

基于聚类中心优化的 k-means 最佳聚类数确定方法

贾瑞玉, 宋建林

(安徽大学 计算机科学与技术学院, 安徽 合肥 230601)

摘 要: k-means 聚类算法是在确定的聚类数 k 下对数据集进行聚类, 通常凭借先验规则假定一个 k 值, 取值具有很大主观性; 此外, k-means 初始聚类中心的选择一般也是随机的, 这使得 k-means 聚类经常出现聚类指标局部最优, 聚类结果不稳定. 针对以上两个问题, 结合密度法改进了 k-means 初始聚类中心点的选择, 并在此基础上提出了一种确定 k-means 最佳聚类数的方法. 实验结果证明, 该方法可以得到更好的聚类结果, 具有更高的准确性、更好的稳定性以及更优的收敛性.

关键词: k-means 聚类; 初始聚类中心; 样本密度; 聚类数

中图分类号: TP301.6

文献标识码: A

文章编号: 1000-7180(2016)05-0062-05

K-means Optimal Clustering Number Determination Method Based on Clustering Center Optimization

JIA Rui-yu, SONG Jian-lin

(College of Computer Science and Technology, Anhui University, Hefei 230601, China)

Abstract: The traditional K-means clustering algorithm requires a certain number of clustering, usually the value of clustering number is assumed by priori rules. Besides, the choice of K-means initial clustering centers is also randomized, which made the clustering result instability. In view of the above factors, improved the selection of initial centers of K-means clusters, and a new method to determine the optimal clustering number is proposed based on that. The experimental results show that the proposed method can get better clustering results and have higher accuracy, better stability and better astringency.

Key words: words k-means clustering; initial clustering centers; density of samples; clustering number

1 引言

聚类分析是数据挖掘领域研究的一个重要技术, 一直受到该领域研究者的高度关注. 传统的 k-means 算法的聚类结果易受聚类数 k 值的影响, 对初始聚类中心的选择依赖性也比较大, 选择不同的初始聚类中心其聚类结果通常不一样, 结果具有很大的不确定性, 聚类指标往往会收敛于局部最优. 目前对 k-means 算法的研究主要集中在两个方向: 一是研究如何获得更好的初始聚类中心; 二是研究如何获得最佳的聚类数, 即最优的 k 值. 对于 k-means 算法初始聚类中心的选取, 文献[1]中提出了选取 k

个处于高密度分布的点作为初始聚类中心的算法; 在文献[2]中基于韩凌波的算法提出了当最大密度参数值对应的样本点不止一个时如何选取其中的样本点作为初始聚类中心的方法; 文献[3]中又提出了基于多层聚类中心的凝聚聚类算法. 对 k-means 算法最佳聚类数的确定研究, 在文献[4]中运用距离代价函数作为聚类有效性目标函数, 该方法假定当距离代价函数达到最小值时, 聚类结果达到最优, 并在理论上证明了经验规则 $k \leq \sqrt{n}$ 的合理性; 文献[5]中提出了一种新的聚类有效性指标, 称为 BWP 指标, 其值为聚类离差距离与聚类距离的比, 并依据此指标提出了确定 k-means 最佳聚类文献[6]提出了基

于动态遗传算法的确定最佳聚类数的算法;文献[7]中提出用数据分层来得到聚类数搜索上界,并提出一个聚类评价指标来得到最佳聚类数。

本文结合对 k-means 算法这两个方向的研究,受文献[2,5]算法的启发,提出了在优化 k-means 初始聚类中心选择的基础上确定最佳聚类数的值的方法。考虑到 k-means 算法聚类结果易受值和初始聚类中心的影响,本文针对这两个影响聚类结果的因素提出了改进方法,一是结合密度法和距离法提出对初始聚类中心的选择进行优化,二是在初始中心点优化的基础上应用一个新的聚类有效性评价函数来确定 k-means 算法的最佳聚类数。理论研究和实验结果表明,该算法的聚类结果优于原来的算法并且稳定性更好。

