问题中. 聚类集成技术 [21, 22] 通过融合多个异质的聚类结果获得数据的类簇结构, 是提高单一聚类方法性能的重要途径, 近年来受到广泛关注. 本质上, 聚类集成是从多个划分的角度估计样本的共现概率, 即两个样本最终出现在同一类的概率. 而该共现概率也可基于原始数据特征直接估计, 因此, 将样本稳定性概念扩展至聚类分析中是可行的. 基于上述讨论, 本文主要研究聚类分析中的样本稳定性, 并提出相应聚类方法.

文本贡献主要体现在以下 3 方面:

(1) 将聚类集成中的样本稳定性扩展至聚类分析中, 并给出理论上的合理性分析;

(2) 提出基于信息熵的样本稳定性度量函数, 并通过可视化实验验证其合理性;

(3) 提出一种基于样本稳定性的聚类方法. 并在基准数据集上验证了其有效性.

本文在第 2 节给出样本稳定性定义和合理性理论分析, 并提出基于信息熵的样本稳定性度量函数. 第 3 节介绍基于样本稳定性聚类方法的算法流程. 第 4 节从实验上验证基于信息熵样本稳定性度量的合理性和基于样本稳定性聚类方法的有效性. 最后总结全文.

2 聚类分析中的样本稳定性

文献 [20] 研究了聚类集成中的样本稳定性. 基于一组聚类结果, 易得两样本被划分在同一类的频率. 该频率估计了两样本的共现概率. 当两样本共现概率为 1 时, 可确定这两个样本最终在同一类; 同理可确定共现概率为 0 的两个样本最终在不同类. 取值中间的共现概率给判断两样本划分带来困难. 两个样本之间的稳定关系应包含两个方面: (1) 两个样本具有较高的确定性在同一类, 即两样本具有较高的共现概率值; (2) 两个样本具有较高的确定性不在同一类, 即两样本具有较低的共现概率值. 直接用共现概率大小表示关系的稳定性不能反应第 2 方面. 为此, 文献 [20] 提出确定性函数的概念来评价样本关系的确定度, 并将一个样本与其他样本共现概率的平均确定度定义为该样本的稳定性. 样本稳定性可用于发现具有稳定关系的样本子集, 度量样本对挖掘类簇结构的贡献度, 有助于数据聚类问题的有效解决.

样本间共现概率也可基于数据的原始特征估计, 如核相似度、共同邻域等. 因此文献 [20] 中的样本稳定性可扩展至聚类分析中. 本节介绍样本稳定性的基本定义, 从理论上分析其合理性, 并提出基于信息熵的样本稳定性度量函数.

2.1 基本定义

假设两样本的共现概率为 p, 下面定义评价这两个样本关系确定度的确定性函数.

定义1 (确定性函数 [20]) 对于变量 $p \in [0,1]$ 和常量 $t \in (0,1)$, 函数 f 是一个确定性函数, 如果 f 满足条件:

(1) 如果 $p < t$, 则 $f'(p) < 0$; 如果 $p > t$, 则 $f'(p) > 0$;

(2) 如果 $p_b < t < p_d$, $\frac{t-p_b}{p_d-t} = \frac{t}{1-t}$, 则 $f(p_b) = f(p_d)$,

其中 t 为样本间关系最不确定时的共现概率值, 该值可基于共现矩阵 $\boldsymbol{P} = \{p_{ij}|1 \leqslant i \leqslant n, 1 \leqslant j \leqslant n\}$ 获得. 条件 (1) 表示确定性函数在 t 处取最小值, 且离 t 越远取值越大. 条件 (1) 意味着共现概率取值距离 t 越远, 确定性程度越高. 条件 (2) 表示确定性函数在 t 两侧成比例对称, 也就是当位于 t 两侧数值 p_b 和 p_d ($p_b < t < p_d$) 距 t 的距离比值 $\frac{t-p_b}{p_d-t}$ 为 $\frac{t}{1-t}$ 时, p_b 和 p_d 取得相同的函数值.

基于给定的确定性函数 f, 对于数据集 $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$, 样本 $\boldsymbol{x}_i$ 的稳定性定义为该样本与其他样本共现概率值的平均确定度:

$$s(\boldsymbol{x}_i) = \frac{1}{n}\sum_{j=1}^{n} f(p_{ij}), \tag{1}$$

其中 n 为数据集中样本数, $p_{ij} \in \boldsymbol{P}$, $\boldsymbol{P} = \{p_{ij}|1 \leqslant i \leqslant n, 1 \leqslant j \leqslant n\}$.

2.2 样本稳定性合理性理论分析

本小节在高斯 (Gauss) 分布假设下分析样本稳定性的合理性. 首先假设两个服从高斯分布类簇 c_1 和 c_2. 然后计算样本属于各类的概率以及两两样本的共现概率, 通过共现概率和样本位置的关系分析样本稳定性取值的趋势. 最后基于确定性函数 fg 计算样本的稳定性 sg, 并通过预期趋势和 sg 取值趋势的关系来分析其合理性.

假设 c_1 和 c_2 表示两个服从高斯分布的类簇, 样本 $\boldsymbol{x} = \{d_1, d_2, \ldots, d_D\}^{\mathrm{T}}$ 的概率密度为

$$p(\boldsymbol{x}|y = c_1) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{D}{2}}}\frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}}\mathrm{e}^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})};$$
$$p(\boldsymbol{x}|y = c_2) = \mathcal{N}(\boldsymbol{x}; -\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{D}{2}}}\frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}}\mathrm{e}^{-\frac{1}{2}(\boldsymbol{x}+\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}+\boldsymbol{\mu})},$$

其中 $\boldsymbol{\mu}$ 为均值向量, $\boldsymbol{\Sigma}$ 为协方差矩阵, D 为数据维度.

给定样本 $\boldsymbol{x}$, 后验类归属概率 $p(y = c_1|\boldsymbol{x})$ 可由贝叶斯 (Bayes) 公式求得:

$$\begin{aligned} p(y = c_1|\boldsymbol{x}) &= \frac{p(\boldsymbol{x}, y = c_1)}{p(\boldsymbol{x})} \\ &= \frac{p(y = c_1)p(\boldsymbol{x}|y = c_1)}{\sum_{y=c_1,c_2} p(y)p(\boldsymbol{x}|y)} \\ &= \frac{\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})}{\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + \mathcal{N}(-\boldsymbol{\mu}, \boldsymbol{\Sigma})} \\ &= \frac{1}{1 + \mathrm{e}^{-2\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{x}}}. \end{aligned}$$

同理, 后验类归属概率 $p(y = c_2|\boldsymbol{x})$ 为

$$p(y = c_2|\boldsymbol{x}) = \frac{1}{1 + \mathrm{e}^{2\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{x}}}.$$

当 $D = 1$ 时, 后验类归属概率的表现形式与 sigmoid 函数形式一致, 函数曲线为 "S" 型. 图 1 给出 $\mu = 3$, $\sigma = 3$ 时的函数图像示例.

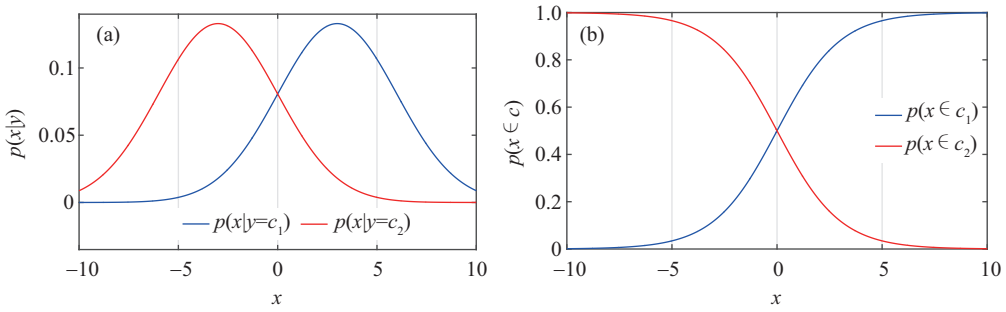


图 1 $\mu = 3$ 和 $\sigma = 3$ 时的函数图像示例

Figure 1 An example with $\mu = 3$ and $\sigma = 3$. (a) Conditional probability; (b) posterior probability

基于上述假设, 样本 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 出现在同一类簇中的概率为

$$\begin{aligned} p_{ij} &= p(y_i = c_1|\boldsymbol{x}_i)p(y_j = c_1|\boldsymbol{x}_j) + p(y_i = c_2|\boldsymbol{x}_i)p(y_j = c_2|\boldsymbol{x}_j) \\ &= \frac{1}{1+\mathrm{e}^{-2\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{x}_i}} \times \frac{1}{1+\mathrm{e}^{-2\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}x_j}} + \frac{1}{1+\mathrm{e}^{2\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{x}_i}} \times \frac{1}{1+\mathrm{e}^{2\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{x}_j}} \\ &= \frac{1+\mathrm{e}^{\boldsymbol{a}\boldsymbol{x}_i+\boldsymbol{a}\boldsymbol{x}_j}}{1+\mathrm{e}^{\boldsymbol{a}\boldsymbol{x}_i}+\mathrm{e}^{\boldsymbol{a}\boldsymbol{x}_j}+\mathrm{e}^{\boldsymbol{a}\boldsymbol{x}_i+\boldsymbol{a}\boldsymbol{x}_j}}, \end{aligned} \tag{2}$$

其中 $\boldsymbol{a} = 2\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}$. 方便起见, 本文剩余部分均用 $\boldsymbol{a}$ 表示 $2\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}$.

