文章编号:1007-5321(2018)04-0029-08 **DOI**:10.13190/j.jbupt.2018-026

增量式模糊 C 有序均值聚类算法

刘永利, 郭呈怡, 王恒达, 晁 浩
(河南理工大学 计算机科学与技术学院, 河南 焦作 454000)

摘要: 针对传统聚类算法难以处理大规模数据和对噪声数据敏感等问题,基于模糊 C 有序均值聚类算法(FCOM),结合 single-pass 和 online 增量架构,分别提出了 single-pass 模糊 C 有序均值聚类算法(SPFCOM)和 online 模糊 C 有序均值聚类算法(OFCOM). SPFCOM 和 OFCOM 算法首先对 FCOM 算法加权,然后以数据块为单位对数据集合进行增量式处理. 实验结果表明,相较于对比算法,SPFCOM 和 OFCOM 算法在聚类准确率方面得到了提高,还具有更强的鲁棒性.

关 键 词: 模糊聚类; 增量聚类; 鲁棒性
中图分类号: TP391.1 **文献标志码**: A

Incremental Fuzzy C-Ordered Means Clustering

LIU Yong-li, GUO Cheng-yi, WANG Heng-da, CHAO Hao
(School of Computer Science and Technology, Henan Polytechnic University, Henan Jiaozuo 454000, China)

Abstract: Because traditional clustering algorithms are difficult to deal with large-scale data and sensitive to noise data, based on the Fuzzy C-ordered-means clustering (FCOM) algorithm, we propose a single-pass fuzzy C-ordered clustering algorithm, named SPFCOM, and an online fuzzy C-ordered clustering algorithm, named OFCOM, by combining single-pass and online incremental architectures respectively. These two algorithms weight the FCOM algorithm, and incrementally process the large-scale data chunk by chunk. Experimental results show that, compared with other similar prominent algorithms, the SPFCOM and OFCOM algorithms can achieve higher accuracy and better robustness.

Key words: fuzzy clustering; incremental clustering; robustness

模糊 C 均值算法(FCM, fuzzy C-means clustering)[1]是模糊聚类的代表性算法,得到了广泛应用. 然而,该算法对噪声比较敏感,极易造成聚类精度的下降. 于是,在 FCM 基础上,衍生出许多改进算法,其改进思路一般包括 2 种,即采用 Huber 的 M-estimators[2]或 Yager 的 OWA 运算符[3]. 2016 年,Leski 提出了模糊 C 有序均值算法(FCOM, fuzzy C-ordered-means clustering)[4],该算法结合了 FCM 算法和有序加权平均运算符,将 M-estimator 和 OWA 运算符同时应用于模糊聚类中,通过排序提高了算法的健壮性和鲁棒性. 但是,与多数传统聚类算法类似,FCOM 算法难以处理大规模数据或数据流. 为了解决此类问题,Hore 等[5-6]提出了 SPFCM(single-pass fuzzy C-means)和 OFCM(online fuzzy C-means) 2 种增量算法,采用的 single-pass 和 online 框架逐渐发展成为增量聚类算法研究中的重要方法. single-pass 方法的基本思想是将数据分成若干数据块(chunk),对每个数据块依次进行聚类,聚类

收稿日期: 2018-01-26
基金项目: 河南省高等学校青年骨干教师项目(2015GGJS-068); 河南省科技攻关计划项目(172102210279); 河南省高校基本科研业务费专项资金项目(NSFRF1616)
作者简介: 刘永利(1980—), 男, 副教授, 硕士生导师, E-mail: yongli.buaa@gmail.com.

后得到的质心参与到下一个数据块的聚类运算中,直至得到最终结果;online 方法则采用并行方式,针对每个数据块聚类之后得到的质心再次聚类,进而得到最终的聚类结果. SPFCM 和 OFCM 2 种算法虽然可以处理大规模数据,但依然是类 FCM 算法,仍然对噪声数据比较敏感.

基于 single-pass 和 online 增量框架分别提出一种模糊 C 有序均值聚类算法:SPFCOM 算法(single-pass fuzzy C-ordered-means clustering)和 OFCOM 算法(online fuzzy C-ordered-means clustering). SPFCOM 和 OFCOM 算法同时继承了 FCOM 和增量算法的优点:一方面通过采用 FCOM 的思想保证算法的鲁棒性;另一方面通过采用增量方法可有效处理大规模数据.

1 相关工作

1.1 FCOM 算法

FCOM 算法的目标函数为

$$J(U,V)=\sum_{c=1}^{C}\sum_{i=1}^{N}\beta_{ci}(u_{ci})^{m}D(x_i,v_c) \tag{1}$$

其中:参数 β_{ci} 表示第 i 个数据对第 c 个簇的典型性,$D(x_i,v_c)$ 可由 H_{cij} 和 E_{cij} 计算得出,即

$$D(x_i,v_c)=\sum_{j=1}^{K}D(x_{ij},v_{cj})=\sum_{j=1}^{K}H_{cij}(E_{cij})^2$$

该目标函数受限于

$$\forall_{\substack{1\leqslant c\leqslant C\\1\leqslant i\leqslant N}}u_{ci}\in[0,1] \tag{2}$$

$$\sum_{c=1}^{C}\beta_{ci}u_{ci}=F_i \tag{3}$$

其中 F_i 表示第 i 个数据对于所有簇的综合典型性.

1.2 基于 FCM 的增量模糊聚类

single-pass 和 online 是聚类研究中常用的 2 种增量化思想,最早分别应用在 SPFCM 算法[5]与 OFCM 算法[6]中. 这 2 种算法均以 WFCM(weighted fuzzy C-means clustering)为基础.