2 k-means 聚类算法

2.1 k-means 聚类算法的基本思想

传统的 k-means 算法是经典的基于划分的聚类算法,该算法把样本数为 n 的数据集划分为 k 个子集合,其中的每个子集合代表一个簇或类,同一个簇中的样本具有高相似度,不同簇间的样本具有低相似度。输入数据集 $S = \{x_1, x_2, \dots, x_n\}$,数据集有 n 个样本对象,然后根据先验规则给定一个 k 值作为划分的簇的数目,随机选取 k 个样本对象作为初始聚类中心,假设聚类中心为 $(c_1, c_2, \dots, c_k)$;然后把中心点外的样本点划分到离它们距离最近的一个中心点所在簇;再从形成的簇中重新计算聚类中心,重新划分簇;上述过程不断迭代,直到目标函数收敛或者新的聚类中心和上一次的聚类中心很接近,达到聚类终止的条件。

k-means 算法最常用的计算两个样本对象间的距离度量为欧氏距离,定义如下:

$$d(x_i, x_j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \dots + (x_{ip} - x_{jp})^2} \quad (1)$$

式中, $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})$, $x_j = (x_{j1}, x_{j2}, \dots, x_{jp})$, 表示两个 p 维属性的数据对象。文献[1]中定义了计算所有样本点平均距离的方法,公式如下:

$$\text{Meandist}(S) = \frac{2}{n(n-1)} \times \sum_{i \neq j, i, j=1}^n d(x_i, x_j) \quad (2)$$

式中, n 为数据集样本对象总数, $d(x_i, x_j)$ 为样本点 x_i 和 x_j 的欧氏距离。而 k-means 算法最常用的目标函数为平方误差准则函数,定义如下:

$$E = \sum_{i=1}^k \sum_{j \in c_i} \|x_j - c_i\|^2 \quad (3)$$

式中, C_i 表示第 i 个簇集合, c_i 表示第 i 个簇的中心。 E 表示所有数据样本对象与它的聚类中心的误差平方和。k-means 算法的具体实现步骤如下所示。

输入: 包含 n 个样本对象的数据集 S , 要划分的簇的数目 k 。

输出: 符合目标函数最优的 k 个簇。

① 从数据集 S 中随机选取 k 个对象点作为初始中心点。

② 把数据集中心点以外的点按式(1)计算各点到聚类中心的距离,然后划分到离它们最近的簇。

③ 计算簇中样本点平均值重新确定各个簇的中心,然后重新划分簇。

④ 计算 E 的值,如果 E 值收敛,则返回 $(c_1, c_2, \dots, c_k)$,输出聚类结果。否则,返回步骤②继续迭代。

2.2 k-means 算法最佳聚类数的研究

由于 k 值的选择对 k-means 的聚类结果有较大的影响,所以确定一个最佳的聚类数对聚类的准确性以及提高聚类的稳定性有很大的帮助。最佳聚类数的确定通常是通过一种迭代过程来实现,该过程先假定一个聚类有效性指标,使用不同的参数(通常为 k 的值),运用指定的聚类算法对数据集进行不同 k 值下的聚类,然后计算 k 值不同时的聚类有效性指标的值,通过比较各个指标的值,当有效性指标值达到评价最优时所对应的聚类数 k 的值即为最佳聚类数。目前已经提出的聚类有效性指标主要有:Ca-linski-Harabasz (简称 CH) 指标、Davies-Bouldin (简称 DB) 指标、Weighted Inter-intra (简称 Wint 指标),文献[5]提出了一种新的聚类有效性指标(称为 BWP 指标,其值为聚类离差距离与聚类距离的比值),并依据此指标提出了确定 k-means 最佳聚类数的方法。相关定义如下[5]所示。

定义 1 存在 n 个数据对象的数据集 $S = \{x_1, x_2, \dots, x_n\}$, 假设 n 个数据对象被划分到 k 个类簇, 定义第 j 类的对象 i 的类间距离 $b(j, i)$ 为该样本到其他每个类中样本平均值中的最小值, 如下:

$$b(j, i) = \min_{1 \leq c \leq k, c \neq j} \left(\frac{1}{n_c} \sum_{p=1}^{n_c} \|x_p^c - x_i^j\|^2 \right) \quad (4)$$

式中, c 和 j 表示类标, n_c 表示类 c 的元素个数, x_p^c 表示第 c 类的第 p 个对象, $x_i^{(j)}$ 表示第 j 类的第 i 个对象。 $\| \|^2$ 表示欧氏距离的平方。