性质1 如果 $\boldsymbol{a}\boldsymbol{x} > 0$, 则 $p(y = c_1|\boldsymbol{x}) > p(y = c_2|\boldsymbol{x})$; 如果 $\boldsymbol{a}\boldsymbol{x} < 0$, 则 $p(y = c_1|\boldsymbol{x}) < p(y = c_2|\boldsymbol{x})$; 如果 $\boldsymbol{a}\boldsymbol{x} = 0$, 则 $p(y = c_1|\boldsymbol{x}) = p(y = c_2|\boldsymbol{x}) = \frac{1}{2}$.

证明

$$\begin{aligned} &\boldsymbol{a}\boldsymbol{x} > 0 \Rightarrow \mathrm{e}^{-\boldsymbol{a}\boldsymbol{x}} < \mathrm{e}^{\boldsymbol{a}\boldsymbol{x}} \Rightarrow p(y = c_1|\boldsymbol{x}) > p(y = c_2|\boldsymbol{x}), \\ &\boldsymbol{a}\boldsymbol{x} < 0 \Rightarrow \mathrm{e}^{-\boldsymbol{a}\boldsymbol{x}} > \mathrm{e}^{\boldsymbol{a}\boldsymbol{x}} \Rightarrow p(y = c_1|\boldsymbol{x}) < p(y = c_2|\boldsymbol{x}), \\ &\boldsymbol{a}\boldsymbol{x} = 0 \Rightarrow \mathrm{e}^{-\boldsymbol{a}\boldsymbol{x}} = \mathrm{e}^{\boldsymbol{a}\boldsymbol{x}} = 1 \Rightarrow p(y = c_1|\boldsymbol{x}) = p(y = c_2|\boldsymbol{x}) = \frac{1}{2}. \end{aligned}$$

综上, 性质 1 成立.

事实上, $\boldsymbol{a}\boldsymbol{x}$ 为样本 $\boldsymbol{x}$ 到类簇 c_2 的马氏距离的平方减去样本 $\boldsymbol{x}$ 到类簇 c_1 的马氏距离平方. 性质 1 表明, 当 $\boldsymbol{a}\boldsymbol{x} > 0$ 时, $\boldsymbol{x}$ 距离类簇 c_1 更近. 反之当 $\boldsymbol{a}\boldsymbol{x} < 0$ 时, $\boldsymbol{x}$ 距离类簇 c_2 更近.

式 (2) 所示的样本共现概率有如下性质.

性质2 如果 $(\boldsymbol{a}\boldsymbol{x}_i)(\boldsymbol{a}\boldsymbol{x}_j) < 0$, 则 $p_{ij} < \frac{1}{2}$; 如果 $(\boldsymbol{a}\boldsymbol{x}_i)(\boldsymbol{a}\boldsymbol{x}_j) = 0$, 则 $p_{ij} = \frac{1}{2}$; 如果 $(\boldsymbol{a}\boldsymbol{x}_i)(\boldsymbol{a}\boldsymbol{x}_j) > 0$, 则 $p_{ij} > \frac{1}{2}$.

证明 由式 (2) 可得

$$p_{ij} - \frac{1}{2} = \frac{1+\mathrm{e}^{\boldsymbol{a}\boldsymbol{x}_i+\boldsymbol{a}\boldsymbol{x}_j}}{1+\mathrm{e}^{\boldsymbol{a}\boldsymbol{x}_i}+\mathrm{e}^{\boldsymbol{a}\boldsymbol{x}_j}+\mathrm{e}^{\boldsymbol{a}\boldsymbol{x}_i+\boldsymbol{a}\boldsymbol{x}_j}} - \frac{1}{2} = \frac{(1-\mathrm{e}^{\boldsymbol{a}\boldsymbol{x}_i})(1-\mathrm{e}^{\boldsymbol{a}\boldsymbol{x}_j})}{1+\mathrm{e}^{\boldsymbol{a}\boldsymbol{x}_i}+\mathrm{e}^{\boldsymbol{a}\boldsymbol{x}_j}+\mathrm{e}^{\boldsymbol{a}\boldsymbol{x}_i+\boldsymbol{a}\boldsymbol{x}_j}} \times \frac{1}{2}.$$

当 $(\boldsymbol{a}\boldsymbol{x}_i)(\boldsymbol{a}\boldsymbol{x}_j) \neq 0$ 时, 根据 $(1-\mathrm{e}^{\boldsymbol{a}\boldsymbol{x}})(\boldsymbol{a}\boldsymbol{x}) < 0$, 有 $(1-\mathrm{e}^{\boldsymbol{a}\boldsymbol{x}_i})(1-\mathrm{e}^{\boldsymbol{a}\boldsymbol{x}_j}) \times (\boldsymbol{a}\boldsymbol{x}_i)(\boldsymbol{a}\boldsymbol{x}_j) > 0$. 即上述公式分子 $(1-\boldsymbol{a}\boldsymbol{x}_i)(1-\boldsymbol{a}\boldsymbol{x}_j)$ 和 $(\boldsymbol{a}\boldsymbol{x}_i)(\boldsymbol{a}\boldsymbol{x}_j)$ 同号. 进而, 当 $(\boldsymbol{a}\boldsymbol{x}_i)(\boldsymbol{a}\boldsymbol{x}_j) < 0$ 时, $p_{ij} < \frac{1}{2}$; 当 $(\boldsymbol{a}\boldsymbol{x}_i)(\boldsymbol{a}\boldsymbol{x}_j) > 0$ 时, $p_{ij} > \frac{1}{2}$.

当 $(\boldsymbol{a}\boldsymbol{x}_i)(\boldsymbol{a}\boldsymbol{x}_j) = 0$ 时, 有 $(1-\mathrm{e}^{\boldsymbol{a}\boldsymbol{x}_i})(1-\mathrm{e}^{\boldsymbol{a}\boldsymbol{x}_j}) = 0$, 则 $p_{ij} = \frac{1}{2}$.

综上, 性质 2 成立.

性质 2 表明, 当两样本离同一类簇近时, 共现概率 $p_{ij} > \frac{1}{2}$; 当两样本离不同类簇近时, 共现概率 $p_{ij} < \frac{1}{2}$.

性质3 对于任意 $\boldsymbol{ax}_k \neq 0$, 有 $\boldsymbol{ax}_i = -\boldsymbol{ax}_j$ 当且仅当 $p_{ik} + p_{jk} = 1$.

证明 根据式 (2) 可得

$$p_{ik} = \frac{\mathrm{e}^{\boldsymbol{ax}_i}}{1+\mathrm{e}^{\boldsymbol{ax}_i}} \times \frac{\mathrm{e}^{\boldsymbol{ax}_k}}{1+\mathrm{e}^{\boldsymbol{ax}_k}} + \frac{1}{1+\mathrm{e}^{\boldsymbol{ax}_i}} \times \frac{1}{1+\mathrm{e}^{\boldsymbol{ax}_k}},$$

$$p_{jk} = \frac{\mathrm{e}^{\boldsymbol{ax}_j}}{1+\mathrm{e}^{\boldsymbol{ax}_j}} \times \frac{\mathrm{e}^{\boldsymbol{ax}_k}}{1+\mathrm{e}^{\boldsymbol{ax}_k}} + \frac{1}{1+\mathrm{e}^{\boldsymbol{ax}_j}} \times \frac{1}{1+\mathrm{e}^{\boldsymbol{ax}_k}},$$

则

$$\begin{aligned}
&p_{ik} + p_{jk} = 1\\
&\Leftrightarrow \frac{2+\mathrm{e}^{\boldsymbol{ax}_i}+\mathrm{e}^{\boldsymbol{ax}_j}+\mathrm{e}^{\boldsymbol{ax}_i+\boldsymbol{ax}_k}+\mathrm{e}^{\boldsymbol{ax}_j+\boldsymbol{ax}_k}+2\mathrm{e}^{\boldsymbol{ax}_i+\boldsymbol{ax}_j+\boldsymbol{ax}_k}}{(1+\mathrm{e}^{\boldsymbol{ax}_i})(1+\mathrm{e}^{\boldsymbol{ax}_j})(1+\mathrm{e}^{\boldsymbol{ax}_k})} = 1\\
&\Leftrightarrow \mathrm{e}^{\boldsymbol{ax}_i+\boldsymbol{ax}_j} = 1\\
&\Leftrightarrow \boldsymbol{ax}_i = -\boldsymbol{ax}_j.
\end{aligned}$$

综上, 性质 3 成立.

