一种具有最优保证特性的贝叶斯可能性聚类方法

刘解放*①② 王士同① 王 骏① 邓赵红①
①(江南大学数字媒体学院 无锡 214122)
②(湖北交通职业技术学院交通信息学院 武汉 430079)

摘 要：该文结合概率理论和可能性理论，提出一种具有最优保证特性的贝叶斯可能性聚类新方法。首先，将未知隶属度和聚类中心作为随机变量，为每个随机变量选择一个合适的概率分布，提出贝叶斯可能性聚类模型；在此基础上，基于贝叶斯推理和和蒙特卡洛采样方法，通过最大后验概率框架求解贝叶斯可能性聚类模型中的未知参数，从而提出一种具有最优保证特性的贝叶斯可能性聚类新方法。并对算法收敛性、算法复杂度等方面作了理论探讨。在合成数据集和真实数据集上的实验表明，所提算法扩展了传统可能性聚类性能，改进了聚类结果。

关键词：可能性聚类；贝叶斯推理；最大后验概率；蒙特卡洛方法
中图分类号： TP391 **文献标识码**: A **文章编号**：1009-5896(2017)07-1554-09
DOI: 10.11999/JEIT160908

Bayesian Possibilistic Clustering Method with Optimality Guarantees

LIU Jiefang①② WANG Shitong① WANG Jun① DENG Zhaohong①
①(*School of Digital Media*, *Jiangnan University*, *Wuxi* 214122, *China*)
②(*School of Traffic Information*, *Hubei Communications Technical College*, *Wuhan* 430079, *China*)

Abstract: A novel Bayesian possibilistic clustering method with optimality guarantees based on probability theory and possibilistic theory is proposed. First, the unknown membership degree and cluster center are represented as random variables. Given the specific constraints and uncertainty associated with each random variable, an appropriate probability distribution for each random variable is selected and the Bayesian possibilistic clustering model is proposed. On this basis, a novel Bayesian possibilistic clustering method with the optimal guarantee properties is propsed based on Bayesian theory and Monte Carlo sampling method using a Maximum-*A-Posteriori* (MAP) framework. Then, the convergence of the algorithm and the complexity of the algorithm are discussed. Experimental results on synthetic and real data sets show that the proposed method extends the traditional possibilistic clustering performance, and improves the clustering results.

Key words: Possibilistic clustering; Bayesian inference; Maximum-*A-Posteriori* (MAP); Monte Carlo method

1 引言

可能性聚类是在可能性理论框架下提出的一种实用性较广泛的聚类方法，它基于可能性理论框架，引入了可能性隶属度，从而放松了经典模糊聚类算法中样本在所有类中隶属度和必须为 1 的条件约束。可能性聚类继承了模糊聚类的实用性和灵活性，同时极大增强了处理带有噪声或异常点数据的聚类性能，目前已成为机器学习领域的研究热点。

研究人员针对可能性聚类模型与算法存在的问题主要从以下几个方面进行研究：(1)避免重合聚类的产生；(2)降低算法对初始化参数的依赖性；(3)利用可能性聚类进行无监督聚类分析；(4)基于核方法的可能性聚类。可能性聚类算法最突出的问题是易产生重合聚类[1]。针对该问题，Pal 等人[2,3]结合模糊方法和可能性方法，先后提出了 FPCM 和 PFCM 混合模型。模型中每个样本点对聚类原型同时产生隶属度和典型性，既解决了 FCM 对噪声敏感的问题，又避免了可能性聚类的重合聚类问题，但 FPCM 模型中约束样本点对聚类的可能性和为 1，使得在处理大数据集的时候产生不合适的典型值；之后提出的 PFCM 放松了对典型性值的约束，但目标函数中含有过多的参数，使得模型对参数的设置有一定的依赖性。同时，文献[4]和指出可能性聚类对初始化极其敏感，若要获得较好的聚类结果

收稿日期：2016-09-09；改回日期：2017-02-10；网络出版：2017-03-21
*通信作者：刘解放 ljf-it@163.com
基金项目：国家自然科学基金(61572236)，江苏省杰出青年基金(BK20140001)，江苏省自然科学基金(BK20151299)
Foundation Items: The National Natural Science Foundation of China (61572236), Jiangsu Province Outstanding Youth Fund (BK20140001), Natural Science Foundation of Jiangsu Province (BK20151299)

必须合理地选择初始化中心与参数 η_c。文献[5]将模糊方法用于可能性聚类得到改进的可能性聚类算法IPCM，用模糊隶属度确定聚类数目的同时通过可能性方法获得鲁棒性，并增强对初始化的不敏感性，能够在初始化不合理的情况下给出正确的聚类结果。文献[6]提出鲁棒性自动融合的可能性聚类算法AM-PCM，不但能够解决参数选择和初始化问题，还能自动确定聚类个数，但也存在着不能保证确定的聚类个数总是最优的问题。另外，关于(3)和(4)的相关研究具体可参考文献[7,8]。

近年来，概率理论大量应用于国民经济、工农业生产和各学科领域，成功地解决了机器学习领域中的一系列问题[9−11]。其中，在聚类领域研究中，众多学者已将概率方法和模糊方法进行有效的融合，并使得融合后的概率模糊聚类较之不管是概率聚类还是模糊聚类都有了突破性的进展。文献[12−14]分别提出了几种典型的概率模糊聚类，不但扩展了原有的模糊聚类性能，且改进其聚类结果。然而，关于可能性理论和概率理论的结合以及采用概率方法实现可能性聚类的研究还较为少见。

受上述思想的启发，本文融合两大理论提出了一种可能性聚类概率模型，也即贝叶斯可能性聚类(Bayesian Possibilistic Clustering, BPC)模型。在BPC 模型中，我们将未知参数可能性隶属度 $\boldsymbol{u}_n$ 和聚类中心 $\boldsymbol{y}_c$ 作为随机变量；考虑到具体的约束和关联每个随机变量的不确定性，我们为每个随机变量选择了一个合适的概率分布。然后，基于马尔科夫链蒙特卡洛(Markov Chain Monte Carlo, MCMC)采样方法和贝叶斯推理，通过最大后验概率(Maximum-*A*-*Posteriori*, MAP)框架求解BPC模型中聚类参数全局最优解，从而开发了一种新的具有最优保证特性的贝叶斯可能性聚类算法。与传统的可能性聚类方法相比，所提 BPC 方法具有以下优点：

(1)BPC 从概率的角度研究并实现可能性聚类。它融合了概率和可能性两大理论，汇聚了两者的优点，这是传统的可能性聚类所不具有的；

(2)由于 BPC 模型参数的求解过程使用了具有最优保证特性的 MCMC 采样方法，而不是定点迭代的策略来解决最优化问题，难免算法对初始化参数敏感，且容易陷入局部极值问题；从理论上讲，该方法可以获得全局最优解；