定义 2 存在 n 个数据样本的数据集 $S =$

$\{x_1, x_2, \dots, x_n\}$, 假设 n 个数据对象被划分到 k 个簇中, 定义第 j 类的对象 i 的类内距离 $w(j, i)$ 为该数据对象到 j 类中其他数据对象距离的平均值, 如下:

$$w(j, i) = \left(\frac{1}{n_c - 1} \right) \sum_{p=1, p \neq i}^{n_j} \|x_p^j - x_i^j\|^2 \quad (5)$$

定义 3 存在 n 个数据对象的数据集 $S = \{x_1, x_2, \dots, x_n\}$, 假设 n 个数据对象被划分到 k 个类簇, 定义第 j 类的数据对象 i 的聚类有效性指标为聚类离差距离与聚类距离的比:

$$\text{BWP}(j, i) = \frac{b(j, i) - w(j, i)}{b(j, i) + w(j, i)} \quad (6)$$

通过 BWP 指标值来确定最佳聚类数, 通过分析可知, $b(j, i)$ 越大说明类间分离性越好, $w(j, i)$ 越小类内一般分布越紧密; 样本对象的 $\text{BWP}(j, i)$ 指标值越大则说明数据对象 i 的聚类效果越好. 通过比较聚类后数据集 n 个数据对象的 BWP 的平均值的大小来确定最佳聚类数, 即下面式中指标值最大时所对应的聚类数为最佳聚类数.

$$\text{avg}_{\text{bwp}}(k) = \frac{1}{n} \sum_{j=1}^k \sum_{i=1}^{n_j} \text{BWP}(j, i) \quad (7)$$

最佳聚类数为 $\text{avg}_{\text{bwp}}(k)$ 取值最大时对应的聚类数, 文献[5]提出的这个确定最佳聚类数的方法存在不足之处, 其初始聚类中心的选择是随机的, 这有可能选择一些噪声点作为初始聚类中心, 导致聚类结果不稳定; 此外, 聚类数 k 的值变化时, 不同聚类数聚类时的初始中心点各不一样, 聚类结果对比性不强, 把其聚类结果做对比会存在一定的偏差. 本文提出了基于初始聚类中心选择优化的 k-means 最佳聚类数确定方法, 该方法先产生一个比较好的初始聚类中心候选集, 每次聚类从这个集合中选取初始聚类中心点, 当搜索聚类数时只变动部分中心点, 从而使聚类结果更稳定, 对比性更强, 聚类效果更好.

3 聚类中心优化的 k-means 最佳聚类数确定方法

3.1 获得初始中心点候选集的方法

对 k-means 算法初始中心点的选择, 要求选取的中心点能很好的反映数据集中对象之间的关系和数据的分布特征, 这对聚类结果的准确性和稳定性以及聚类效率都有很大的影响. k-means 算法对数据对象之间的相似性度量一般采用的是欧氏距离, 当数据对象在给定的空间范围内数据点越多, 说明该数据对象作为聚类中心点更有利于目标函数的收

敛. 受文献[2]提出的算法启发, 本文依据密度法提出了一种产生一个 k-means 算法初始聚类中心点候选集的方法. 文献[1]中提出了计算样本密度的方法, 定义如下:

定义 4 存在 n 个数据对象的数据集 $S = \{x_1, x_2, \dots, x_n\}$, 定义数据对象 x_i 的密度参数为 $\text{density}(x_i, \text{Meandist})$, 表示以 x_i 为中心, 式(2)定义的样本平均距离 Meandist 为半径的区域中对象点的个数:

$$\text{density}(x_i, \text{meandist}) = \sum_{j=1}^n u(\text{meandist} - d(x_i, x_j)) \quad (8)$$

式中, $u(x) = \begin{cases} 1, & x > 0 \\ 0, & \text{other} \end{cases}$, 计算集合 S 中所有样本对象的密度参数, 组成密度参数集合 D , $D = \{\text{density}(x_i, \text{meandist}), i \in (1, 2 \dots n)\}$. 利用下面步骤产生一个用于搜索聚类数时的初始聚类中心点候选集合 U , 依据文献[4]中证明的先验规则, 设定 U 中元素个数为 $\text{int}\sqrt{n}$, 其元素为在取不同的聚类数 k 时可能选取到的初始聚类中心, U 的生成步骤如下所示.