1.2.1 WFCM 算法

WFCM 是加权的 FCM 算法,即在 FCM 算法迭代过程中,对得到的质心进行加权,为重要性高的对象赋予较高权重. 假设目标数据集为 $\boldsymbol{X}=[x_1,x_2,\cdots,x_n]$,则 WFCM 的目标函数 J_{WFCM} 为

$$J_{\mathrm{WFCM}}=\sum_{c=1}^{C}\sum_{i=1}^{N}w_iu_{ci}^{m}\parallel x_i-v_c\parallel^2 \tag{4}$$

其中 w_i 为第 i 个数据的权重. 通过拉格朗日乘子法可得 u_{ci} 和 v_c 的迭代公式为

$$u_{ci}=\frac{\parallel x_i-v_c\parallel^{\frac{-2}{m-1}}}{\sum_{k=1}^{C}\parallel x_i-v_k\parallel^{\frac{-2}{m-1}}} \tag{5}$$

$$v_c=\frac{\sum_{i=1}^{N}w_iu_{ci}^{m}x_i}{\sum_{i=1}^{N}w_iu_{ci}^{m}} \tag{6}$$

1.2.2 SPFCM 与 OFCM 算法

SPFCM 基于 WFCM 算法进行聚类. 将所有对象的权重初始化为 1,即 $\boldsymbol{w}_{\mathrm{data}}=[1,1,\cdots,1]^{\mathrm{T}}$,假定数据集划分为 h 个数据块,即 $\boldsymbol{X}=[\boldsymbol{X}^1,\boldsymbol{X}^2,\cdots,\boldsymbol{X}^h]$,$\boldsymbol{X}^t$ 表示第 t 个数据块($1\leqslant t\leqslant h$). 对 $\boldsymbol{X}^1$ 进行聚类时,得到质心 $\Delta=[v_1,v_2,\cdots,v_c]$ 及其权重 $\boldsymbol{w}_c=\sum_{i=1}^{n}(u_{ci})\boldsymbol{w}_{\mathrm{data}}$,其中 n 为该数据块的对象个数;当 $t>1$ 时,将第 $t-1$ 个数据块的质心加入到第 t 个数据块中,得到新数据块 $\boldsymbol{X}^{t'}=[\Delta^{t-1},\boldsymbol{X}^t]$,同时更新质心权重 $\boldsymbol{w}_c^t=\sum_{i=1}^{n}(u_{ci})[\boldsymbol{w}_c^{t-1},\boldsymbol{w}_{\mathrm{data}}]$.

OFCM 同样基于 WFCM 算法进行聚类. 首先对每个数据块单独聚类,然后将每个数据块得到的质心和权重组成一个总的数据块和权重矩阵,再次聚类得到最终质心,其中每个数据块聚类后的权重表示为 $\boldsymbol{w}=[\boldsymbol{w}_1^1,\boldsymbol{w}_2^1,\cdots,\boldsymbol{w}_c^1,\boldsymbol{w}_1^2,\boldsymbol{w}_2^2,\cdots,\boldsymbol{w}_c^2,\cdots,\boldsymbol{w}_1^h,\boldsymbol{w}_2^h,\cdots,\boldsymbol{w}_c^h]$.

2 增量式模糊 C 有序均值聚类算法

分别基于 single-pass 和 online 增量框架,提出了增量式模糊 C 有序均值聚类算法:SPFCOM 和 OFCOM. 在 SPFCOM 和 OFCOM 算法中,首先设计加权的 FCOM 算法 WFCOM(weighted fuzzy C-ordered-means clustering),其目标函数为

$$J(U,V)=\sum_{c=1}^{C}\sum_{i=1}^{N}w_i\beta_{ci}(u_{ci})^{m}D(x_i,v_c) \tag{7}$$

约束条件为

$$\forall_{\substack{1\leqslant c\leqslant C\\1\leqslant i\leqslant N}}u_{ci}\in[0,1] \tag{8}$$

$$\sum_{c=1}^{C}\beta_{ci}u_{ci}=F_i \tag{9}$$

最小化该目标函数,得到 u_{ci} 和 v_{cj} 的迭代公式为

$$\forall_{\substack{1<i<N\\1<c<C}}u_{ci}=\frac{F_iD(x_i,v_s)^{\frac{1}{1-m}}}{\sum_{t=1}^{C}\beta_{ti}D(x_i,v_t)^{\frac{1}{1-m}}} \tag{10}$$

$$\underset{\substack{1<c<C\\1<j<K}}{\forall} v_{cj}=\frac{\sum_{i=1}^{N} w_i\beta_{ci}(u_{ci})^m H_{cij}x_{ij}}{\sum_{i=1}^{N} w_i\beta_{ci}(u_{ci})^m H_{cij}} \tag{11}$$

2.1　SPFCOM 算法

SPFCOM 算法在 WFCOM 基础上，采用 single-pass 增量框架进行实现. 具体实现过程为：首先初始化权重，将所有数据对象的权重初始化为 1，即 $\boldsymbol{w}=[1,1,\cdots,1]^{\mathrm{T}}$，并假设数据集划分为 h 个数据块，即 $\boldsymbol{X}=[\boldsymbol{X}^1,\boldsymbol{X}^2,\cdots,\boldsymbol{X}^h]$.

1）当 $t=1$ 时，聚类过程与 FCOM 算法类似，可得质心矩阵 $\boldsymbol{\Delta}^1=[v_1,v_2,\cdots,v_c]$ 和相应的权重矩阵：

$$\boldsymbol{w}'_c=\sum_{i=1}^{n_1}(u_{ci})w_i \tag{12}$$

其中 $1\leqslant c\leqslant C, w_i=1, 1\leqslant i\leqslant n_1$，$n_t$ 为第 t 个数据块的对象个数.