(3)BPC 能够重现和扩展传统可能性聚类性能，并改进聚类结果。也即，该方法进一步突破经典可能性聚类模糊指数 m 必须大于 1 的约束，使其不但可以小于等于 1，甚至可为负值，这是以前可能性聚类(如 PCM)无法实现的。

2　模糊聚类和可能性聚类原理

为了便于讨论相关模型及算法，首先对本文用到的一些数学符号进行统一的说明。

$\boldsymbol{X}=[\boldsymbol{x}_1\ \boldsymbol{x}_2\ \cdots\ \boldsymbol{x}_N]^{\mathrm{T}}$ 是 $N\times D$ 的样本集，其中 N 是样本个数，D 是样本维度；$\boldsymbol{Y}=[\boldsymbol{y}_1\ \boldsymbol{y}_2\ \cdots\ \boldsymbol{y}_C]^{\mathrm{T}}$ 是 $C\times D$ 的聚类中心矩阵，其中 C 是聚类个数；u_{nc} 是第 n 样本点 $\boldsymbol{x}_n$ 在第 c 个聚类中的隶属度，其中，$n=1,2,\cdots,N$，$c=1,2,\cdots,C$；m 是模糊指数，$\boldsymbol{I}$ 是 D 维单位矩阵，$\forall n,c,u_{nc}\in[0,1]$；$\forall c,0<\sum_{n=1}^{N}u_{nc}<N$；$\forall n,\max\limits_{c}u_{nc}>0$。

模糊聚类以 Zadeh 的模糊理论为基础，将样本对聚类的隶属度扩展到闭区间[0,1]，使得样本与所有聚类都建立起联系，聚类结果比硬聚类更易解释。但是，模糊聚类算法中必须满足任意样本对于所有类的隶属度和必须为 1，也即 $\forall n,\sum_{c=1}^{C}u_{nc}=1$。尽管该约束避免求解目标函数时出现平凡解问题，即所有隶属度都为 0 的情况，但由于模糊隶属度是相对数值，它不总是符合归属度或相容性的直观概念，从而导致该类算法无法适用于带有噪声或异常点的数据环境[15]，代表性的算法如 FCM[16]。

针对模糊聚类约束条件所引起的问题，Krishnapuram 等人[17]于 1993 年提出了可能性聚类；它的可能性隶属度是绝对数值，而非相对数值。它表示“代表性”或“相容性”，也即一个样本在某个类中的隶属度不依赖于其他类中的隶属度，经典的方法如 PCM[17]，它不但考虑一般划分聚类的标准——类内间距尽量小，类间间距尽量大；且强调隶属度值尽量大的原则，避免平凡解问题。其思想可由式(1)表示：

$$\left.\begin{aligned}&\min_{\boldsymbol{U},\boldsymbol{Y}}\ \sum_{n=1}^{N}\sum_{c=1}^{C}u_{nc}^{m}d(\boldsymbol{x}_n,\boldsymbol{y}_c)^2+\eta\sum_{n=1}^{N}\sum_{c=1}^{C}f(u_{nc})\\&\text{s.t}\quad \forall n,c,u_{nc}\in[0,1],\ n=1,2,\cdots,N\\&\qquad \forall c,0<\sum_{n=1}^{N}u_{nc}<N,\ c=1,2,\cdots,C\\&\qquad \forall n,\max_{c}u_{nc}>0\end{aligned}\right\}\tag{1}$$

其中，$d(\boldsymbol{x}_n,\boldsymbol{y}_c)$ 表示距离函数(如欧式距离)，类似FCM[16]的目标函数，主要实现上述所说的一般化聚类标准；$f(u_{nc})$ 表示 u_{nc} 的单调递减函数，主要实现上述强调的隶属度值要尽量大的原则；$\eta>0$ 表示对局部优化和全局优化进行权衡的权重因子。

PCM 是可能性聚类思想的经典实现，它的目标函数为

$$\min_{\boldsymbol{U},\boldsymbol{Y}}J=\sum_{n=1}^{N}\sum_{c=1}^{C}u_{nc}^{m}\left\|\boldsymbol{x}_n-\boldsymbol{y}_c\right\|^2+\sum_{n=1}^{N}\sum_{c=1}^{C}\eta_c\left(1-u_{nc}\right)^m\tag{2}$$

通过使用一个交替优化策略，可以解得隶属度和聚类中心。根据 KKT(Karush–Kuhn–Tucker)条件，可导出式(2)的解析解如式(3)和式(4)。

$$u_{nc}=\frac{1}{1+\left(\left\|\boldsymbol{x}_n-\boldsymbol{y}_c\right\|^2/\eta_c\right)^{1/(m-1)}} \tag{3}$$

$$\boldsymbol{y}_c=\frac{1}{\sum_{n=1}^{N}u_{uc}^m}\sum_{n=1}^{N}u_{uc}^m\boldsymbol{x}_n \tag{4}$$

其中，η_c 可通过样本数据集求得，具体可参考文献[17]。根据式(3)和式(4)，可以发现，因为式(2)中第二项没有变量 $\boldsymbol{y}_c$，因此所求参数 $\boldsymbol{y}_c$ 的迭代公式完全等同于 FCM 的推导，也即为样本加权平均值，而参数 u_{nc} 的迭代式仅与对应类中心的距离有关，而与其他类中心的距离无关，这也正符合可能性隶属度是绝对的而非相对的思想。

3 贝叶斯可能性聚类(BPC)

3.1 BPC 模型

根据第 2 节对可能性聚类原理的深入分析，本节给出 BPC 模型，它由 3 部分组成：可能性数据似然、可能性隶属度先验和聚类中心先验，分别定义如下：

定义 1 可能性数据似然(PDL)：

$$\begin{aligned}p(\boldsymbol{X}\mid\boldsymbol{U},\boldsymbol{Y})&=\prod_{n=1}^{N}\mathrm{PDL}(\boldsymbol{x}_n\mid\boldsymbol{u}_n,\boldsymbol{Y})\\&=\prod_{n=1}^{N}\frac{1}{Z(\boldsymbol{u}_n,m,\boldsymbol{Y})}\\&\quad\cdot\prod_{c=1}^{C}\mathcal{N}(\boldsymbol{x}_n\mid\boldsymbol{\mu}=\boldsymbol{y}_c,\boldsymbol{\Lambda}=u_{nc}^m\boldsymbol{I})\end{aligned} \tag{5}$$

可能性数据似然正比于 C 个正态似然的乘积，且每个正态分布具有不同精度。这可看作是 C 个正态分布同时生成的数据似然。针对每个样本，各正态分量具有不同的精度 u_{nc}^m，因此低精度将加大正态分量的偏差。该数据似然中，每个样本具有唯一的参数向量，因此每个样本可认为具有其自身的生成概率分布。然而，对于所有样本，通过分组，他们共享均值 $\boldsymbol{y}_c$。另外，$Z(\boldsymbol{u}_n,m,\boldsymbol{Y})$ 是归一化常数，因为多个高斯函数相乘所得的可能性数据似然已不再是标准正态分布。然而，在所研究的 MAP 推理算法中，该归一化常量被式(6)可能性隶属度先验消去，因此，它不必计算。