(1) 利用式(2)计算样本间的平均距离 Meandist, n 为数据集样本对象总数, $d(x_i, x_j)$ 为利用式(1)计算的样本点 x_i 和 x_j 的距离.

(2) 计算数据集中每个样本的密度参数 $\text{density}(x_i, \text{Meandist})$, 并形成密度参数集合 D .

(3) 找出 D 中密度参数值最大的样本参数 $\max I$, 如果 $\max I = 1$, 找出 D 中最大密度参数 D_i 对应的数据集 S 中的样本点 i , 并把样本点 i 添加到初始中心点候选集合 U 中, 然后转到步骤(5)执行.

(4) 若 $\max I > 1$ 并且对应的样本点之间的距离小于 Meandist, 则计算最大密度参数值对应的各样本点 Meandist(S) 领域内所有点与 i 的距离和 $\text{sum}(D_i) = \sum_{j=0}^{n_i} d(i, j)$, $\text{sum}(D_i)$ 为数据样本点 i 的 Meandist(S) 邻域内所有数据对象到 i 的距离的和, 把 sum 取最小值时对应的数据对象添加到集合 U 中.

(5) 把数据样本点 i 及 D 中与样本点 i 距离小于 Meandist(S) 的样本点从 D 集合中删除.

(6) 重复步骤(3) ~ (5), 直到找出 $\text{int}\sqrt{n}$ 个候选中心点, 最终初始中心点候选集合 U 中有 $\text{int}\sqrt{n}$ 个数据对象.

3.2 新的聚类有效性指标

一个好的聚类划分应该尽可能反映数据集的内在结构,使类内样本最大相似,类间样本相似度最小.本文在确定最佳聚类数的搜索时为了确保聚类结果稳定性以及聚类结果的对比性,使每次变动因子最少,聚类时初始中心点的选择都是在中心点候选集 U 中选取,选取 U 中的前 k 个样本对象为其初始中心点.本文提出了一种新的聚类有效性指标,称之为改进的类间类内划分 (Improved Between-Within Proportion, 文中简称为 IBWP 指标).新指标相关概念定义如下:

定义 5 样本对象的类内距离定义如下,存在 n 个数据样本的数据集 $S = \{x_1, x_2, \dots, x_n\}$,假设 n 个数据样本被划分到 k 个类簇,定义第 j 类的样本对象 i 的类内距离 $iw(j, i)$ 为该样本到 j 类中其他样本距离的平均值,计算公式如下:

$$iw(j, i) = \left(\frac{1}{n_c - 1} \right) \sum_{p=1, p \neq i}^{n_j} d(x_p^j, x_i^j) \quad (9)$$

定义 6 存在 n 个数据对象的数据集,假设 n 个数据对象被划分到 k 个类簇,定义第 j 类的对象 i 的类间距离(文中称为 $ib(j, i)$)为该样本到其他每个类中心点距离最小值,如下:

$$ib(j, i) = \min_{1 \leq p, i \leq k, p \neq j} \{d(i, c_p), c_p \in U\} \quad (10)$$

定义 7 存在 n 个数据对象的数据集 $S = \{x_1, x_2, \dots, x_n\}$,假设 n 个数据对象被划分到 k 个类簇,定义第 j 类的对象 i 的聚类距离(文中称为 $cd(j, i)$)为 i 的类间聚类和类内距离之和:

$$cd(j, i) = iw(j, i) + ib(j, i) \quad (11)$$

定义 8 存在 n 个数据对象的数据集 $S = \{x_1, x_2, \dots, x_n\}$,假设 n 个数据对象被划分到 k 个类簇,定义第 j 类的对象 i 的聚类离差距离(文中称为 $ibarw(j, i)$)为 i 的类间距离和类内距离之差:

$$ibarw(j, i) = ib(j, i) - iw(j, i) \quad (12)$$

定义 9 存在 n 个数据对象的数据集 $S = \{x_1, x_2, \dots, x_n\}$,假设 n 个数据对象被划分到 k 个类簇,定义第 j 类的对象 i 的聚类有效性指标为

$$IBWP(j, i) = \frac{ib(j, i) - iw(j, i)}{ib(j, i) + iw(j, i)} \quad (13)$$

对新聚类指标进行分析发现,样本对象 i 的类内距离 $iw(j, i)$ 能比较好地反映样本点与类中其他样本点的关系,能很好地反映类簇内的结构;样本对象 i 的类间距离 $ib(j, i)$ 可以比较好地反映样本对象与其他类的关系,如果样本 i 与最近的类簇满足类间远离原则,则与其他类更能满足这个原则,此外,