最小化目标函数，对数据块 $\boldsymbol{X}^1$ 聚类后隶属度 u_{ci} 和质心 v_{cj} 的公式为

$$u_{ci}=\frac{F_i D(x'_i,v_c)^{\frac{1}{1-m}}}{\sum_{t=1}^{C}\beta_{ti}D(x'_i,v_t)^{\frac{1}{1-m}}} \tag{13}$$

$$v_{cj}=\frac{\sum_{i=1}^{n_1} w_i\beta_{ci}(u_{ci})^m H_{cij}x_{ij}}{\sum_{i=1}^{n_1} w_i\beta_{ci}(u_{ci})^m H_{cij}} \tag{14}$$

其中 $1\leqslant j\leqslant K$.

2）当 $t>1$ 时，将数据块 $\boldsymbol{X}^{t-1}$ 的质心 Δ^{t-1} 加权后加入数据块 $\boldsymbol{X}^t$ 中，得到新的数据块 $\boldsymbol{X}^{t'}=[\Delta^{t-1},\boldsymbol{X}^t]$. 对新的数据块进行聚类，相应权重更新为

$$\boldsymbol{w}'_c=\sum_{i=1}^{c+n_t}(u_{ci})[\boldsymbol{w}'_c,\boldsymbol{w}_c] \tag{15}$$

其中 $1\leqslant c\leqslant C, w_c=1, 1\leqslant i\leqslant n_t$.

最小化目标函数，对数据块 $\boldsymbol{X}^t$ 聚类后得到的隶属度 u_{ci} 和质心 v_{cj} 的公式为

$$u_{ci}=\frac{F_i D(x'_i,v_c)^{\frac{1}{1-m}}}{\sum_{t=1}^{C}\beta_{ti}D(x'_i,v_t)^{\frac{1}{1-m}}} \tag{16}$$

$$v_{cj}=\frac{\sum_{i=1}^{n_t+c} w'_i\beta_{ci}(u_{ci})^m H_{cij}x'_{ij}}{\sum_{i=1}^{n_t+c} w'_i\beta_{ci}(u_{ci})^m H_{cij}} \tag{17}$$

其中 $1\leqslant j\leqslant K, 1\leqslant i\leqslant n_t+c$.

SPFCOM 算法的聚类过程可用伪码描述如下.

1　确定簇数目 c 和权重指数 $m\in(1,\infty)$，选择 ε-敏感性相异性度量方法，初始化隶属度矩阵 $\boldsymbol{U}^{(0)}$，$\beta_{ci}=1$，$H_{cij}=1$，$F_i=1$，设置迭代次数 $z=1$；

2　将数据集划分为 h 块，即 $\boldsymbol{X}=[\boldsymbol{X}^1,\boldsymbol{X}^2,\cdots,\boldsymbol{X}^h]$，初始化权重 $\boldsymbol{w}=[1,1,\cdots,1]^{\mathrm{T}}$. 设置 $t=1$；

3　利用式(14)或(17)更新簇中心矩阵 $\boldsymbol{V}^{(z)}$；

4　计算残差 E_{cij}；

5　对残差 E_{cij} 进行排序；

6　计算综合典型性参数 F_i；

7　利用式(13)或(16)计算隶属度矩阵 $\boldsymbol{U}^{(z)}$；

8　① $t<h$：若 $\|\boldsymbol{U}^{(z)}-\boldsymbol{U}^{(z-1)}\|_{\mathrm{F}}>\xi$，令 $z\leftarrow z+1$ 并转到步骤 3；否则把质心和权重添加到下一个数据块的数据矩阵和权重矩阵中，获得新的数据块矩阵和权重矩阵，令 $t\leftarrow t+1$ 并且转到步骤 3；

② $t=h$：若 $\|\boldsymbol{U}^{(z)}-\boldsymbol{U}^{(z-1)}\|_{\mathrm{F}}>\xi$，令 $z\leftarrow z+1$ 并转到步骤 3；否则结束.

2.2　OFCOM 算法

基于 WFCOM 算法，采用 online 增量框架，设计了 OFCOM 算法. 在 OFCOM 算法中，同样地，初始化权重得到 $\boldsymbol{w}=[1,1,\cdots,1]^{\mathrm{T}}$，并将 h 个数据块记为 $\boldsymbol{X}=[\boldsymbol{X}^1,\boldsymbol{X}^2,\cdots,\boldsymbol{X}^h]$.

1）对 h 个数据块依次聚类，得到质心矩阵 $\boldsymbol{\Delta}^t=[v_1,v_2,\cdots,v_c]$ 和相应权重矩阵：

$$\boldsymbol{w}_c^t=\sum_{i=1}^{n_t}u_{ci}\boldsymbol{w}_i \tag{18}$$

其中 $1\leqslant c\leqslant C, w_i=1, 1\leqslant i\leqslant n_t$，$n_t$ 为第 t 个数据块的对象个数.

最小化目标函数，对数据块 $\boldsymbol{X}^t$ 聚类后隶属度 u_{ci} 和质心 v_{cj} 的表达式为

$$u_{ci}=\frac{F_i D(x_i,v_c)^{\frac{1}{1-m}}}{\sum_{t=1}^{C}\beta_{ti}D(x_i,v_t)^{\frac{1}{1-m}}} \tag{19}$$

$$v_{cj}=\frac{\sum_{i=1}^{n_t} w_i\beta_{ci}(u_{ci})^m H_{cij}x_{ij}}{\sum_{i=1}^{n_t} w_i\beta_{ci}(u_{ci})^m H_{cij}} \tag{20}$$

其中 $1\leqslant j\leqslant K$.