性质 3 表明如果两个样本到两个类簇的距离差互为相反数, 那么这两个样本与任意非 0 样本的共现概率的和为 1. 性质 3 意味着到两个类簇距离差的绝对值相同的样本应具有相同的稳定性取值.

性质4 对于 $\boldsymbol{ax}_k \neq 0$, 有 $|\boldsymbol{ax}_i| < |\boldsymbol{ax}_j|$ 当且仅当 $|p_{ik} - \frac{1}{2}| < |p_{jk} - \frac{1}{2}|$.

证明 首先讨论 $(\boldsymbol{ax}_i)(\boldsymbol{ax}_j) > 0$ 的情况. 此时, $\boldsymbol{ax}_i$, $\boldsymbol{ax}_j$ 和 $\boldsymbol{ax}_k$ 的取值关系可分为以下 4 种情况讨论:

(1) $\boldsymbol{ax}_i > 0$, $\boldsymbol{ax}_j > 0$, $\boldsymbol{ax}_k > 0$. 根据性质 2, 有

$$\begin{aligned}
&\left|p_{ik} - \frac{1}{2}\right| < \left|p_{jk} - \frac{1}{2}\right|\\
&\Leftrightarrow p_{ik} - p_{jk} < 0\\
&\Leftrightarrow \frac{(1-\mathrm{e}^{\boldsymbol{ax}_k})(\mathrm{e}^{\boldsymbol{ax}_j}-\mathrm{e}^{\boldsymbol{ax}_i})}{(1+\mathrm{e}^{\boldsymbol{ax}_i})(1+\mathrm{e}^{\boldsymbol{ax}_j})(1+\mathrm{e}^{\boldsymbol{ax}_k})} < 0\\
&\Leftrightarrow \mathrm{e}^{\boldsymbol{ax}_j} - \mathrm{e}^{\boldsymbol{ax}_i} > 0.
\end{aligned}$$

因为 $\boldsymbol{ax}_i > 0$, $\boldsymbol{ax}_j > 0$, 则 $\mathrm{e}^{\boldsymbol{ax}_j} - \mathrm{e}^{\boldsymbol{ax}_i} > 0 \Leftrightarrow |\boldsymbol{ax}_i| < |\boldsymbol{ax}_j|$.

(2) $\boldsymbol{ax}_i > 0$, $\boldsymbol{ax}_j > 0$, $\boldsymbol{ax}_k < 0$. 根据性质 2, 有

$$\left|p_{ik} - \frac{1}{2}\right| < \left|p_{jk} - \frac{1}{2}\right| \Leftrightarrow p_{jk} - p_{ik} < 0 \Leftrightarrow \mathrm{e}^{\boldsymbol{ax}_i} - \mathrm{e}^{\boldsymbol{ax}_j} < 0.$$

因为 $\boldsymbol{ax}_i > 0$, $\boldsymbol{ax}_j > 0$, 则 $\mathrm{e}^{\boldsymbol{ax}_i} - \mathrm{e}^{\boldsymbol{ax}_j} < 0 \Leftrightarrow |\boldsymbol{ax}_i| < |\boldsymbol{ax}_j|$.

(3) $\boldsymbol{ax}_i < 0$, $\boldsymbol{ax}_j < 0$, $\boldsymbol{ax}_k > 0$. 根据性质 2, 有

$$\left|p_{ik} - \frac{1}{2}\right| < \left|p_{jk} - \frac{1}{2}\right| \Leftrightarrow p_{jk} - p_{ik} < 0 \Leftrightarrow \mathrm{e}^{\boldsymbol{ax}_i} - \mathrm{e}^{\boldsymbol{ax}_j} > 0.$$

因为 $\boldsymbol{ax}_i < 0$, $\boldsymbol{ax}_j < 0$, 则 $\mathrm{e}^{\boldsymbol{ax}_i} - \mathrm{e}^{\boldsymbol{ax}_j} > 0 \Leftrightarrow |\boldsymbol{ax}_i| < |\boldsymbol{ax}_j|$.

(4) $\boldsymbol{a}\boldsymbol{x}_i < 0$, $\boldsymbol{a}\boldsymbol{x}_j < 0$, $\boldsymbol{a}\boldsymbol{x}_k < 0$. 根据性质 2, 有

$$\left|p_{ik} - \frac{1}{2}\right| < \left|p_{jk} - \frac{1}{2}\right| \Leftrightarrow p_{ik} - p_{jk} < 0 \Leftrightarrow \mathrm{e}^{\boldsymbol{a}\boldsymbol{x}_j} - \mathrm{e}^{\boldsymbol{a}\boldsymbol{x}_i} < 0.$$

因为 $\boldsymbol{a}\boldsymbol{x}_i < 0$, $\boldsymbol{a}\boldsymbol{x}_j < 0$, 则 $\mathrm{e}^{\boldsymbol{a}\boldsymbol{x}_j} - \mathrm{e}^{\boldsymbol{a}\boldsymbol{x}_i} < 0 \Leftrightarrow |\boldsymbol{a}\boldsymbol{x}_i| < |\boldsymbol{a}\boldsymbol{x}_j|$.

接下来讨论 $(\boldsymbol{a}\boldsymbol{x}_i)(\boldsymbol{a}\boldsymbol{x}_j) < 0$ 的情形. 令 $\boldsymbol{a}\boldsymbol{x}_{i'} = -(\boldsymbol{a}\boldsymbol{x}_i)$, 根据性质 3, 有 $|p_{ik} - \frac{1}{2}| = |p_{i'k} - \frac{1}{2}|$. 由于 $|\boldsymbol{a}\boldsymbol{x}_{i'}| = |\boldsymbol{a}\boldsymbol{x}_i|$, 等价于证明 $|\boldsymbol{a}\boldsymbol{x}_{i'}| < |\boldsymbol{a}\boldsymbol{x}_j| \Leftrightarrow |p_{i'k} - \frac{1}{2}| \leqslant |p_{jk} - \frac{1}{2}|$. 由于 $(\boldsymbol{a}\boldsymbol{x}_{i'})(\boldsymbol{a}\boldsymbol{x}_j) > 0$, 则该部分证明与第 1 种情况相同.

综上, 性质 4 成立.

性质 4 表明, 样本与两个类簇的距离差的绝对值越大, 则该样本与其他 $\boldsymbol{a}\boldsymbol{x} \neq 0$ 的样本出现在同类的概率越接近 $\frac{1}{2}$. 性质 4 意味着与两个类簇的距离差的绝对值越大的样本点应具有更高的稳定性取值.

接下来, 分析样本稳定性的合理性. 假设在高斯分布下的确定性函数为 fg, 则在服从高斯分布的均衡类簇假设下, 样本 $\boldsymbol{x}_i$ 的稳定性定义为

$$\mathrm{sg}(\boldsymbol{x}_i) = \frac{1}{n}\sum_{\boldsymbol{x}_j \in X} \mathrm{fg}(p_{ij}), \tag{3}$$

其中 $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$ 表示样本集合. 在高斯分布假设下, 显然此时样本间关系最不确定的共现概率值为 $t = \frac{1}{2}$. 稳定性函数 sg 具有以下性质.

性质5 如果 $|\boldsymbol{a}\boldsymbol{x}_i| < |\boldsymbol{a}\boldsymbol{x}_j|$, 则 $\mathrm{sg}(\boldsymbol{x}_i) < \mathrm{sg}(\boldsymbol{x}_j)$.

证明 根据式 (3), $\mathrm{sg}(\boldsymbol{x}_i) = \frac{1}{n}\sum_{\boldsymbol{x}_k \in \boldsymbol{x}} \mathrm{fg}(p_{ik})$, $\mathrm{sg}(\boldsymbol{x}_j) = \frac{1}{n}\sum_{\boldsymbol{x}_k \in \boldsymbol{x}} \mathrm{fg}(p_{jk})$.