如果样本点 i 不是划分在类 j 中,则一定是划分在离它最近的这个类中. IBWP 指标很好地反映了单个的样本点的聚类有效性指标, IBWP 值越大,说明单个样本点的聚类效果越好.通过求数据集的 n 个样本的 IBWP 值,然后求其平均值 $AVG_{ibwp}(k)$, k 表示聚类数.

$$avg_{ibwp}(k) = \frac{1}{n} \sum_{j=1}^k \sum_{i=1}^{n_j} IBWP(j, i) \quad (14)$$

$$k_{opt} = \max_{2 \leq k \leq \sqrt{n}} avg_{ibwp}(k) \quad (15)$$

3.3 k-means 最佳聚类数确定方法描述

本文结合传统 k-means 算法和文中的初始中心点选择优化方法以及 IBWP 聚类有效性指标,提出了一种新的确定最佳聚类数的方法,如下所示.

(1) 应用本文的初始中心点选择方法,获得一个包含有 $\text{int} \sqrt{n}$ 个样本对象的初始聚类中心候选集合 U .

(2) 聚类数的搜索,搜索范围为 $[k_{\min}, k_{\max}]$, $k_{\min} = 2, k_{\max} = \text{int} \sqrt{n}$.

(3) k 值从 $k_{\min}$ 循环至 $k_{\max}$, 步骤如下:

① 从集合 U 中选择前 k 个数据对象点作为初始聚类中心点.

② 把数据集 S 中除选择的初始聚类中心点外的其他样本点划分到最近的簇中.

③ 聚类后利用式(12) 分别计算数据集中 n 个样本对象的 IBWP 指标值.

④ 利用式(13) 计算本次聚类的平均 IBWP 值.

(4) 依据式(14) 求出最佳聚类数 k_{opt} .

(5) 输出最佳聚类数 k_{opt} , 最佳聚类数所对应的平均 IBWP 值, 对应的初始聚类中心点.

4 实验结果与分析

为了检验提出的聚类方法的有效性和效果,选取了专用于测试聚类算法性能的 UCI 数据库中的 iris 数据集、seeds 数据集、wine 数据集作为实验数据如图 1~3 所示, iris 数据集中有 150 条数据,每条数据有 4 个属性; seeds 数据集有 210 条数据,每条数据有 8 个属性; wine 数据集有 178 条数据,每条数据有 13 个属性.实验对数据集的聚类数搜索范围为 $[2, k_{\max}]$, 根据普遍使用的经验规则取 $k_{\max} = \text{int} \sqrt{n}$. 由于 k-means 聚类算法易受初始中心点选择的影响,算法先优化了初始中心点的选择,求出一个比较好的初始聚类中心点候选集合,使聚类数不同时只变动部分初始中心点,增强了聚类的稳定性,也增加

了不同聚类数时计算出的指标值的对比性,此外也能确定一个较为合适的聚类初始中心点.实验对比了在这3个UCI数据集上几个常用的聚类指标值的聚类评价结果.如表1所示.

表1 几种聚类指标在三个 UCI 数据集上的评价结果

数据集	样本数	属性数	最佳聚类数			
			CH 指标	DB 指标	BWP 指标	IBWP 指标
IRIS	150	4	2	2	2	2
WINE	178	13	3	2	2	2
SEEDS	210	8	3	3	3	3