2）将每个数据块聚类后得到的质心和权重分别组成总的数据矩阵 $\boldsymbol{X}'$ 和权重矩阵 $\boldsymbol{w}'$，并对这 2 个矩阵再次聚类得到最终隶属度矩阵和质心矩阵. 分别记数据矩阵 $\boldsymbol{X}'$ 和权重矩阵 $\boldsymbol{w}'$ 为 $\boldsymbol{X}'=[\Delta^1,\Delta^2,\cdots,\Delta^h]$ 和 $\boldsymbol{w}'=[\boldsymbol{w}_1^1,\boldsymbol{w}_2^1,\cdots,\boldsymbol{w}_c^1,\boldsymbol{w}_1^2,\boldsymbol{w}_2^2,\cdots,\boldsymbol{w}_c^2,\cdots,\boldsymbol{w}_1^h,\boldsymbol{w}_2^h,\cdots,\boldsymbol{w}_c^h]$.

最小化目标函数,得到隶属度和质心的表达式:

$$u_{ci}=\frac{F_i D(x'_i,v_c)^{\frac{1}{1-m}}}{\sum_{t=1}^{C}\beta_{ti}D(x'_i,v_t)^{\frac{1}{1-m}}} \tag{21}$$

$$v_{cj}=\frac{\sum_{i=1}^{n_x} w'_i\beta_{ci}(u_{ci})^m H_{cij}x'_{ij}}{\sum_{i=1}^{n_x} w'_i\beta_{ci}(u_{ci})^m H_{cij}} \tag{22}$$

其中 $1\leqslant j\leqslant K$,$1\leqslant i\leqslant n_x$,$n_x$为数据个数.

OFCOM 算法的聚类过程可用伪码描述如下.

1 确定簇数目 c 和权重指数 $m\in(1,\infty)$. 选择 ε—敏感性相异性度量方法. 初始化隶属度矩阵 $\boldsymbol{U}^{(0)}$,$\beta_{ci}=1$,$H_{cij}=1$,$F_i=1$,质心矩阵 l 和权重矩阵 m,且设置迭代次数 $z=1$;

2 将数据集划分为 h 块,即 $\boldsymbol{X}=[\boldsymbol{X}^1,\boldsymbol{X}^2,\cdots,\boldsymbol{X}^h]$,初始化权重 $\boldsymbol{w}=[1,1,\cdots,1]^{\mathrm{T}}$. 设置 $t=1$;

3 利用式(20)或(22)更新簇中心矩阵 $\boldsymbol{V}^{(z)}$;

4 计算残差 E_{cij};

5 对残差 E_{cij}进行排序;

6 计算综合典型性参数 F_i;

7 利用式(19)或(21)计算隶属度矩阵 $\boldsymbol{U}^{(z)}$;

8 ① $t<=h$:若 $\|\boldsymbol{U}^{(z)}-\boldsymbol{U}^{(z-1)}\|_{\mathrm{F}}>\xi$,令 $z\leftarrow z+1$ 并转到步骤 3;否则把质心和权重添加到质心矩阵 $\boldsymbol{l}$ 和权重矩阵 $\boldsymbol{m}$ 中,令 $t\leftarrow t+1$ 并转到步骤 3;

② $t>h$:若 $\|\boldsymbol{U}^{(z)}-\boldsymbol{U}^{(z-1)}\|_{\mathrm{F}}>\xi$,令 $z\leftarrow z+1$ 并转到步骤 3;否则结束.

3 实验与分析

3.1 数据集介绍

实验数据集均来自于 UCI 数据库,各数据集具体信息见表 1. 为定量评估聚类结果的准确性,采用 F 度量(F-measure)[7]和熵(entropy)2 种评价标准.

表 1 实验数据集简介

数据集	样本/个	属性/个	类别/个
Breast Tissue(BT)	106	9	6
PID	768	8	2
Mammographic-Mass(Ma)	1 077	68	8
Waveform	5 000	21	3
Ohsumed(Oh)	1 000	500	10
Epileptic Seizure Recognition (ESR)	11 500	178	5

3.2 实验结果

为了验证 SPFCOM 和 OFCOM 算法的效果,在 UCI 数据库的数据集上进行了 2 组实验:第 1 组实验重点考查算法的准确率,并与 SPFCM、SPHFCM(single-pass hyperspherical fuzzy C means)[8]、OFCM 和 OHFCM(online hyperspherical fuzzy C means)[8]算法进行了比较;第 2 组实验验证算法的鲁棒性.

1) 算法准确率

实验中,通过 UCI 数据库的 6 个数据集评估算法的准确率. 为便于比较,采用与 FCOM 算法相同的参数设置. 将 3 种 single-pass 算法(SPFCM、SPHFCM、SPFCOM)和 3 种 online 算法(OFCM、OHFCM、OFCOM)分别进行比较. 为了考查算法在不同数据块块大小条件下的表现,将数据块大小分别设定为各数据集样本总数的 5%、10%、20% 和 50%. 各实验运行 10 次后取平均值,实验结果如图 1 ~4 所示.

从图 1 ~4 可以看出,在各数据集的实验结果中,当数据块大小分别取 5%、10%、20% 和 50% 时,SPFCOM 和 OFCM 算法均可以取得较高的聚类准确率,总体表现较优,用 F 度量和熵表示的平均聚类结果见表 2.