定义 2 可能性隶属度先验(PMP)：

$$\begin{aligned}p(\boldsymbol{U}\mid\boldsymbol{Y})&=\prod_{n=1}^{N}\mathrm{PMP}(\boldsymbol{u}_n\mid\boldsymbol{Y})\\&=\prod_{n=1}^{N}Z(\boldsymbol{u}_n,m,\boldsymbol{Y})\prod_{c=1}^{C}\exp\left\{-\frac{1}{2}\eta_c\left(1-u_{nc}\right)^m\right\}\end{aligned} \tag{6}$$

可能性隶属度先验包含两项。第 1 项是平衡因子，用于消去式(5)的归一化常量，而第 2 项是一个特殊的可能性隶属度分布，$\forall n,c,u_{nc}\in[0,1]$，$\max\limits_{c}u_{nc}>0$，$\eta_c>0$。

定义 3 聚类中心先验：

$$p(\boldsymbol{Y})=\prod_{c=1}^{C}\mathcal{N}\left(\boldsymbol{y}_c\mid\boldsymbol{\mu}_{\boldsymbol{y}},\boldsymbol{\Sigma}_{\boldsymbol{y}}\right) \tag{7}$$

BPC 模型假定聚类中心参数分布符合高斯分布 $\mathcal{N}\left(\boldsymbol{y}_c\mid\boldsymbol{\mu}_{\boldsymbol{y}},\boldsymbol{\Sigma}_{\boldsymbol{y}}\right)$。根据经验贝叶斯方法，可设置超参 $\boldsymbol{\mu}_{\boldsymbol{y}}$ 为所有样本的平均值：

$$\boldsymbol{\mu}_{\boldsymbol{y}}=\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}_n \tag{8}$$

超参 $\boldsymbol{\Sigma}_{\boldsymbol{y}}$ 为所有样本的协方差：

$$\boldsymbol{\Sigma}_{\boldsymbol{y}}=\frac{\gamma}{N}\sum_{n=1}^{N}\left(\boldsymbol{x}_n-\boldsymbol{\mu}_{\boldsymbol{y}}\right)\left(\boldsymbol{x}_n-\boldsymbol{\mu}_{\boldsymbol{y}}\right)^{\mathrm{T}} \tag{9}$$

其中，γ 为用户预设定参数，它影响高斯分布的密度。大量实验表明 $\gamma=3$ 较为合适。若考虑整个 BPC 模型的对数似然作为一个可能性聚类的目标函数，聚类中心先验则充当一个正则化项。

通过将式(5)、式(6)和式(7)相乘，可得数据 $\boldsymbol{X}$ 和参数 $\boldsymbol{U},\boldsymbol{Y}$ 的联合似然如式(10)：

$$\begin{aligned}p(\boldsymbol{X},\boldsymbol{U},\boldsymbol{Y})&=p(\boldsymbol{X}\mid\boldsymbol{U},\boldsymbol{Y})p(\boldsymbol{U}\mid\boldsymbol{Y})p(\boldsymbol{Y})\\&\propto\exp\left\{-\frac{1}{2}\sum_{n=1}^{N}\sum_{c=1}^{C}u_{nc}^m\left\|\boldsymbol{x}_n-\boldsymbol{y}_c\right\|^2\right\}\\&\quad\cdot\prod_{n=1}^{N}\prod_{c=1}^{C}\exp\left\{-\frac{1}{2}\eta_c\left(1-u_{nc}\right)^m\right\}\\&\quad\cdot\exp\left\{-\frac{1}{2}\sum_{c=1}^{C}\left(\boldsymbol{y}_c-\boldsymbol{\mu}_{\boldsymbol{y}}\right)^{\mathrm{T}}\boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1}\left(\boldsymbol{y}_c-\boldsymbol{\mu}_{\boldsymbol{y}}\right)\right\}\end{aligned} \tag{10}$$

它正比于参数的后验分布 $p(\boldsymbol{U},\boldsymbol{Y}\mid\boldsymbol{X})\propto p(\boldsymbol{X},\boldsymbol{U},\boldsymbol{Y})$。联合似然的目标函数形式是它自身的负对数，为了简化，乘上因子 2，得目标函数如式(11)：

$$\begin{aligned}J(\boldsymbol{X},\boldsymbol{U},\boldsymbol{Y})=&\sum_{n=1}^{N}\sum_{c=1}^{C}u_{nc}^m\left\|\boldsymbol{x}_n-\boldsymbol{y}_c\right\|^2+\sum_{n=1}^{N}\sum_{c=1}^{C}\eta_c\left(1-u_{nc}\right)^m\\&+\sum_{c=1}^{C}\left(\boldsymbol{y}_c-\boldsymbol{\mu}_{\boldsymbol{y}}\right)^{\mathrm{T}}\boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1}(\boldsymbol{y}_c-\boldsymbol{\mu}_{\boldsymbol{y}})\end{aligned} \tag{11}$$

3.2 BPC 算法

我们的主要任务是找出满足 BPC 模型的最大后验概率所需参数。这等同于找到可能性聚类目标函数中参数的全局最优值。为了完成 MAP 推理，我们采用了 MCMC 采样方法[18]，它具有最优保证特性。使用 Metropolis-Hastings[19]采样方法生成 BPC 模型的后验分布样本。样本生成后，通过后验概率评估，保留最好的样本。表 1 的算法 1 给出了贝叶斯可能性聚类算法的具体过程。

表 1　BPC 算法

算法 1　贝叶斯可能性聚类(BPC)算法

输入：样本 $\boldsymbol{X}$，模糊指数 m，聚类个数 C，迭代次数 N_{iter}

输出：隶属度 $\boldsymbol{U}^*$ 和聚类中心 $\boldsymbol{Y}^*$

采用式(8)，式(9)初始化参数 $\boldsymbol{\mu}_y$，$\boldsymbol{\Sigma}_y$

初始化 $\boldsymbol{u}_n \sim \text{Uniform}(0,1), \forall n$

初始化 $\boldsymbol{y}_c \sim \mathcal{N}(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y), \forall c$

$\boldsymbol{u}_n^* = \boldsymbol{u}_n$，$\boldsymbol{y}_c^* = \boldsymbol{y}_c$ //把当前样本赋给 MAP 样本

for iter=1,2,…,N_{iter}

　//采样 $\boldsymbol{U} \sim p(\boldsymbol{U} \mid \boldsymbol{X}, \boldsymbol{Y}) \propto p(\boldsymbol{X}, \boldsymbol{U}, \boldsymbol{Y})$

　for n=1,2,…,N

　　采用式(12)生成新的建议隶属度样本 $\boldsymbol{u}_n^+$

　　采用式(14)，以概率 $R_{\boldsymbol{u}}$ 接受 $\boldsymbol{u}_n = \boldsymbol{u}_n^+$

　　if $p(\boldsymbol{x}_n, \boldsymbol{u}_n^+ \mid \boldsymbol{Y}^*) > p(\boldsymbol{x}_n, \boldsymbol{u}_n^* \mid \boldsymbol{Y}^*)$ //依据式(13)