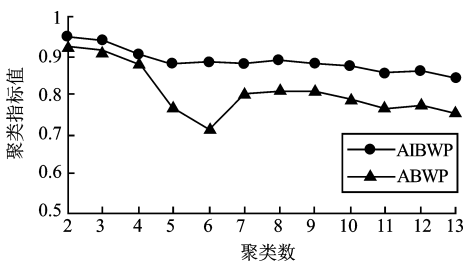


图1 IRIS 数据集聚类数—指标值图

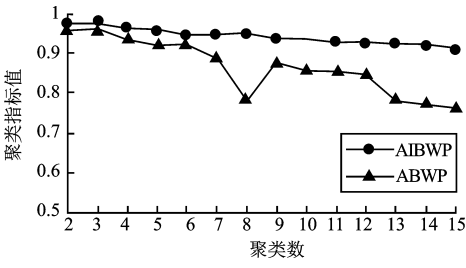


图2 SEEDS 数据集聚类数—指标值图

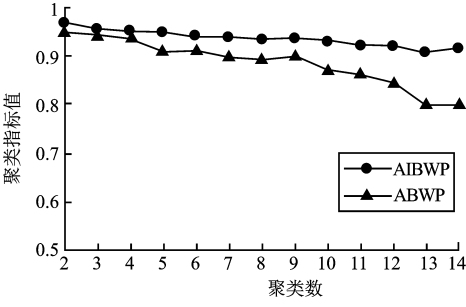


图3 WINE 数据集聚类数—指标值图

从图1~图3的实验结果可以看出,本文的方法得出的聚类结果比较稳定,而文献[5]提出的方法每次聚类的初始聚类中心是随机选择的,因而得出的聚类指标值波动较大;由上面3个实验结果图可以看出,AIBWP值比ABWP值保持在更稳定的水平,说明本文算法的聚类结果稳定性更好,此外,AIBWP一直保持在一个比ABWP更大的值,说明

本文的聚类算法对样本的聚类准确性更高,此外,还能生成一个比较好的初始中心点候选集,如表2所示.

表2 UCI 数据集的初始中心点候选集

数据集	初始中心点候选集合
IRIS	$S = \{x_{80}, x_{99}, x_{77}, x_{147}, x_{135}, x_{149}, x_{115}, x_{142}, x_{71}, x_{98}, x_{91}, x_{107}, x_{130}\}$
WINE	$S = \{x_{167}, x_{141}, x_{41}, x_{68}, x_{153}, x_{112}, x_{36}, x_{51}, x_{39}, x_{34}, x_{12}, x_{61}, x_{55}, x_{19}\}$
SEEDS	$S = \{x_{136}, x_{17}, x_{198}, x_{15}, x_{202}, x_{52}, x_{10}, x_{77}, x_{61}, x_{172}, x_{142}, x_{110}, x_{79}, x_{171}, x_{115}\}$

表2是本文算法求出的UCI中3个实验数据集中比较适合作为初始聚类中心点的样本对象集合,当搜索聚类数时,从中心点候选集中选取初始中心点.这样可以增强聚类的准确性以及每次搜索聚类数聚类时的对比性.实验结果表明,本文方法可以有效地确定比较好的初始中心点,以及得到正确的最佳聚类数.

5 结束语

k-means 聚类算法对聚类数 k 值和初始聚类中心点的选择依赖性较强,算法聚类结果易受 k 值和中心点的影响,而在大多数情况下 k 值是依据先验规则给定的,聚类中心点是随机选择的,所以 k-means 聚类算法的聚类结果不稳定.本文针对影响 k-means 聚类有效性的这两个问题进行研究,优化了初始中心点的选择,在此基础上提出了一种聚类有效性指标 (IBWP 指标),提出一个确定 k-means 聚类算法最佳聚类数的方法.理论和实验分析验证了方法的有效性和良好性能,既能确定 k-means 聚类算法的最佳聚类数,同时又能确定比较好的初始聚类中心点.本文的聚类数搜索范围为 $[k_{\min}, k_{\max}]$, $k_{\min}=2, k_{\max}=\text{int} \sqrt{n}$,是根据经验规则确定的,聚类算法的时间复杂度大;因此,未来的研究工作主要是对聚类数搜索范围进行优化,以获得更好更准确的搜索范围.

参考文献:

[1] 韩凌波,王强.一种改进的 k-means 初始聚类中心选取算法[J].计算机工程与应用,2010,46(17):150-152.
[2] 黄敏,何中市.一种新的 k-means 聚类中心选取算法[J].计算机工程与应用,2011,47(35):132-134.