根据定义 1, 当 $t = \frac{1}{2}$ 时, 条件 (2) 表明 fg 关于 $\frac{1}{2}$ 对称. 结合条件 (1) 易得 $|p_b - \frac{1}{2}| < |p_d - \frac{1}{2}| \Rightarrow \mathrm{fg}(p_b) < \mathrm{fg}(p_d)$.

当 $\boldsymbol{a}\boldsymbol{x}_k \neq 0$ 时, 根据性质 4, 有 $|\boldsymbol{a}\boldsymbol{x}_i| < |\boldsymbol{a}\boldsymbol{x}_j| \Rightarrow |p_{ik} - \frac{1}{2}| < |p_{jk} - \frac{1}{2}| \Rightarrow \mathrm{fg}(p_{ik}) < \mathrm{fg}(jk)$.

当 $\boldsymbol{a}\boldsymbol{x}_k = 0$ 时, 根据性质 2, 有 $p_{ik} = p_{jk} = \frac{1}{2}$, 此时 $\mathrm{fg}(p_{ik}) = \mathrm{fg}(p_{jk})$.

综上, 性质 5 成立.

性质 5 反应了 sg 的取值趋势. 从性质 5 可知, sg 在 $\boldsymbol{a}\boldsymbol{x}_i = 0$ 处取得最小值, 到两类簇距离差的绝对值大的样本将获得更高的 sg 值.

性质6 如果 $\boldsymbol{a}\boldsymbol{x}_i = -\boldsymbol{a}\boldsymbol{x}_j$, 则 $\mathrm{sg}(\boldsymbol{x}_i) = \mathrm{sg}(-\boldsymbol{x}_j)$.

证明 根据式 (3), 有

$$\mathrm{sg}(\boldsymbol{x}_i) = \frac{1}{n}\sum_{\boldsymbol{x}_k \in \boldsymbol{x}} \mathrm{fg}(p_{ik}).$$

根据性质 3 和 $\boldsymbol{a}\boldsymbol{x}_i = -\boldsymbol{a}\boldsymbol{x}_j$, 得

$$\begin{aligned}\mathrm{sg}(\boldsymbol{x}_j) &= \frac{1}{n}\sum_{\boldsymbol{x}_k \in \boldsymbol{x}} \mathrm{fg}(p_{jk}) \\ &= \frac{1}{n}\sum_{\boldsymbol{x}_k \in \boldsymbol{x}} \mathrm{fg}(1 - p_{ik}).\end{aligned}$$

根据定义 1, 有 $f(p_{ik}) = f(1 - p_{ik})$. 则 $\mathrm{sg}(\boldsymbol{x}_i) = \mathrm{sg}(-\boldsymbol{x}_j)$.

性质 6 说明到两类簇距离差的绝对值相同的样本具有相同的稳定性取值. 性质 5 和 6 说明在高斯假设下的样本稳定性取值趋势与性质 3 和 4 的预期趋势一致.

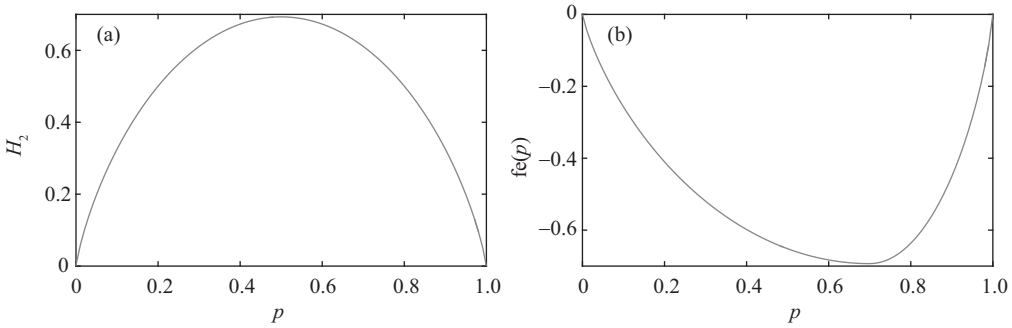


图 2 信息熵和 fe ($t = 0.7$) 的函数图像

Figure 2 The cuver of (a) H_2 and (b) fe ($t = 0.7$)

2.3 基于信息熵的确定性函数

在信息论中, Shannon 信息熵是度量信息不确定性程度的重要方法 [23]. Shannon 信息熵已经广泛应用于机器学习任务中, 如在 ID3 和 C4.5 等决策树算法中起重要作用 [24]. 若信息源有 n 个取值, 概率分别为 $\{p_1, p_2, \ldots, p_n\}$, Shannon 信息熵定义为

$$H = -\sum_{i=1}^{n} p_i \log p_i.$$

接下来, 借鉴信息熵的函数形式构造满足定义 1 的确定性函数. 对于共现与非共现的二元信源, 信息熵函数表达式为

$$H_2 = -(p \log p + (1-p)\log(1-p)),$$

其中 p 为共现概率. 图 2(a) 展示了 H_2 的函数图像. 从图 2(a) 可知, 函数 H_2 与确定性函数有相反的单调性. 此外, 函数 H_2 的极值点在 $t = \frac{1}{2}$ 处取得. 而在处理真实数据时, 不同数据最难以区分的共现概率值不一定在 $t = \frac{1}{2}$ 处取得. 最不确定的共现概率 t 应从两两样本共现概率分布中获得. 因此, 根据定义 1, 将基于二元信源的信息熵函数转换为确定性函数需要进行两个变换: (1) 改变单调性; (2) 改变极值点位置, 使极值点在 t 处取得, 其中 $t \in (0, 1)$. 变换后, 基于信息熵的确定性函数为

$$\mathrm{fe}(p) = \begin{cases} \left(\frac{p}{2t}\right)\log\left(\frac{p}{2t}\right) + \left(1-\frac{p}{2t}\right)\log\left(1-\frac{p}{2t}\right), & p < t, \\ \left(1-\frac{(1-p)}{2(1-t)}\right)\log\left(1-\frac{(1-p)}{2(1-t)}\right) + \left(\frac{(1-p)}{2(1-t)}\right)\log\left(\frac{(1-p)}{2(1-t)}\right), & p \geqslant t. \end{cases} \tag{4}$$

对于样本共现概率集合, 本文通过使用大津法 (Otsu) [25] 求解阈值 t. Otsu 算法是一种经典的图像二值化方法, 原理是寻找阈值将像素划分为两类使得类间方差最大. 因此, 该方法也可作为自适应的阈值确定方法. 令 $\boldsymbol{P} = \{p_{ij} | 1 \leqslant i \leqslant n, 1 \leqslant j \leqslant n\}$ 表示共现概率矩阵, 样本间关系最不确定的共现概率值为 $t = \mathrm{Otsu}(\boldsymbol{P})$.

图 2(b) 显示了 $t = 0.7$ 时 fe 的函数图像, 从图 2(b) 初步可知函数 fe 满足确定性函数的单调性要求. 接下来, 证明 fe 是确定性函数.

性质7 式 (4) 为确定性函数.

证明 对式 (4) 求导可得

$$\mathrm{fe}'(p)=\begin{cases}\frac{1}{2t}\log(\frac{p}{2t-p}), & p<t,\\ \frac{1}{2(1-t)}\log(\frac{1+p-2t}{1-p}), & p\geqslant t.\end{cases}$$

易知 $p<\ t$ 时, $\mathrm{fe}'(p)<0$; $p>t$ 时, $\mathrm{fe}'(p)>0$. 因此, 式 (4) 满足确定性函数的第 1 条约束.

由 $p_b<t<p_d$, $\frac{t-p_b}{p_d-t}=\frac{t}{1-t}$ 可推出 $\frac{p_b}{t}=\frac{1-p_d}{1-t}$, 带入式 (4) 可得 $\mathrm{fe}(p_b)=\mathrm{fe}(p_d)$. 因此, 式 (4) 满足确定性函数的第 2 条约束.

综上, 式 (4) 是确定性函数.

对于数据集 $X=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_n\}$ 中的样本 $\boldsymbol{x}_i$, 其基于信息熵的样本稳定性为

$$\mathrm{se}(\boldsymbol{x}_i)=\frac{1}{n}\sum_{j=1}^{n}\mathrm{fe}(p_{ij}),\tag{5}$$

其中 $p_{ij}\in\boldsymbol{P}$, $\boldsymbol{P}=\{p_{ij}|1\leqslant i\leqslant n,1\leqslant j\leqslant n\}$.

显然, 数据集中样本间关系的不确定性程度不同, 导致不同样本的稳定性大小不同, 对挖掘数据潜在团簇结构起到的作用也不同. 因此, 本文提出基于样本稳定性的聚类方法, 对稳定样本集与不稳定样本集采取针对性的处理策略. 具体算法将在第 3 节给出.