　　$\boldsymbol{u}_n^* = \boldsymbol{u}_n^+$

　　endif

　endfor

　//采样 $\boldsymbol{Y} \sim p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{U}) \propto p(\boldsymbol{X}, \boldsymbol{U}, \boldsymbol{Y})$

　for $c = 1,2,\cdots,C$

　　采用式(15)生成新的建议聚类中心样本 $\boldsymbol{y}_c^+$

　　采用式(17)，以概率 $R_{\boldsymbol{y}}$ 接受 $\boldsymbol{y}_c = \boldsymbol{y}_c^+$

　　if $p(\boldsymbol{X}, \boldsymbol{y}_c^+ \mid \boldsymbol{U}^*) > p(\boldsymbol{X}, \boldsymbol{y}_c^* \mid \boldsymbol{U}^*)$ //依据式(16)

　　$\boldsymbol{y}_c^* = \boldsymbol{y}_c^+$

　　endif

　endfor

　//检查整个样本的最大似然

　if $p(\boldsymbol{X}, \boldsymbol{U}, \boldsymbol{Y}) > p(\boldsymbol{X}, \boldsymbol{U}^*, \boldsymbol{Y}^*)$ //依据式(10)

　　$\boldsymbol{U}^* = \boldsymbol{U}$，$\boldsymbol{Y}^* = \boldsymbol{Y}$

　endif

endfor

给定样本和聚类中心，使用 Metropolis-Hastings 采样方法对隶属度条件分布 $p(\boldsymbol{U} \mid \boldsymbol{X}, \boldsymbol{Y})$ 进行采样。为了简化，使用式(12)的均匀分布作为可能性隶属度建议分布。

$$\boldsymbol{u}_n^+ \sim \text{Uniform}(0,1), \forall n \tag{12}$$

当给定聚类中心，隶属度条件分布 $p(\boldsymbol{U} \mid \boldsymbol{X}, \boldsymbol{Y})$ 正比于联合分布 $p(\boldsymbol{X}, \boldsymbol{U}, \boldsymbol{Y})$，另外，对于任何新生成的建议隶属度 $\boldsymbol{u}_n^+$，其他相关隶属度和聚类中心不变。因此，只需估计式(13)所示概率。

$$\begin{aligned} p(\boldsymbol{x}_n, \boldsymbol{u}_n \mid \boldsymbol{Y}) &= p(\boldsymbol{x}_n \mid \boldsymbol{u}_n, \boldsymbol{Y})\, p(\boldsymbol{u}_n \mid \boldsymbol{Y}) \\ &\propto \prod_{c=1}^{C} \exp\left\{-\frac{1}{2} u_{nc}^m \left\|\boldsymbol{x}_n - \boldsymbol{y}_c\right\|^2\right\} \\ &\quad \cdot \exp\left\{-\frac{1}{2}\eta_c \left(1 - u_{nc}\right)^m\right\} \end{aligned} \tag{13}$$

新生成的建议隶属度样本 $\boldsymbol{u}_n^+$ 以概率 $R_{\boldsymbol{u}}$ 接受并替代 $\boldsymbol{u}_n$。

$$R_{\boldsymbol{u}} = \min\left\{1, \frac{p(\boldsymbol{x}_n, \boldsymbol{u}_n^+ \mid \boldsymbol{Y})}{p(\boldsymbol{x}_n, \boldsymbol{u}_n \mid \boldsymbol{Y})}\right\} \tag{14}$$

因为建议分布不依赖于当前生成的建议样本，所以不必进行 Hastings 校正。

给定样本和隶属度，采用 Metropolis-Hastings 采样方法对聚类中心条件分布 $p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{U})$ 进行采样。当给定隶属度时，该条件分布正比于联合分布 $p(\boldsymbol{X}, \boldsymbol{U}, \boldsymbol{Y})$。聚类中心建议分布来自于高斯分布，它以马尔可夫链当前状态为中心，较小值为方差，如式(15)：

$$\boldsymbol{y}_c^+ \sim \mathcal{N}\left(\boldsymbol{y}_c, \frac{1}{\delta}\boldsymbol{\Sigma}_y\right) \tag{15}$$

其中，δ 是用户预设定参数，用来控制生成建议聚类中心的紧密程度，它以当前状态为中心；它的大小关系到样本的接受率。在应用中，我们设定 $\delta = 10$。对于任何新生成的建议聚类中心 $\boldsymbol{y}_c^+$，它独立于其他聚类中心和隶属度；因此，只需估计式(16)所示概率：

$$\begin{aligned} p(\boldsymbol{X}, \boldsymbol{y}_c \mid \boldsymbol{U}) &= p(\boldsymbol{X} \mid \boldsymbol{U}, \boldsymbol{y}_c)\, p(\boldsymbol{y}_c) \\ &\propto \exp\left\{-\frac{1}{2}\sum_{n=1}^{N} u_{nc}^m \left\|\boldsymbol{x}_n - \boldsymbol{y}_c\right\|^2\right\} \\ &\quad \cdot \exp\left\{-\frac{1}{2}(\boldsymbol{y}_c - \boldsymbol{\mu}_y)^{\mathrm{T}} \boldsymbol{\Sigma}_y^{-1} (\boldsymbol{y}_c - \boldsymbol{\mu}_y)\right\} \end{aligned} \tag{16}$$

新生成的建议聚类中心样本 $\boldsymbol{y}_c^+$，以概率 $R_{\boldsymbol{y}}$ 接受并替代 $\boldsymbol{y}_c$。

$$R_{\boldsymbol{y}} = \min\left\{1, \frac{p(\boldsymbol{X}, \boldsymbol{y}_c^+ \mid \boldsymbol{U})}{p(\boldsymbol{X}, \boldsymbol{y}_c \mid \boldsymbol{U})}\right\} \tag{17}$$