通过该组实验可知,SPFCOM 算法与 OFCOM 算法在实验数据集上获得了比对比算法更高或相当的聚类准确率,且部分数据集上提升幅度较为明显.

2) 算法鲁棒性

为了验证算法的鲁棒性,第 2 组实验在 BT 数据集上测试噪声数据对 SPFCOM 与 OFCOM 算法的影响. 将数量为 BT 数据集样本总数 10%、20%、30% 和 40% 的噪声样本分别添加到数据集中,实验结果如图 5 所示.

图 5(a)和图 5(b)分别对比了 SPFCM、SPHFCM 和 SPFCOM 3 种算法在 F 度量和熵两项指标上的表现. 当噪声数据逐渐增多时,3 种 single-pass 算法的 F 度量指标逐渐下降. 在加入 10% 的噪声样本时,3 种算法的 F 度量值均下降较多,其中 SPFCM 算法的下降幅度较大,而 SPHFCM 算法直接降至该算法的最低点,此时 SPFCOM 算法的 F 度量指标仍比其他 2 种算法分别提高了 6.6% 和 8.5%;继续加入噪声样本后,3 种算法的 F 度量指标则趋于稳定;在噪声样本增加至 40% 后,SPFCOM 算法 F 度量指标的提高幅度分别为 5.3% 和 6.3%. 在熵指标方面,如图 5(b),可以得到与 F 度量指标类似的实验结果;在加入 10% 的噪声点时,3 种算法的熵值上升