3 基于样本稳定性的聚类过程

基于样本稳定聚类方法的核心思想在于先将数据集划分为稳定样本集与不稳定样本集, 再采用针对性策略分别处理这两类样本. 由于稳定样本间关系明确, 其潜在的团簇结构较清晰, 容易获取较准确的聚类结果. 直接挖掘不稳定样本难度较大, 将不稳定样本划分至稳定样本集的类结构中相对较容易, 且有望提升不稳定样本类归属决策的性能. 具体地, 基于样本稳定性的聚类方法包括 3 个主要部分: (1) 将数据集划分为稳定样本集与不稳定样本集; (2) 挖掘稳定样本集的潜在团簇结构; (3) 将不稳定样本划分至该结构中.

首先, 本文借助第 2 节所提出的样本稳定性将数据集划分为稳定样本集与不稳定样本集. 由第 2 节可知, 样本稳定性的度量依赖于样本对的共现概率矩阵. 基于获得数据特征信息, 共同 K 近邻 (KNN) 是度量样本共现概率的一个简单直观的策略. 对于数据集 $X=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_n\}$, 样本 $\boldsymbol{x}_i$ 的 KNN 邻域集合为

$$\mathrm{KNN}(\boldsymbol{x}_i)=A,\quad |A|=K,\quad A\subseteq X,\quad \forall\boldsymbol{x}_j\subseteq A,\quad \forall\boldsymbol{x}_k\in X-A,\quad \mathrm{dis}(\boldsymbol{x}_i,\boldsymbol{x}_j)\leqslant\mathrm{dis}(\boldsymbol{x}_i,\boldsymbol{x}_k).$$

对于 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$, 基于共同 KNN 的共现概率为共同邻域样本所占的比例, 计算如下:

$$p_{ij}=\frac{|\{\boldsymbol{x}_p|\boldsymbol{x}_p\in\mathrm{KNN}(\boldsymbol{x}_i),\boldsymbol{x}_p\in\mathrm{KNN}(\boldsymbol{x}_j)\}|}{K},\tag{6}$$

其中邻域大小 K 取值过大或过小均不利于共现概率的刻画, 取值过大导致共现概率趋同, 取值过小则导致局部性问题. 在本文后续实验中, 邻域大小 K 设置为 $[n/(2k)]$, 其中 k 为数据集预期的类簇个数. 在数据平衡度未知的情况下, 如果将数据视为平衡数据处理, $[n/k]$ 为每类的样本数. 邻域大小 K 设置为 $[(n/k)/2]$, 一方面可以缓解数据不平衡带来的影响, 另一方面可以缓解 K 取值过大或过小对刻画共现概率的影响.

根据式 (5), 可得所有样本的稳定性 $\boldsymbol{S}=\{s_1,s_2,\ldots,s_n\}$. 通过阈值 t_s 将样本划分为稳定样本集 SS 和不稳定样本集 NS, 具体如下:

$$\mathrm{SS}=\{\boldsymbol{x}_i|s_i\geqslant t_s,1\leqslant i\leqslant n\};\tag{7}$$

$$\mathrm{NS}=\{\boldsymbol{x}_i|s_i<t_s,1\leqslant i\leqslant n\},\tag{8}$$

其中 t_s 通过 Otsu 算法求得, 即 $t_s=\mathrm{Otsu}(\boldsymbol{S})$.

将数据集划分为稳定样本集和不稳定样本集的过程见算法 1 中第 1~3 行.

然后, 挖掘稳定样本集的潜在团簇结构. 稳定样本集 SS 中样本间关系较确定, 共现概率的取值大都分布于两端, 即接近 0 或 1. 因此, 稳定样本集的潜在类簇结构相对较清晰, 可采用已有聚类分析算法挖掘其类簇结构. 去除不稳定样本会存在打散原有团簇的情况, 导致潜在类数目可能会多于预期, 因此采取可自动确定类个数的聚类算法. 此外, 在计算样本间的共现概率矩阵时已经获得了样本间的距离矩阵, 因此基于距离矩阵的聚类算法将降低该部分的计算时间消耗. 综上考虑, 在挖掘稳定样本集的潜在团簇结构时采用 AP 算法 [8]. 该部分过程见算法 1 中第 4 行.

接下来处理不稳定样本集. 对于较难判断类别归属的不稳定样本, 将其划分至稳定样本集的类结构中. 主要思想为将不稳定样本归属于最相似的类簇中. 样本 $\boldsymbol{x}_i$ 与类簇 c 的相似度计算如下:

$$\mathrm{sim}(\boldsymbol{x}_i,c)=\max_{\boldsymbol{x}_j\in c}(\mathrm{sim}(\boldsymbol{x}_i,\boldsymbol{x}_j)).$$

由于不稳定样本中存在与任何类簇都不相似的情况, 因此, 采用逐层处理不稳定样本的策略. 该策略优先划分类簇周围的不稳定样本, 并用确定类归属的样本扩充类簇, 重复执行这两个过程直到所有样本都确定类归属. 其中, 样本 $\boldsymbol{x}$ 与类簇的接近度为

$$\mathrm{ps}(\boldsymbol{x})=\max_{1\leqslant i\leqslant k}(\mathrm{sim}(\boldsymbol{x},c_i)).$$

需划分的样本集合为接近度大于阈值 t_{ps} 的样本:

$$\mathrm{NS}^{>}=\{x|\mathrm{ps}(\boldsymbol{x})>t_{\mathrm{ps}},\ \boldsymbol{x}\in\mathrm{NS}\},\tag{9}$$

其中 $t_{\mathrm{ps}}=\mathrm{Otsu}(\mathrm{PS})$, $\mathrm{PS}=\{\mathrm{ps}(\boldsymbol{x})|\boldsymbol{x}\in\mathrm{NS}\}$.

需划分的样本被分配到最近的类簇中并扩展该类簇大小. 类簇 c_i 的扩展如下计算:

$$c_i=c_i\cup\{\boldsymbol{x}|\mathrm{sim}(\boldsymbol{x},c_i)>\mathrm{sim}(\boldsymbol{x},c_j),\ j\neq i\}.\tag{10}$$

重复执行上述分配并扩展类簇的过程可将所有不稳定样本都分配到类簇中. 算法 1 中第 5~9 行显示了不稳定样本的分配过程.

如果获取的类簇个数多于预期, 采用层次聚类 (hierarchical clustering, HC) 逐层合并相似类簇以达期望. 类簇 c_i 和 c_j 的相似度为

$$\mathrm{sim}(c_i,c_j)=\frac{1}{|c_i||c_j|}\sum_{\boldsymbol{x}_i\in c_i}\sum_{\boldsymbol{x}_j\in c_i}\mathrm{sim}(\boldsymbol{x}_i,\boldsymbol{x}_j).\tag{11}$$

基于样本稳定性聚类方法的流程如算法 1 所示.

算法 1 基于样本稳定性的聚类方法

输入: 数据集 $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$, 预期类个数 k.

输出: 聚类结果 $C = \{c_1, c_2, \ldots, c_k\}$.

基于式 (6) 得到样本对的共现概率矩阵 $\boldsymbol{P} = \{p_{ij} | 1 \leqslant i \leqslant n, 1 \leqslant j \leqslant n\}$;
根据式 (5) 计算 n 个样本的稳定性 $\boldsymbol{S} = \{s_1, s_2, \ldots, s_n\}$;
根据式 (7) 和 (8) 得到稳定样本集 SS 和不稳定样本集 NS;
利用 AP 算法对稳定样本集 SS 聚类: $C_{ss} = \{c_1^{\rm ss}, c_2^{\rm ss}, \ldots, c_{k'}^{\rm ss}\} \leftarrow \mathrm{AP(SS)}$;
while $|\mathrm{NS}| \neq 0$ **do**
    根据式 (9) 得 $\mathrm{NS}^{>}$;
    更新稳定样本集和不稳定样本集: $\mathrm{SS} = \mathrm{SS} \cup \mathrm{NS}^{>}$, $\mathrm{NS} = \mathrm{NS} - \mathrm{NS}^{>}$;
    根据式 (10) 扩充 $C_{\rm ss}$ 中的每个类簇;
end while
if $k' > k$ **then**
    根据式 (11) 和层次聚类算法 HC 合并 $C_{\rm SS}$ 中相似类簇直至类个数为 k: $C \leftarrow \mathrm{HC}(C_{\rm SS}, k)$;
else
    $C \leftarrow C_{\rm SS}$.
end if

算法 1 主要消耗时间的部分包括: (1) 计算共现概率矩阵; (2) 利用 AP 算法对稳定样本集聚类; (3) 划分不稳定样本集; (4) 利用层次聚类合并类簇. 其中, 计算共现概率矩阵需要计算样本间距离矩阵, 每个样本的近邻, 以及样本间的共同邻域. 该部分的时间复杂度为 $O(2n^2 + n\log n)$. 利用 AP 算法对稳定样本集聚类的时间复杂度为 $O(|\mathrm{SS}|^2 \log |\mathrm{SS}|)$. 划分不稳定样本集的时间复杂度为 $O(T|\mathrm{SS}||\mathrm{NS}|)$, 其中 T 表示迭代的次数. 利用层次聚类合并类簇的时间复杂度为 $O(n^2)$. 需要注意的是样本间距离矩阵在计算共现概率矩阵时已经计算, 在对稳定样本集聚类、划分不稳定样本集、合并类簇时不需重复计算. 综上, 算法 1 的时间复杂度为 $O(2n^2 + n\log n + |\mathrm{SS}|^2 \log |\mathrm{SS}|)$.