因为高斯建议分布是对称的，所以也不必进行 Hastings 校正。

在BPC算法中，第1~4步根据经验贝叶斯方法和参数建议分布初始化MAP样本 $\{\boldsymbol{U}^*, \boldsymbol{Y}^*\}$；第6~12步求解隶属度参数，其中第7步根据式(12)均匀建议分布生成新的建议隶属度样本 $\boldsymbol{u}_n^+$，第8步根据式(14)的概率接受 $\boldsymbol{u}_n = \boldsymbol{u}_n^+$，第9~11步根据式(13)比较每一个建议隶属度向量 $\boldsymbol{u}_n^+$，如果 $\boldsymbol{u}_n^+$ 增大了当前MAP样本的似然，它将成为新的 $\boldsymbol{u}_n^*$；第13~19步求解聚类中心参数，其中第14步根据高斯建议分布式(15)生成新建议样本 $\boldsymbol{y}_c^+$；第15步根据式(17)的概率接受 $\boldsymbol{y}_c = \boldsymbol{y}_c^+$；第16~18步根据式(16)比较每一个建议聚类中心向量 $\boldsymbol{y}_c^+$，如果 $\boldsymbol{y}_c^+$ 增大了当前MAP样本的似然，它将成为新的 $\boldsymbol{y}_c^*$；第20~22步根据式(10)对比当前概率样本 $\{\boldsymbol{U}, \boldsymbol{Y}\}$ 和当前MAP样本 $\{\boldsymbol{U}^*, \boldsymbol{Y}^*\}$ 的似然，如果概率样本具有更高的联合似然，保留作为新的MAP样本。我们把以概率接受的样本与MAP样本的比较看作是“加速搜索”，它为MAP样本的改进提供了更多可能性。

4 BPC 相关讨论

4.1 收敛性分析

定理 1 BPC 算法一致收敛。

证明 文献[20–22]已经证明了 MCMC 方法收敛，其中包括 Metropolis-Hastings 和 Gibbs 采样方法，且 MCMC 方法具有最优保证特性，即随着迭代次数的增加，MCMC 方法能保证收敛到全局最优解。而本文提出的 BPC 算法正是基于 MCMC 框架下开发的，因此，定理 1 成立。 证毕

根据上述分析可知本文算法一致收敛，且具有 MCMC 最优保证特性。第 5 节的实验也验证了它的一致收敛性。但是，收敛速度及结果是不可预测的，它依赖于数据的结构和所寻参数个数，及模型和建议分布中参数设置。关于更多 MCMC 方法收敛速度的相关讨论，可以参考文献[23]。

4.2 复杂度分析

该算法的时间复杂度包括两部分：分别为搜索 MAP 样本 $\boldsymbol{U}^*$ 和 $\boldsymbol{Y}^*$ 所耗时间，因此单次迭代的渐进计算时间复杂度为 $O(NCD+CD^2)$，其中 N 是样本个数，C 是聚类个数，D 是数据维数。其中 D^2 由式(9)中的协方差矩阵产生。实际应用中，一般为对角协方差矩阵，因此，单次迭代($N_{\text{iter}}=1$)的时间复杂度降为 $O(NCD)$ 。

4.3 参数分析

BPC 算法包含一些参数，他们必须在运行前设置。下面给出了各相关参数的讨论及设置说明。

模糊指数 m：参数 m 等同于经典 FCM 算法中模糊指数的作用，因此 m 接近于 1，隶属度变得更硬。然而，因为本文算法没有使用闭合解形式，所以可自由设置模糊指数 m 为 1，获得硬聚类；甚至可以小于 1 或负值。尽管模糊指数的设置依赖于特定数据集及应用，然而大量的实验表明对于 BPC 算法 $m=1.2$ 是一个合适的选择。

聚类个数 C：聚类个数是 BPC 算法中关键参数。正如其他大部分聚类算法一样，用户必须预先设定。该参数也依赖于特定数据集及应用。

聚类中心先验方差参数 γ：参数 γ 控制聚类中心高斯先验分布的协方差大小，如式(9)。较小值(如小于 1)将致使中心点接近数据集均值，而较大值将减少先验对结果的影响。实验中，设为 3。

5 实验分析

5.1 实验设置

为了验证所提BPC算法的有效性，本节采用合成数据集和真实数据集对其进行实验，主要从以下两个方面进行研究：(1)与经典FCM[16]算法比较，观察BPC算法在处理带有噪声或异常点数据上的性能；(2)与经典PCM[17]算法相比，观察BPC算法是否扩展了性能、改进了结果。实验所得结果均采用五重交叉验证获得平均值。实验平台：Intel i7-4770 4核CPU, 8 GB内存 Windows 7操作系统，算法采用Matlab2010a编写。

实验中，经典PCM算法和本文BPC算法的参数 η_c 通过式(18)进行设置。

$$\eta_c=\sum_{n=1}^{N}u_{uc}^{m}\left\|\boldsymbol{x}_n-\boldsymbol{y}_c\right\|^2\Bigg/\left(\sum_{n=1}^{N}u_{uc}^{m}\right) \tag{18}$$

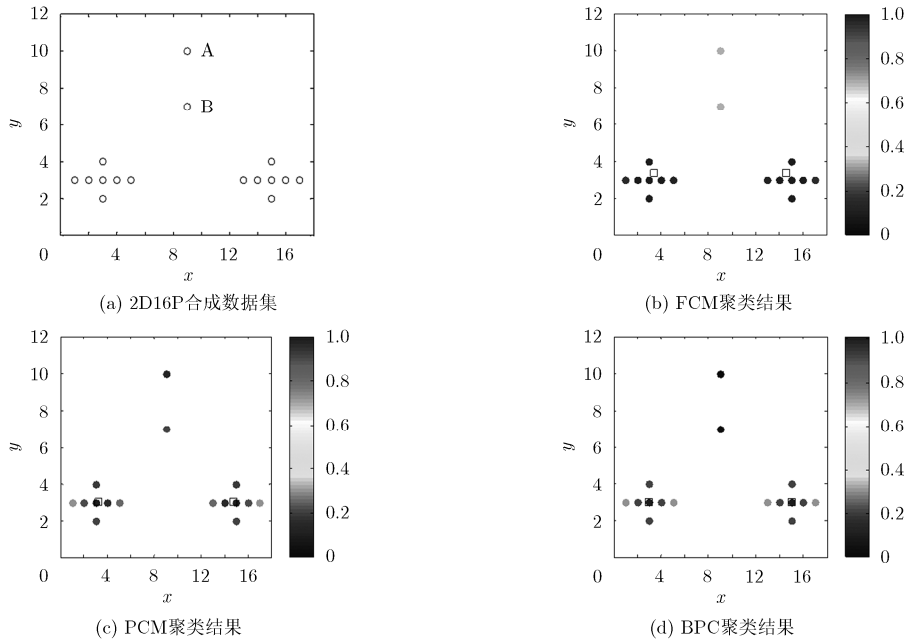
所提BPC算法中其他参数均根据4.3节进行设置。

5.2 合成数据集

该合成数据集含有 2 类，每类各有 7 个样本点，另加 2 个噪声点/异常点 A, B 组成共有 16 个样本的数据集，其分布如图 1(a)所示，x 和 y 分别表示样本点的横坐标和纵坐标。为了更好地描述和讨论，我们命名该合成数据集为 2D16P。