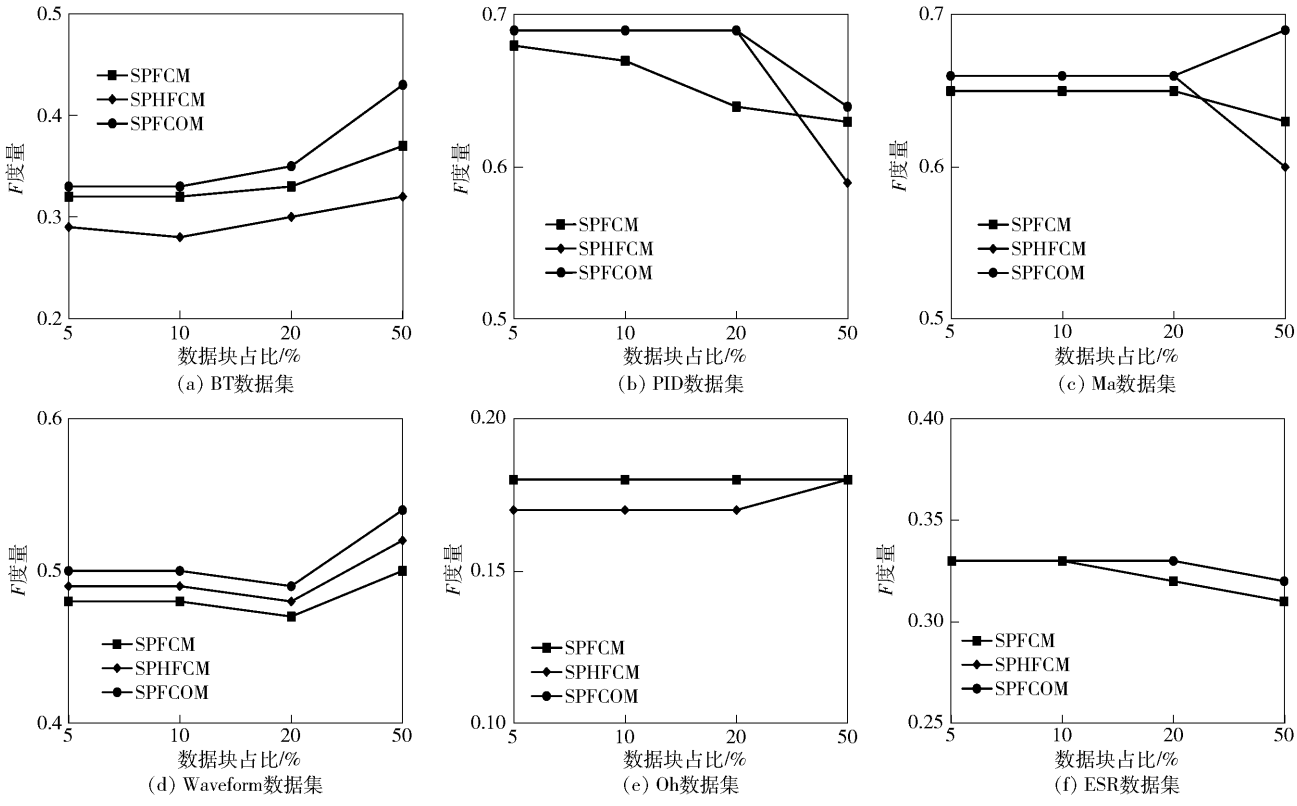

图 1　SPFCM、SPHFCM 和 SPFCOM 算法的 *F* 度量对比

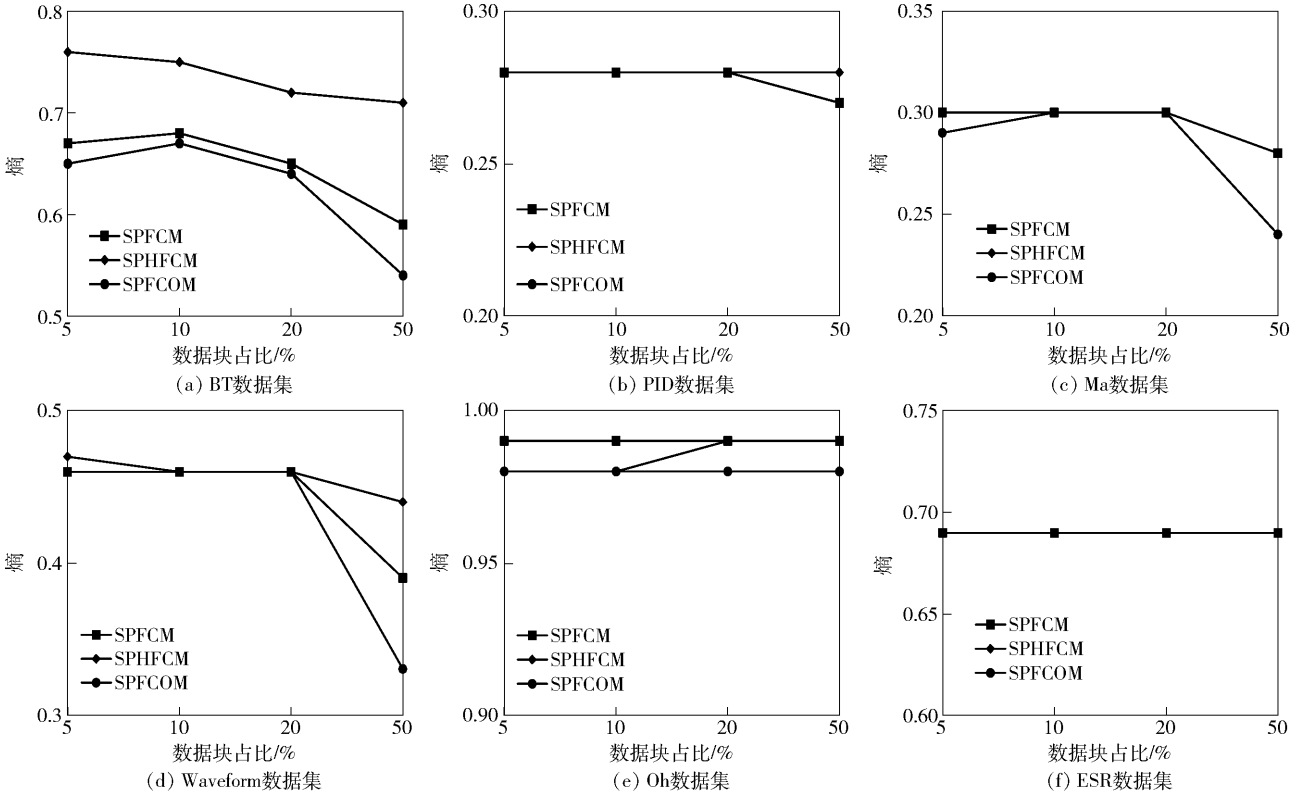

图 2　SPFCM、SPHFCM 和 SPFCOM 算法的熵对比

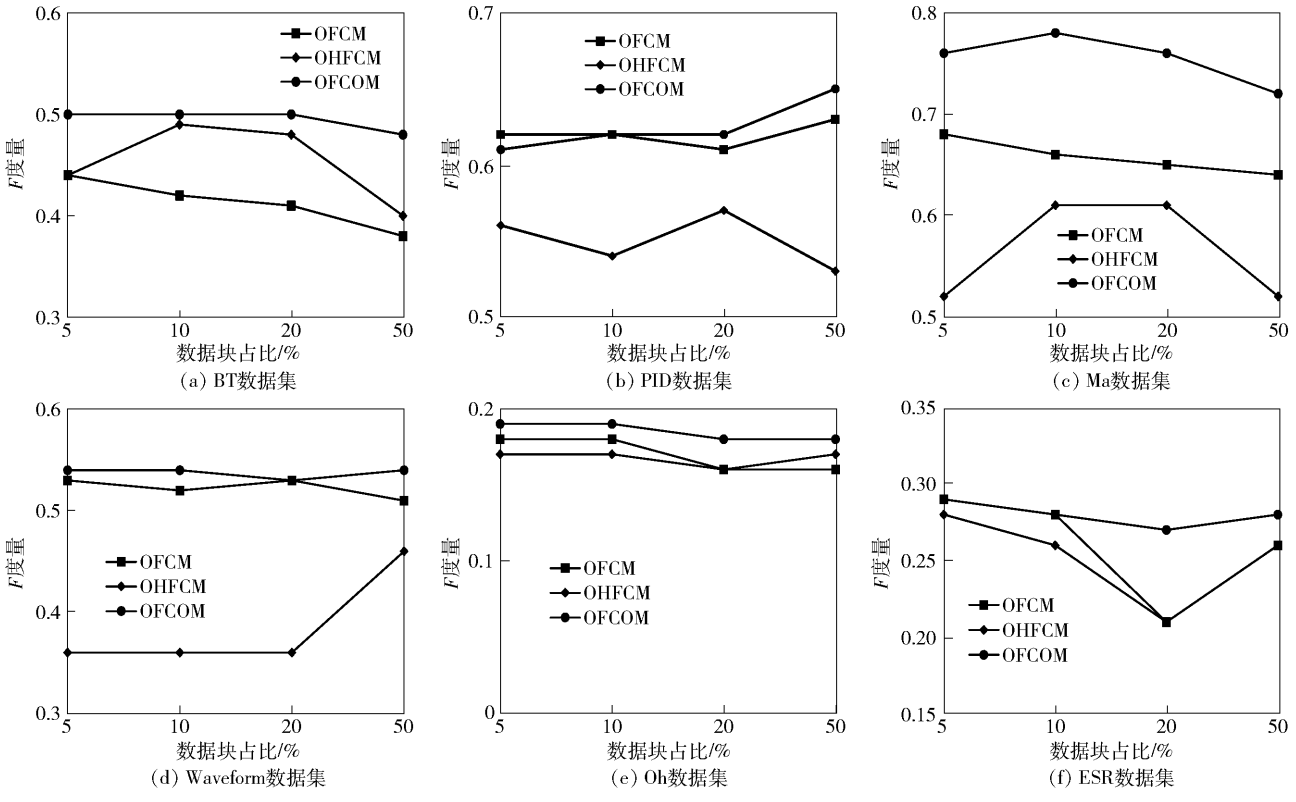

图3 OFCM、OHFCM 和 OFCOM 算法的 *F* 度量对比

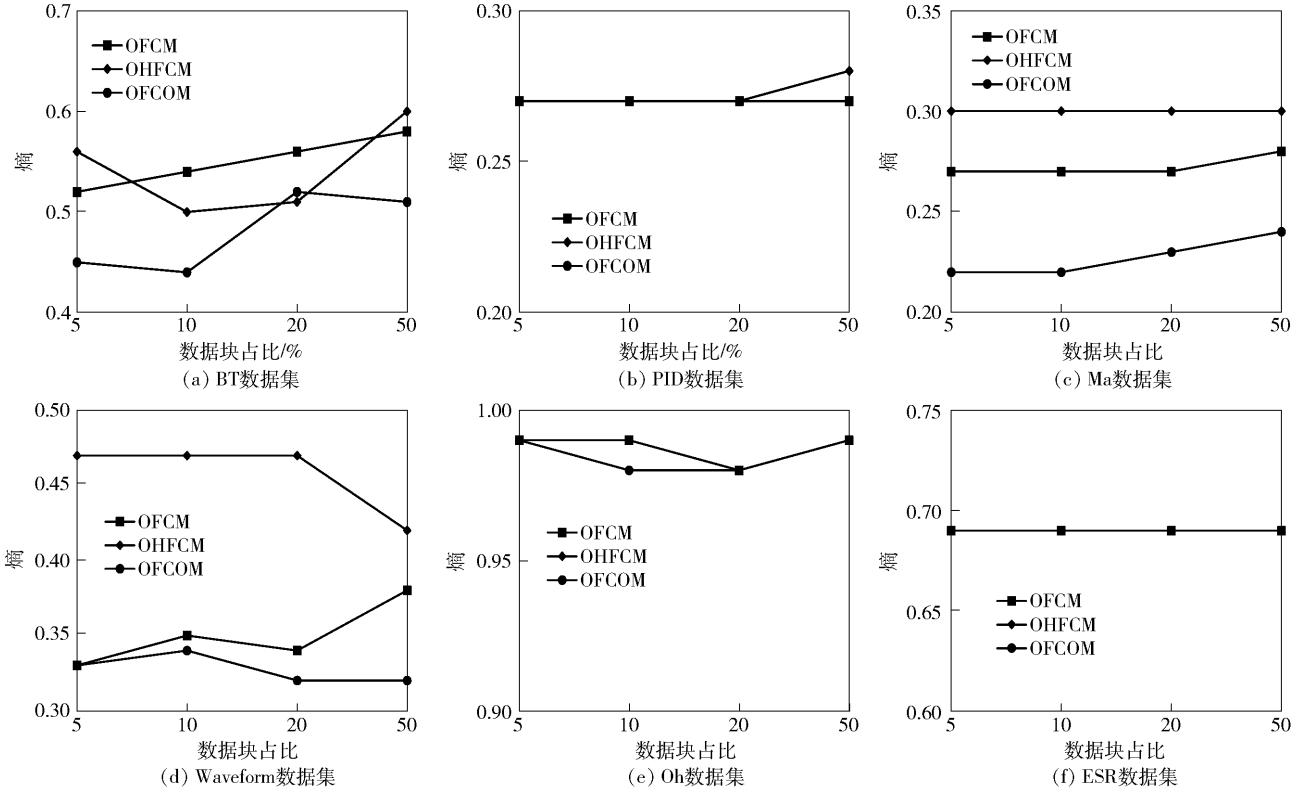

图4 OFCM、OHFCM 和 OFCOM 算法的熵对比

表 2　算法准确率比较

算法	*F* 度量		熵	
	均值	提高幅度/%	均值	降低幅度/%
SPFCM	0.410 8	3.04	0.614 2	1.29
SPHFCM	0.412 1	2.72	0.615 0	1.42
SPFCOM	0.423 3	—	0.606 3	—
OFCM	0.400 0	6.98	0.592 9	2.04
OHFCM	0.347 9	23.00	0.615 8	5.68
OFCOM	0.427 9	—	0.580 8	—