4 实验分析

本文实验分析主要包括两方面: (1) 在人造数据集和图像分割数据集上直观展示基于信息熵的稳定性度量的合理性; (2) 在基准数据集上验证基于稳定性聚类方法的有效性.

4.1 合理性实验分析

本小节采用两组数据验证聚类中稳定性的合理性. 第 1 组为 4 个二维人造数据, 数据分布如图 3 第 1 行所示; 第 2 组为 4 幅来源于 BSD500 的图像分割数据, 图 4 第 1 行显示了其具体图像.

对于每个人造数据集, 基于式 (5) 计算每个样本的稳定性, 并根据式 (7) 和 (8) 分别得到稳定样本集 SS 和不稳定样本集 NS, 实验结果如图 3 第 2 行所示. 在图 3 第 2 行中, 不稳定样本集 NS 由灰色区域表示, 稳定样本集 SS 由其他颜色表示. 显然, 图 3 第 2 行显示, 在 4 组人造数据中, 根据基于信息熵的样本稳定性指标找到的不稳定样本主要处于不同类簇的交界区域. 在聚类任务中, 这类样本属于较难判断类别归属的样本点.

对于图像分割数据, 将每个像素点视为样本点, RGB 取值视为该样本点的特征值. 在该实验中, 根据式 (8) 得到每幅图像中的不稳定样本点构成不稳定区域, 并根据式 (7) 得到稳定样本点构成稳定区域. 对于稳定区域, 采用图像分割算法 Chen-Vese Method [26] 将其分割. 该部分实验结果如图 4 第 2 行所示, 图中灰色区域展示了不稳定区域, 稳定区域的分割结果由黑色区域与白色区域表示. 显然,

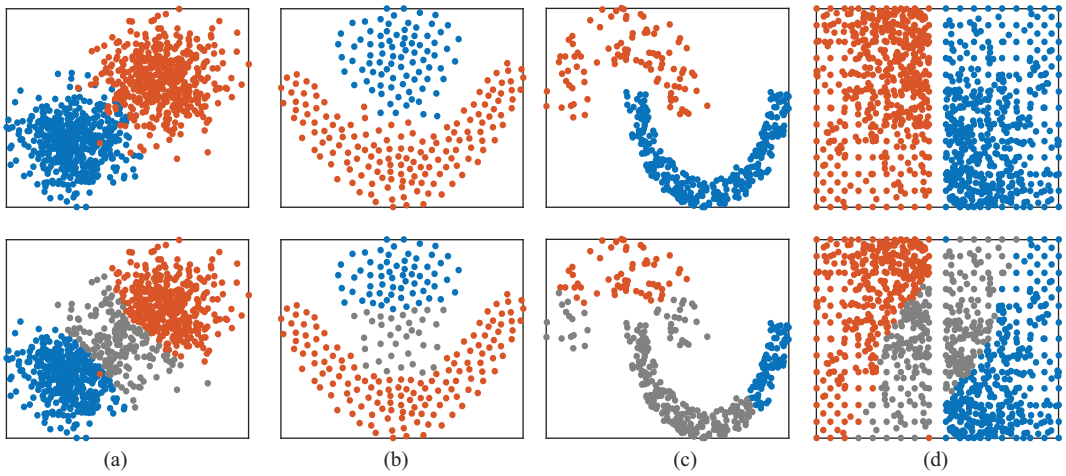


图 3 人造数据实验

Figure 3 Experiments on synthetic data sets. (a) 2d2k; (b) Flame; (c) Jain; (d) WingNut

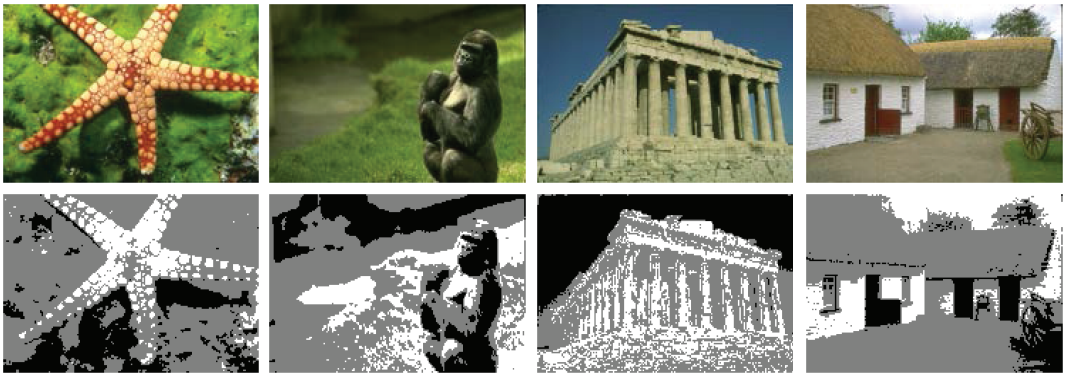


图 4 图像分割数据实验

Figure 4 Experiments on image segmentation data sets

如图 4 第 2 行所示, 图像中的不稳定区域主要为物体的边缘部分, 该部分在图像分割任务中属于难处理区域.

以上实验直观上显示了基于信息熵的样本稳定性度量可有效区分稳定样本集与不稳定样本集, 进而验证了本文所提样本稳定性度量的合理性.

4.2 有效性实验分析

为验证基于样本稳定性的聚类方法在处理聚类任务时的性能, 采用来源于 UCI 数据集的 8 组数据和来源于 CLUTO 的 4 组文本数据, 12 组数据的详细信息如表 1 所示. 为对照性地显示基于样本稳定性聚类算法 (sample's stability-based clustering, SSC) 的聚类性能, 8 个具有代表性的聚类算法被选为性能参照算法, 包括 K-means 算法、K-means* 算法 [27]、全局 K-means 算法 (global K-means, GK-means) [28]、AP 算法、自适应仿射传播聚类算法 (adaptive affinity propagation, AD-AP) [29]、DP 算法、改进的密度峰值聚类算法 (grid-division-based density peak clustering, GDPC) [30]、边界剥离聚类算法 (border-peeling clustering, BPC) [31]. 其中, 各类算法均采用固定聚类个数的版本. 为降低 K-means

表 1　基准数据集

Table 1　Benchmark data sets

Number	Data set	Number of samples	Number of attributes	Number of clusters
1	Iris	150	4	3
2	Wine recognition data	178	13	3
3	Seeds data set	210	7	3
4	Glass identification database	214	9	6
5	Ecoli	336	7	8
6	Cardiotocography data set	2126	40	10
7	Image segmentation data	2310	19	7
8	Waveform database generator	5000	21	3
9	tr45	690	8261	10
10	tr41	878	7454	10
11	wap	1560	8460	20
12	re1	1657	3758	25

表 2　C^b 和 C^d 的交叉表

Table 2　The cross tabulation of C^b and C^d

C^d	C^b				
	c_1^b	c_2^b	$\cdots$	c_k^b	SUM
c_1^d	n_{11}	n_{12}	$\cdots$	n_{1k}	$n_{1\cdot}$
c_2^d	n_{21}	n_{22}	$\cdots$	n_{2k}	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
c_k^d	n_{k1}	n_{k2}	$\cdots$	n_{kk}	$n_{k\cdot}$
SUM	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot k}$	n

算法和 K-means* 算法中随机性对性能评测的影响, 运行这两种方法 100 次并展示其平均性能. 本文采用余弦距离处理文本数据.