该实验将 BPC 算法分别与经典 FCM[16]和 PCM[17]算法进行比较。实验结果如图 1 和表 1 所示，其中图 1(b)~1(d)中小正方形代表聚类中心。从实验结果中可以发现：(1)经典 FCM 由于存在隶属度约束条件 $\forall n,\sum_{c=1}^{C}u_{nc}=1$，致使每类中几乎所有样本的隶属度值都接近于 1，最小者也为 0.5。在这种情况下，对于两个噪声点 A 和 B，尽管根据样本分布可以很直观地发现他们应该具有不等且相对较小的隶属度，但是 FCM 计算的结果都为 0.5，且聚类中心的偏差较大。(2)经典的 PCM 算法由于放松了隶属度约束条件 $\forall n,\sum_{c=1}^{C}u_{nc}=1$，每类各样本点的隶属度仅与该类的中心的距离有关，它能够使得距离中心较远的的噪声点 A 和 B 获得不等且相对较小的隶属度，分别为 0.0014 和 0.0180，且聚类中心的偏差较小。显然，PCM 受噪声的影响要好于 FCM。(3) 对比经典的 PCM 算法，本文 BPC 算法能够使得距离聚类中心较近的样本的隶属度尽量大，而较远的样本的隶属度尽量小，这使得 BPC 获得样本隶属度具有更好的可解释性；从聚类中心的位置可以发现，它几乎不受噪声的影响。

表 2 量化了 FCM, PCM 和 BPC 3 种算法的聚类中心和隶属度。可以发现 FCM 受噪声的影响最大，中心点较大程度地偏向噪声点 A 和 B; PCM 次之，而本文 BPC 算法由于采用了概率和可能性理论相结合策略及全局最优保证特性，它几乎不受噪声的影响；另一个细节的发现是它能够使得(1,3)和(5,3), (3,2)和(3,4), (2,3)和(4,3)等对称点获得几乎相等的隶属度。总之，对于样本归属度问题，本文所提的 BPC 较之 FCM 和 PCM 有更合理的解释，能够给我们更多的关于样本数据的位置信息，具有更强的抗噪性能。



(a) 2D16P合成数据集　(b) FCM聚类结果

(c) PCM聚类结果　(d) BPC聚类结果

图 1　2D16P 合成数据集及各算法的聚类结果

表 2　3 种算法在 2D16P 数据集上的聚类结果

(x,y)	FCM		PCM		BPC	
	C1	C2	C1	C2	C1	C2
(1,3)	**0.9686**	0.0314	**0.7188**	0.0640	**0.7339**	0
(2,3)	**0.9866**	0.0134	**0.8922**	0.0737	**0.9255**	0
(3,3)	**0.9976**	0.0024	**0.9950**	0.0857	**1.0000**	0
(4,3)	**0.9957**	0.0043	**0.9579**	0.1006	**0.9257**	0.0001
(5,3)	**0.9720**	0.0280	**0.8082**	0.1198	**0.7342**	0.0004
(3,4)	**0.9959**	0.0041	**0.9318**	0.0852	**0.9256**	0
(3,2)	**0.9850**	0.0150	**0.9162**	0.0850	**0.9256**	0
(13,3)	0.0280	0.9720	0.1198	**0.8082**	0.0004	**0.7342**
(14,3)	0.0043	**0.9957**	0.1006	**0.9579**	0.0001	**0.9257**
(15,3)	0.0024	**0.9976**	0.0857	**0.9950**	0	**1.0000**
(16,3)	0.0134	**0.9866**	0.0737	**0.8922**	0	**0.9255**
(17,3)	0.0314	**0.9686**	0.0640	**0.7188**	0	**0.7339**
(15,4)	0.0041	**0.9959**	0.0852	**0.9318**	0	**0.9256**
(15,2)	0.0150	**0.9850**	0.0850	**0.9162**	0	**0.9256**
B (9,7)	0.5000	0.5000	0.2102	0.2102	0.0180	0.0180
A (9,10)	0.5000	0.5000	0.1373	0.1373	0.0014	0.0014
centers	(3.4186,3.3793)	(14.5814,3.3793)	(3.2487,3.0592)	(14.7513,3.0592)	(3.0005,3.0003)	(14.9995,3.0003)

5.3 UCI 数据集

Iris plants 数据集[24]在模式识别理论实践验证中是一个非常著名且广泛被使用的数据集。它包含 3 类，每类 50 个样本点，共 150 个样本点，每个样本具有 4 个属性。每类代表一种鸢尾属植物，其中一类和另外两类线性可分，而另外两类线性不可分。

为了便于观察，图 2 给出了样本的第 2 和第 4 个属性的分布图。

该实验结果分别展示在图 2 和表 2 中，为便于讨论，我们假定第 1~50 个样本为第 1 类 C1，第 51~100 个样本为第 2 类 C2，剩余的样本为第 3 类 C3。首先，从图 2(b)~2(d)中可以有如下发现：(1) 图 2(b)表明 FCM 能够很好地划分出第 1 类 C1，而 C2 和 C3 的样本隶属度均有交叉，说明两类中均有错误划分的样本。(2)图 2(c)表明 PCM 也能较好的划分出 C1，但是 C2 和 C3 中同样有大量样本的隶属度交叉较为严重，略逊色于 FCM。(3)图 2(d)表明每个类的样本仅在该类中的隶属度值较大，而在另外两类中的隶属度值几乎接近于 0，仅有 C3 中少数几个样本存在隶属度交叉，这证明 BPC 几乎完美地划分了 Iris plants 数据集。

表 3 给出了 3 种算法的量化聚类结果，它由 5 种常用的聚类性能指标 Accuracy, RI, NMI, Pure 和 F-measure 组成。表 2 的结果也进一步验证了上述结论。该实验同样表明可能性聚类引入概率理论后改进了聚类结果。

表 3　3 种算法在 Iris plants 数据集上的聚类结果

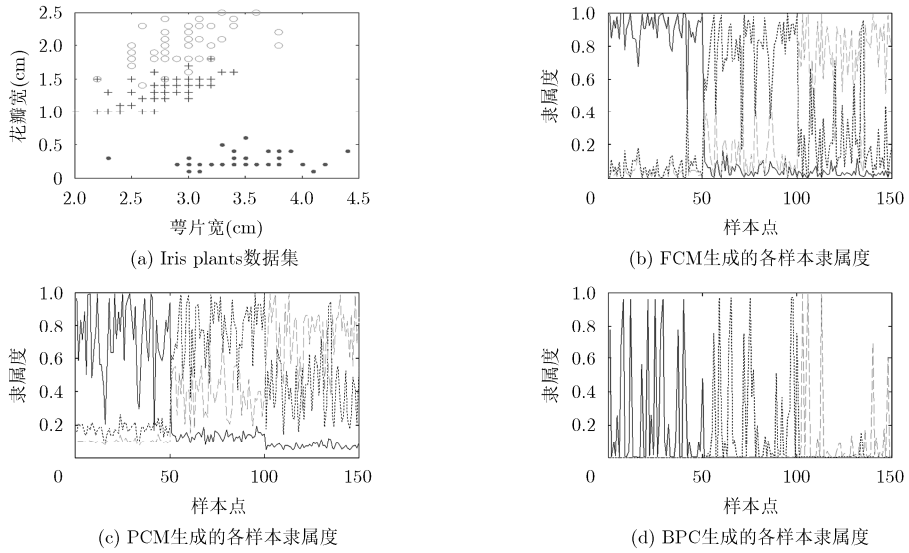
聚类指标	FCM	PCM	BPC
Accuracy	0.8933	0.8680	**0.9200**
RI	0.8597	0.8751	**0.9045**
NMI	0.7496	0.7651	**0.7732**
PURE	0.9028	0.8745	**0.9250**
F-measure	**0.7540**	0.6988	0.6758