较多；继续添加噪声样本后，熵指标趋于平稳，直至噪声样本占比达到 40%；在此过程中，SPFCOM 算法的熵指标均低于对比算法.

图 5(c)和图 5(d)中给出了 OFCM、OHFCM 和 OFCOM 3 种算法的对比. 从图 5(c)可以看出，当噪声样本增多时，3 种算法的 *F* 度量指标逐渐降低；当加入 10% 噪声样本时，OFCM 和 OHFCM 算法的 *F* 度量指标均大幅降低，而 OFCOM 算法则变化较小，其 *F* 度量值相比其他 2 种算法分别高出 24.2% 和 8.7%；继续加入噪声样本，3 种算法的 *F* 度量指标则渐趋平稳；噪声样本数量达到 40% 时，OFCOM 算法 *F* 度量值的提高幅度分别为 16.6% 和 4.6%. 在图 5(d)给出的熵指标对比中，可得到相似结论：当加入 10% 的噪声样本时，OFCM 和 OHFCM 算法的熵指标逐渐上升，而 OFCOM 算法则变化较小，在 10% 处其熵值相比其他 2 种算法分别降低 22.2% 和 7.6%；继续添加噪声样本，3 种算法的熵指标则比较平稳，而 OFCOM 算法的熵指标一直表现最优；最终在 40% 噪声样本点处，OFCOM 的熵值比对比算法分别低 15.9% 和 3.0%.