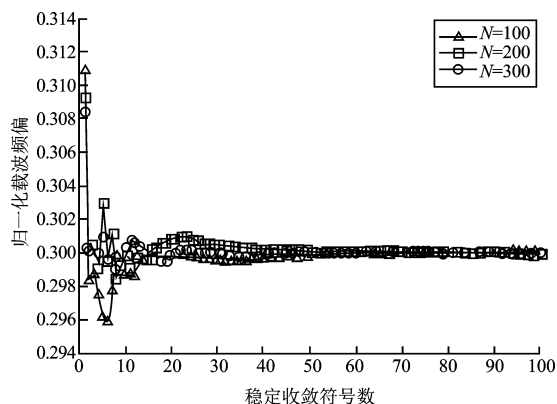


图1 不同粒子数频偏估值收敛状态

下对算法的估计方差和误码率进行仿真. 可知算法对不同频偏的估计性能具有较好的鲁棒性,同时也验证了样本容量越多,通过粒子权值调整的后验分布与状态变量的真实后验密度越近似,算法估计性能更好. 本文算法在高斯和非高斯噪声(ϵ 混合噪声)环境下频偏估值为 $fT=0.3$ 时的估计情况,可以看到在非高斯噪声下该算法仍具有较好的频偏跟踪性能.

6 结束语

通过建立载波频偏非线性状态空间模型,提出一种基于 Unscented 高斯粒子滤波算法进行载波频偏估计算法. 该算法引入 UKF 思想对高斯免重采样粒子滤波进行优化,使用最新观测值提高采样粒子质量,从而更好改善算法性能. 仿真表明,该算法提高了估计精度和估计收敛速度,在不同频偏以及非高斯噪声条件下均具有较好的估计性能. 针对更

多信号调制方式及恶劣信道环境考察算法性能是未来需要继续研究的内容.

参考文献:

- [1] Lopez Salcedo J A, Del Peral Rosado J A, Seco Granda G. Survey on robust carrier tracking techniques [J]. IEEE Communications Surveys & Tutorials, 2014, 16(2):670-688.
- [2] 李妍妍, 张伟. 基于 UPF 的 GPS/SINS 组合导航系统设计[J]. 微电子学与计算机, 2014, 31(9):80-83.
- [3] 吴天琳, 彭华, 黄艳艳, 等. 高阶 QAM 信号多状态自适应卡尔曼滤波载波同步算法[J]. 信号处理, 2014, 30(6):718-727.
- [4] 赵琳. 非线性系统滤波理论[M]. 北京:国防工业出版社, 2012:3-5.
- [5] 王法胜, 鲁明羽, 赵清杰, 等. 粒子滤波算法[J]. 计算机学报, 2014, 37(8):104-107.
- [6] Hussein Hijazi. Joint CFO and time-varying channel estimation by particle filtering in OFDM systems[C]//The 3rd International Conference on Communications and Information Technology. Beirut, 2013:241-245.
- [7] 郑作虎, 王首勇. 一种基于 GHF 的高斯粒子滤波算法[J]. 控制与决策, 2014, 29(9):1698-1702.

作者简介:

焦玲女, (1991-), 硕士研究生. 研究方向为载波同步技术. E-mail: thelevelplain@sina.cn

蒋磊男, (1974-), 博士, 副教授. 研究方向为超宽带通信关键技术.

许华男, (1976-), 博士, 副教授. 研究方向为信号盲处理、通信信号分析技术.

(上接第 66 页)

- [3] 马儒宁, 王秀丽, 丁军娣. 多层核心集凝聚算法[J]. 软件学报, 2013, 24(3):490-506.
- [4] 杨善林, 李永森, 胡笑旋. k-means 算法中的 k 值优化问题研究[J]. 系统工程理论与实践, 2006, 26(2):97-101.
- [5] 周世兵, 徐振源, 唐旭清. k-means 算法最佳聚类数确定方法[J]. 计算机应用, 2010, 30(8):1995-1998.
- [6] 何宏, 谭永红. 一种基于动态遗传算法的聚类新方法[J]. 电子学报, 2012, 40(2):254-259.

- [7] 王勇, 唐靖. 高效率的 K-means 最佳聚类数确定算法[J]. 计算机应用, 2014, 34(5):1331-1335.

作者简介:

贾瑞玉女, (1965-), 副教授, 硕士生导师. 研究方向为智能算法与数据挖掘.

宋建林(通讯作者)男, (1990-), 硕士研究生. 研究方向为数据挖掘. E-mail: 435247994@qq.com.