为评价聚类结果性能, 这里采用两个经典的外部评价指标: 标准化互信息 (normalized mutual information, NMI) [21] 和调整兰德指数 (adjusted rand index, ARI) [32]. 这两个评价指标通过度量聚类结果与参照结果的相似度来度量聚类性能, 其计算可基于两个聚类结果的交叉表. 聚类结果 $C^b = \{c_1^b, c_2^b, \ldots, c_k^b\}$ 和 $C^d = \{c_1^d, c_2^d, \ldots, c_k^d\}$ 的交叉表如表 2 所示.

基于表 2, NMI 计算为

$$\mathrm{NMI}(C^b, C^d) = \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} n_{ij}\log(\frac{nn_{ij}}{n_{i\cdot}n_{\cdot j}})}{\sqrt{(\sum_{i=1}^{k} n_{i\cdot}\log(\frac{n_{i\cdot}}{n}))(\sum_{j=1}^{k} n_{\cdot j}\log(\frac{n_{\cdot j}}{n}))}}.$$

基于表 2, ARI 计算为

$$\mathrm{ARI}(C^b, C^d) = \frac{r_0 - r_3}{\frac{1}{2}(r_1 + r_2) - r_3},$$

表 3 9 个算法的 NMI 值

Table 3 Index NMI from the nine clustering algorithms

Number	K-means	K-means*	GK-means	AP	AD-AP	DP	GDPC	BPC	SSC
1	0.7116±0.0622	0.7309 ±0.0142	0.7419	0.7777	0.7777	0.6586	0.6586	0.7630	**0.8031**
2	0.8414±0.0118	**0.8423±0.0184**	0.8347	0.7507	0.7315	0.5885	0.7415	0.7955	0.8364
3	0.6743±0.0000	0.6725±0.0037	0.6654	0.6873	0.6790	0.6797	0.5246	0.6959	**0.7043**
4	0.4107±0.0292	0.3409±0.0313	0.4008	0.3392	0.3065	0.3265	0.2757	0.3918	**0.4481**
5	0.5939±0.0283	0.5970±0.0185	0.6224	0.6311	0.5653	0.4985	0.4773	0.6333	**0.6768**
6	0.8972±0.0227	0.9177±0.0368	0.9478	0.9270	0.8118	0.8020	0.8199	**1.0000**	0.9797
7	0.6114±0.0159	0.6065±0.0112	0.6109	0.6116	0.6095	0.6599	0.6666	0.6664	**0.6683**
8	0.3642±0.0000	0.3642±0.0000	0.3642	0.2989	0.2760	0.2214	0.2128	0.3592	**0.3728**
9	0.4669±0.0406	0.2708±0.0244	0.4245	0.4062	0.3681	0.2673	0.3209	0.4491	**0.4752**
10	0.4829±0.0389	0.3160±0.0312	0.4292	0.4003	0.3917	0.3507	0.4268	0.4311	**0.5160**
11	**0.5058±0.0211**	0.4668±0.0163	0.4792	0.3603	0.3603	0.2728	0.2608	0.4523	0.4995
12	0.4360±0.0170	0.3031±0.0148	0.4194	0.4464	0.4461	0.3628	0.3524	0.4192	**0.4477**
Average rank	4.0000	5.9167	4.3333	4.6667	6.2500	7.5000	7.4167	3.5833	1.3333

其中

$$r_0=\sum_{i=1}^{K}\sum_{j=1}^{K}\binom{n_{ij}}{2},\quad r_1=\sum_{i=1}^{K}\binom{n_{i\cdot}}{2},\quad r_2=\sum_{j=1}^{K}\binom{n_{\cdot j}}{2},\quad r_3=\frac{2r_1r_2}{n(n-1)}.$$

在实验中, 数据的真实划分为评价指标的参照划分. 因此, 越高的指标值意味着越优的聚类性能. 9 个聚类算法在 12 组数据集上的聚类结果评价由表 3 和 4 表示. 其中表 3 展示了 NMI 评价指标值, 表 4 展示了 ARI 评价指标值. 在表 3 和 4 中, 每个数据的最优聚类结果的评价值由下划线加粗标识, 最后一行展示了每个算法的平均序值.

表 3 和 4 显示, SSC 的平均序值低于其他算法. 表 3 显示, SSC 在 NMI 评价指标下得到 12 组数据中 9 个数据的最优聚类结果. 表 4 显示, 在 ARI 评价指标下, SSC 取得 12 组数据中 8 个数据上的最优聚类结果. 为进一步分析表 3 和 4 结果, 采用 Nemenyi 后续检验. Nemenyi 检验比较两个算法的平均序值之差与临界阈值的大小来检验两个算法性能是否相同. 临界阈值的计算如下:

$$\mathrm{CD}=q_\alpha\sqrt{\frac{A(A+1)}{6D}},\tag{12}$$

其中 A 为算法个数, D 为数据集个数. 置信度为 95%, 算法个数为 9 时, $q_\alpha=3.102$. 根据式 (12) 可得, CD = 3.468. 图 5 显示了 Nemenyi 检验结果. 图中, 红色圆点表示了对应算法的平均序值, 蓝线为对应算法的临界值域, 黑色虚线为 SSC 算法的上临界值. 如图 5 所示, SSC 算法显著优于 K-means*, AP, AD-AP, DP, GDPC. 图 5 直观显示 SSC 取得了更靠前的平均序值.

以上实验结果验证了基于样本稳定性的聚类方法在数据聚类上的有效性.

5 结束语

聚类分析是处理无标记数据的重要技术. 数据的类型多样化、分布复杂化发展导致样本间关系的不确定性增强, 给有效挖掘数据的潜在类簇结构带来挑战. 为此, 本文将聚类集成中的样本稳定性概

表 4　9 个算法的 ARI 值

Table 4　Index ARI from the nine clustering algorithms

Number	K-means	K-means*	GK-means	AP	AD-AP	DP	GDPC	BPC	SSC
1	0.6589±0.1179	0.7102±0.0080	0.7163	0.7445	0.7445	0.4531	0.4531	0.7184	**0.7874**
2	0.8451±0.0140	0.8477±0.0196	0.8471	0.7262	0.7144	0.4956	0.7567	0.8025	**0.8516**
3	0.7049±0.0000	0.7026±0.0048	0.6934	0.7064	0.6952	0.7076	0.4726	0.7020	**0.7406**
4	**0.2572±0.0149**	0.1859±0.0278	0.2471	0.1862	0.1583	0.2267	0.2294	0.2397	0.2481
5	0.4238±0.0811	0.4180±0.0334	0.4150	0.4551	0.3412	0.3086	0.2655	0.6142	**0.7080**
6	0.7779±0.0654	0.8049±0.0905	0.8521	0.8558	0.6529	0.6323	0.6372	**1.0000**	0.9771
7	0.4711 ±0.0446	0.4736±0.0354	0.5034	0.5100	0.5117	0.5301	**0.5318**	0.5159	0.5260
8	0.2535±0.0000	0.2535±0.0000	0.2536	0.2167	0.1980	0.1897	0.1875	0.2542	**0.2592**
9	**0.3336±0.0627**	0.1855±0.0421	0.2512	0.2498	0.2217	0.0822	0.1275	0.2848	0.2354
10	0.3024±0.0691	0.2698±0.0305	0.2493	0.2296	0.2349	0.1388	0.2725	0.2695	**0.3066**
11	0.2025±0.0857	0.1408±0.0390	0.1792	0.1126	0.1126	0.1756	0.1964	0.1157	**0.2035**
12	0.2451±0.0181	0.1224±0.0147	0.1598	0.2066	0.2178	0.1303	0.1660	0.2427	**0.2709**
Average rank	3.9167	5.7500	4.9167	5.3333	6.4167	7.0000	6.2500	3.7500	1.6667

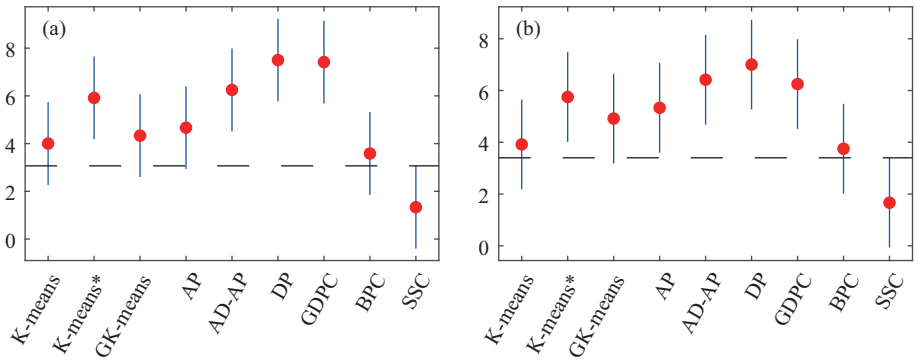


图 5　(网络版彩图) Nemenyi 后续检验

Figure 5　(Color online) Nemenyi post-hoc test based on (a) Table 3 and (b) Table 4

念扩展至聚类分析中, 既可用于度量样本间关系的确定度, 也可发现具有稳定关系的样本集. 本文从理论上给出了样本稳定性的合理性分析, 并提出了一个基于信息熵的样本稳定性度量函数. 在此基础上, 本文提出了一个基于样本稳定性的聚类方法, 该方法先挖掘稳定样本集的类簇结构并将不稳定样本集划分至该结构中. 二维人造数据和图像分割数据验证了样本稳定性度量的合理性, 12 组基准数据上的对比性实验显示了基于样本稳定性聚类方法相比其他 8 个代表性聚类算法的优越性. 样本的稳定性也可为样本对于聚类结果的贡献度提供权重度量方法. 此外值得关注的是如何降低计算样本关系矩阵的时间消耗.