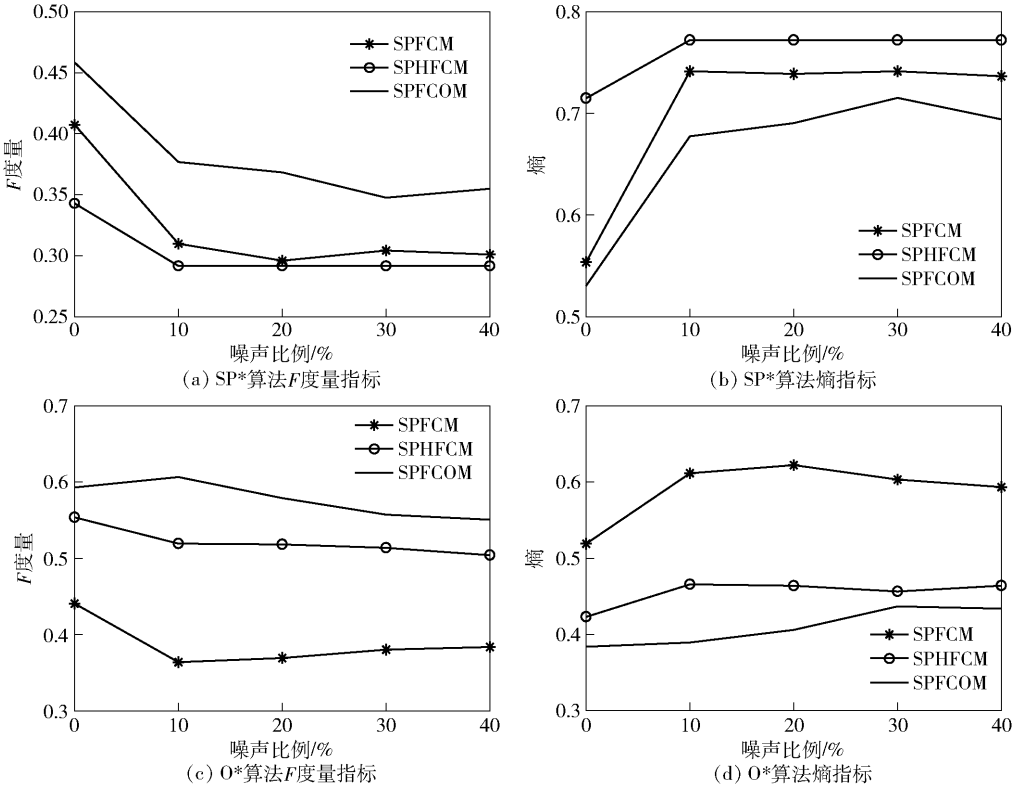

图 5　噪声对聚类的影响

本组实验结果表明，当数据集中存在噪声时，SPFCOM 和 OFCOM 具有较好的鲁棒性.

4　结束语

大数据时代背景下，是否具备处理大规模数据的能力并具备较高的鲁棒性，是衡量一个聚类算法聚类表现的重要标准. 鉴于传统算法在这 2 方面的劣势，提出了 2 种增量式模糊 C 有序均值聚类算法，即 SPFCOM 与 OFCOM. 这 2 种算法一方面引入排序机制，有效降低了对噪声数据的敏感度；另一方面分别采用了 single-pass 和 online 增量架构，能有效处理大规模数据. 为了评估算法的聚类效果，在 6 个真实数据集上进行了实验. 实验结果表明，相较于对比算法，SPFCOM 与 OFCOM 可以获得更高的

聚类准确率,同时具有更强的鲁棒性.

参考文献:

[1] Bezdek J C, Ehrlich R, Full W. FCM: the fuzzy C means clustering algorithm[J]. Computers & Geosciences, 1984, 10(2): 191-203.

[2] Huber P J. Robust statistics[J]. Journal of the American Statistical Association, 1981, 78(381): 1248-1251.

[3] Yager R R. On ordered weighted averaging aggregation operators in multicriteria decisionmaking[J]. Readings in Fuzzy Sets for Intelligent Systems, 1993, 18(1): 80-87.

[4] Leski J M. Fuzzy c-ordered-means clustering[J]. Fuzzy Sets & Systems, 2016(286): 114-133.

[5] Hore P, Hall L O, Goldgof D B. Single pass fuzzy C means[C]// IEEE International Conference on Fuzzy Systems. London: IEEE, 2007: 1-7.

[6] Hore P, Hall L O, Goldgof D B, et al. Online fuzzy C means[C]// NAFIPS 2008. New York: IEEE, 2008: 1-5.

[7] Maratea A, Petrosino A, Manzo M. Adjusted *F*-measure and kernel scaling for imbalanced data learning[J]. Information Sciences, 2014, 257(2): 331-341.

[8] Mei J P, Wang Y, Chen L, et al. Incremental fuzzy clustering for document categorization[C]// IEEE International Conference on Fuzzy Systems. Beijing: IEEE, 2014: 1518-1525.