5.4 图像数据集实验

该实验从 MSRA Salient Object Database[25]中选择了 3 幅图像 Butterfly, Dog 和 Boy，分别如图 3(a1)~3(a3)所示。基于他们的颜色分量，划分他们到不同类中。也即，在每个图像中，具有相同级别的像素被考虑划分到相同的类。在这些图像中，主要的对象几乎共享相同的强度级别，所以通过 3 种聚类方法可以将不同像素划分到不同类中。图 3(b1)~3(b3)，图 3(c1)~3(c3)和 图 3(d1)~3(d3)分别展示了 3 种方法在 3 幅图像的聚类结果。从图 3 中可以直观地发现本文所提 BPC 方法较之其他对比方法可以更好地检测图像中的主要对象。

6　结论

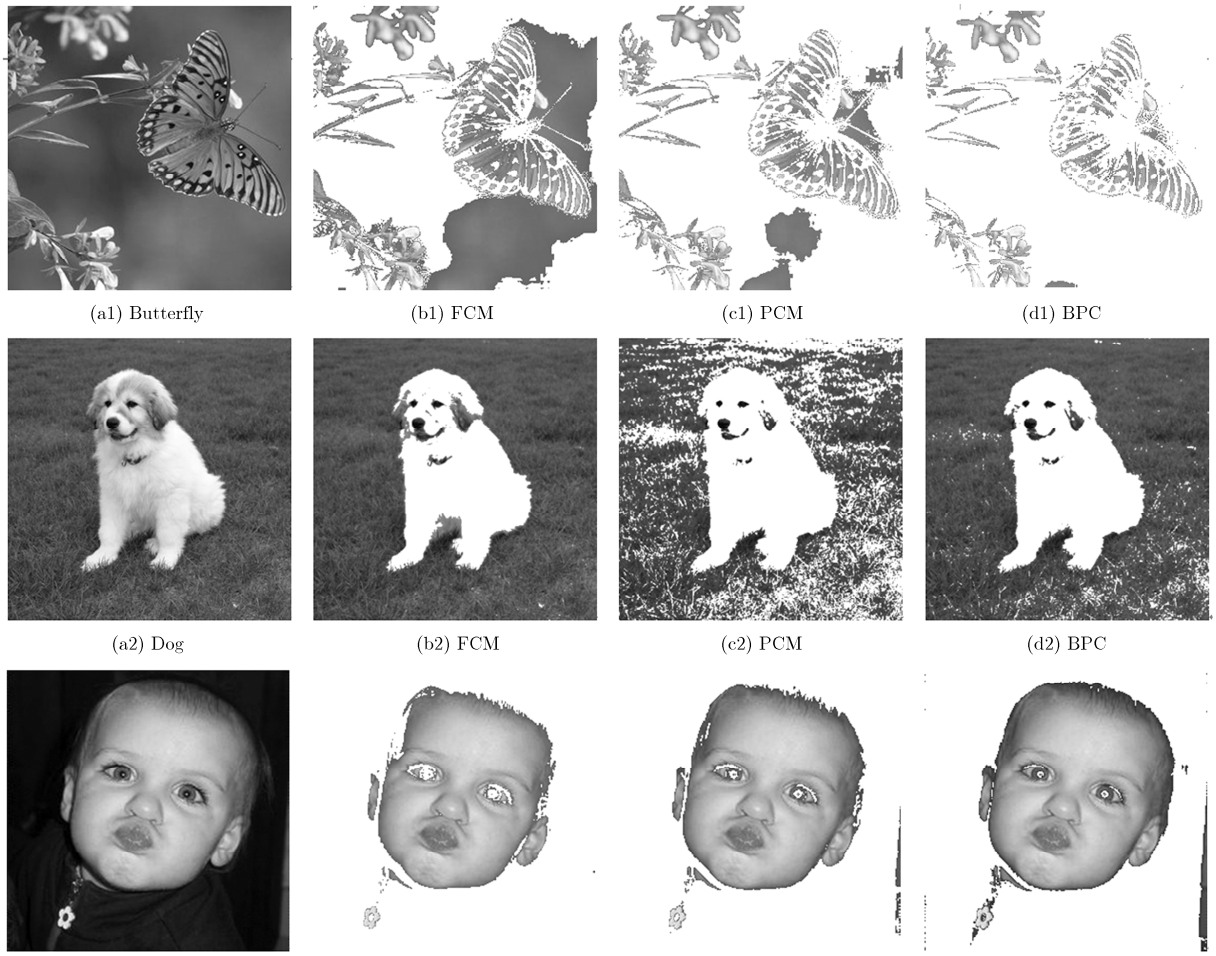
由于可能性和概率思想的盛行，可能性聚类和概率聚类已被广泛地研究，许多相关的算法和模型也已被开发。然而，到目前为止，他们相互关系和协作方面的研究还较为少见。为了利用两者的优点，我们从概率的角度研究了可能性聚类及其实现。本文提出了一种贝叶斯可能性聚类新方法 BPC。与传统可能性聚类模型不同，它将算法中待求解的隶属度和聚类中心作为随机变量，并为之选择一个合适的概率密度分布；在此基础上，基于 MAP 理论框架和 MCMC 采样方法求得该分布中的未知参数。同时对算法复杂度、收敛性作了理论探讨，并给出了参数选取准则。研究表明，本文所提 BPC 方法不但扩展了可能性聚类性能，且改进了聚类结果，在实际聚类问题上具有良好适应性。



(a) Iris plants数据集　(b) FCM生成的各样本隶属度

(c) PCM生成的各样本隶属度　(d) BPC生成的各样本隶属度

图 2　Iris plants 数据集及各算法生成的样本隶属度



(a1) Butterfly　(b1) FCM　(c1) PCM　(d1) BPC

(a2) Dog　(b2) FCM　(c2) PCM　(d2) BPC

(a3) Boy　(b3) FCM　(c3) PCM　(d3) BPC

(a)原图　(b) FCM聚类结果　(c) PCM聚类结果　(d) BPC聚类结果

图 3　3 幅图像及各算法划分的主要对象

参 考 文 献

[1] BARNI M, CAPPELLINI V, and MECOCCI A. Comments on "a possibilistic approach to clustering"[J]. *IEEE Transactions on Fuzzy Systems*, 1996, 4(3): 393–396.