参考文献

1　LeCun Y, Bengio Y, Hinton G. Deep learning. Nature, 2015, 521: 436–444
2　Jain A K, Murty M N, Flynn P J. Data clustering: a review. ACM Comput Surv, 1999, 31: 264–323
3　Zheng L, Yang Y, Tian Q. SIFT meets CNN: a decade survey of instance retrieval. IEEE Trans Pattern Anal Mach

Intell, 2018, 40: 1224–1244
4 Dhillon I S, Modha D S. Concept decompositions for large sparse text data using clustering. Mach Learn, 2001, 42: 143–175
5 Lu D, Tripodis Y, Gerstenfeld L C, et al. Clustering of temporal gene expression data with mixtures of mixed effects models with a penalized likelihood. Bioinformatics, 2018, 35: 778–786
6 Li Z C, Liu J, Yang Y, et al. Clustering-guided sparse structural learning for unsupervised feature selection. IEEE Trans Knowl Data Eng, 2014, 26: 2138–2150
7 Erhan D, Bengio Y, Courville A, et al. Why does unsupervised pre-training help deep learning? J Mach Learn Res, 2010, 11: 625–660
8 Frey B J, Dueck D. Clustering by passing messages between data points. Science, 2007, 315: 972–976
9 Rodriguez A, Laio A. Clustering by fast search and find of density peaks. Science, 2014, 344: 1492–1496
10 Otto C, Wang D, Jain A K. Clustering millions of faces by identity. IEEE Trans Pattern Anal Mach Intell, 2018, 40: 289–303
11 Zhang D, Wang F, Si L, et al. Maximum margin multiple instance clustering with applications to image and text clustering. IEEE Trans Neural Netw, 2011, 22: 739–751
12 Brockmeier A J, Mu T, Ananiadou S, et al. Self-tuned descriptive document clustering using a predictive network. IEEE Trans Knowl Data Eng, 2018, 30: 1929–1942
13 Qian Y H, Liang J Y, Pedrycz W, et al. Positive approximation: an accelerator for attribute reduction in rough set theory. Artif Intell, 2010, 174: 597–618
14 Qian Y H, Li F J, Liang J Y, et al. Space structure and clustering of categorical data. IEEE Trans Neural Netw Learn Syst, 2016, 27: 2047–2059
15 Fahy C, Yang S, Gongora M. Ant colony stream clustering: a fast density clustering algorithm for dynamic data streams. IEEE Trans Cybern, 2019, 49: 2215–2228
16 Blomstedt P, Tang J, Xiong J, et al. A Bayesian predictive model for clustering data of mixed discrete and continuous type. IEEE Trans Pattern Anal Mach Intell, 2015, 37: 489–498
17 Douzas G, Bacao F. Self-organizing map oversampling (SOMO) for imbalanced data set learning. Expert Syst Appl, 2017, 82: 40–52
18 Xu J, Han J, Nie F, et al. Re-weighted discriminatively embedded K-means for multi-view clustering. IEEE Trans Image Process, 2017, 26: 3016–3027
19 Yao X, Han J, Zhang D, et al. Revisiting co-saliency detection: a novel approach based on two-stage multi-view spectral rotation co-clustering. IEEE Trans Image Process, 2017, 26: 3196–3209
20 Li F J, Qian Y H, Wang J T, et al. Clustering ensemble based on sample's stability. Artif Intell, 2019, 273: 37–55
21 Strehl A L, Ghosh J. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. J Mach Learn Res, 2003, 3: 583–617
22 Li F J, Qian Y H, Wang J T, et al. Multigranulation information fusion: a Dempster-Shafer evidence theory-based clustering ensemble method. Inf Sci, 2017, 378: 389–409
23 Shannon C E. A mathematical theory of communication. Bell Syst Technical J, 1948, 27: 379–423
24 Hu Q H, Guo M Z, Yu D R, et al. Information entropy for ordinal classification. Sci China Inf Sci, 2010, 53: 1188–1200
25 Otsu N. A threshold selection method from gray-level histograms. IEEE Trans Syst Man Cybern, 1979, 9: 62–66
26 Chan T F, Vese L A. Active contours without edges. IEEE Trans Image Process, 2001, 10: 266–277
27 Malinen M I, Mariescu-Istodor R, Fränti P. K-means*: clustering by gradual data transformation. Pattern Recogn, 2014, 47: 3376–3386
28 Likas A, Vlassis N, Verbeek J J. The global k-means clustering algorithm. Pattern Recogn, 2003, 36: 451–461
29 Fan Z, Jiang J, Weng S, et al. Adaptive density distribution inspired affinity propagation clustering. Neural Comput Applic, 2019, 31: 435–445
30 Xu X, Ding S, Shi Z. An improved density peaks clustering algorithm with fast finding cluster centers. Knowledge-Based Syst, 2018, 158: 65–74
31 Averbuch-Elor H, Bar N, Cohen-Or D. Border-peeling clustering. IEEE Trans Pattern Anal Mach Intell, 2019. doi: 10.1109/TPAMI.2019.2924953
32 Hubert L, Arabie P. Comparing partitions. J Classif, 1985, 2: 193–218

Clustering method based on sample's stability

Feijiang LI[1], Yuhua QIAN[1,2*], Jieting WANG[1], Jiye LIANG[2] & Wenjian WANG[2]

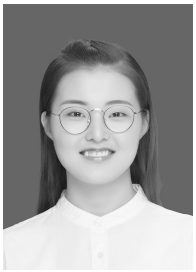
1. *Institute of Big Data Science and Industry, Shanxi University, Taiyuan* 030006*, China*;
2. *Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan* 030006*, China*
* Corresponding author. E-mail: jinchengqyh@126.com

Abstract The complexity of data types and distributions leads to an increase in uncertainty of the relationships between data samples, which bring challenges in discovering the cluster structure inherent in a data set. To address this challenge, this paper presents the concept of the sample's stability in a clustering ensemble, which is extended to the area of clustering analysis. We theoretically analyze the rationality of the sample's stability and propose an entropy-based sample's stability measure. Besides, we propose a clustering method based on the sample's stability. The proposed method divides a data set into stable and unstable samples, discovers the cluster structure of the stable samples, and assigns the unstable samples into this structure. The results of experiments on two-dimensional data sets and an image segmentation data set visually demonstrate the rationality of the sample's stability concept and effectiveness of the proposed clustering method based on the sample's stability measure.

Keywords machine learning, unsupervised learning, clustering analysis, sample's stability, stability theory

Feijiang LI was born in 1990. He received his B.S. degree from the School of Computer Science and Technology, Northeast University, China, in 2012. He is a Ph.D. candidate at the Institute of Big Data Science and Industry, Shanxi University. His research interests include machine learning and pattern recognition.

Yuhua QIAN was born in 1976. He received his M.S. and Ph.D. degrees in computers with applications at Shanxi University in 2005 and 2011, respectively. He is a professor at the Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, China. He is actively pursuing research in pattern recognition, feature selection, rough set theory, granular computing, and artificial intelligence. He is best known for his research in multigranulation rough sets in learning from categorical data and granular computing.



Jieting WANG was born in 1991. She received her B.S. and M.S. degrees from the School of Mathematical Sciences, Shanxi University, China, in 2013 and 2015, respectively. She is a Ph.D. candidate at the Institute of Big Data Science and Industry, Shanxi University. Her research interests include statistical machine learning and data mining.

Jiye LIANG was born in 1962. He received his M.S. and Ph.D. degrees from Xi'an Jiaotong University, Xi'an, China, in 1990 and 2001, respectively. He is a professor in the School of Computer and Information Technology and Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education at Shanxi University. His current research interests include computational intelligence, granular computing, data mining, and knowledge discovery.