[2] PAL N R, PAL K, and BEZDEK J C. A mixed c-means clustering model[C]. Proceedings of the Sixth IEEE International Conference on Fuzzy Systems, Barcelona, Spain, 1997: 11–21.

[3] PAL N R, PAL K, KELLER J M, *et al*. A possibilistic fuzzy c-means clustering algorithm[J]. *IEEE Transactions on Fuzzy Systems*, 2005, 13(4): 517–530. doi: 10.1109/tfuzz. 2004.840 099.

[4] KRISHNAPURAM R and KELLER J M. The possibilistic c-means algorithm: Insights and recommendations[J]. *IEEE Transactions on Fuzzy Systems*, 1996, 4(3): 385–393.

[5] ZHANG J S and LEUNG Y W. Improved possibilistic c-means clustering algorithms[J]. *IEEE Transactions on Fuzzy Systems*, 2004, 12(2): 209–217. doi: 10.1109/tfuzz. 2004.825079.

[6] YANG M S and LAI C Y. A robust automatic merging possibilistic clustering method[J]. *IEEE Transactions on Fuzzy Systems*, 2011, 19(1): 26–41. doi: 10.1109/tfuzz.2010. 2077640.

[7] 范九伦, 裴继红. 基于可能性分布的聚类有效性[J]. 电子学报, 1998, 26(4): 113–115.
FAN Jiulun and PEI Jihong. Cluster validity based on possibilistic distribution[J]. *Acta Electronica Sinica*, 1998, 26(4): 113–115.

[8] ZARANDI M H F, AVAZBEIGI M, and ANSSARI M H. New possibilistic noise rejection clustering algorithm with simulated annealing[C]. 2011 Annual Meeting of the North American Fuzzy Information Processing Society, Canada, 2011: 1–5. doi: 10.1109/nafips.2011.5752004.

[9] DENG Z H, CAO L B, JIANG Y Z, *et al*. Minimax probability TSK fuzzy system classifier: A more transparent and highly interpretable classification model[J]. *IEEE Transactions on Fuzzy Systems*, 2015, 23(4): 813–826. doi: 10.1109/tfuzz.2014.2328014.

[10] 夏建明, 杨俊安, 陈功. 参数自适应调整的稀疏贝叶斯重构算法[J]. 电子与信息学报, 2014, 36(6): 1355–1361. doi: 10.3724/SP.J.1146.2013.00629.
XIA Jianming, YANG Junan, and CHEN Gong. Bayesian sparse reconstruction with adaptive parameters adjustment[J]. *Journal of Electronics & Information Technology*, 2014, 36(6): 1355–1361. doi: 10.3724/SP.J.1146.2013.00629.

[11] 王峰, 向新, 易克初, 等. 基于隐变量贝叶斯模型的稀疏信号恢复[J]. 电子与信息学报, 2015, 37(1): 97–102. doi: 10.11999/JEIT140169.
WANG Feng, XIANG Xin, YI Kechu, *et al*. Sparse signals recovery based on latent variable Bayesian models[J]. *Journal of Electronics & Information Technology*, 2015, 37(1): 97–102. doi: 10.11999/JEIT140169.

[12] WANG S T, CHUNG K F, SHEN H B, *et al*. Note on the relationship between probabilistic and fuzzy clustering[J]. *Soft Computing*, 2004, 8(5): 366–369. doi: 10.1007/s00500-003-0309-8.

[13] YU L, WEI C, and ZHENG G. Adaptive Bayesian estimation with cluster structured sparsity[J]. *IEEE Signal Processing Letters*, 2015, 22(12): 2309–2313. doi: 10.1109 /lsp.2015.2477440.

[14] GLENN T C, ZARE A, and GADER P D. Bayesian fuzzy clustering[J]. *IEEE Transactions on Fuzzy Systems*, 2015, 23(5): 1545–1561. doi: 10.1109/tfuzz.2014.2370676.

[15] ZARINBAL M, ZARANDI M H F, and TURKSEN I B. Relative entropy fuzzy c-means clustering[J]. *Information Sciences*, 2014, 260: 74–97. doi: 10.1016/j.ins.2013.11.004.

[16] BEZDEK J C, EHRLICH R, and FULL W. FCM: The fuzzy c-means clustering algorithm[J]. *Computers & Geosciences*, 1984, 10(2-3): 191–203.

[17] KRISHNAPURAM R and KELLER J M. A Possibilistic approach to clustering[J]. *IEEE Transactions on Fuzzy Systems*, 1993, 1(2): 98–110.

[18] ANDRIEU C, DE FREITAS N, DOUCET A, *et al*. An introduction to MCMC for machine learning[J]. *Machine Learning*, 2003, 50(1): 5–43. doi: 10.1023/A:1020281327116.

[19] CHIB S and GREENBERG E. Understanding the metropolis-hastings algorithm[J]. *The American Statistician*, 1995, 49(4): 327–335.

[20] PLUMMER M, BEST N, COWLES K, *et al*. CODA: Convergence diagnosis and output analysis for MCMC[J]. *R News*, 2006, 6(1): 7–11.

[21] 朱崇军. MCMC样本确定的后验密度的收敛性[J]. 数学杂志, 2002, 22(3): 345–348. doi: 10.3969/j.issn.0255-7797.2002.03.019.
ZHU Chongjun. On the convergences of *a posteriori* density determined by MCMC samplers[J]. *Journal of Math ematics*, 2002, 22(3): 345–348. doi: 10.3969/j.issn.0255-7797.2002.03.019.

[22] ROBERTS G O and SMITH A F M. Simple conditions for the convergence of the Gibbs sampler and Metropolis-Hastings algorithms[J]. *Stochastic Processes and Their Applications*, 1994, 49(2): 207–216.

[23] ZELLNER A and MIN C K. Gibbs sampler convergence criteria[J]. *Journal of the American Statistical Association*, 1995, 90(431): 921–927.

[24] ASUNCION A and NEWMAN D J. UC irvine machine learning repository[OL]. http://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.names, 2015.

[25] LIU T, YUAN Z, SUN J, *et al*. Learning to detect a salient object[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011, 33(2): 353–367. doi: 10.1109/tpami.2010.70.

刘解放： 男，1982年生，博士生，研究方向为模式识别、数据挖掘.

王士同： 男，1964年生，教授，博士生导师，主要研究方向为模式识别、人工智能.

王 骏： 男，1978年生，博士，副教授，主要研究方向为智能计算、数据挖掘.

邓赵红： 男，1981年生，博士，教授，主要研究方向为智能计算、系统建模.