

一种改进的模糊聚类有效性指标

王丽娜, 马晓晓

(南京信息工程大学 电子与信息工程学院, 江苏 南京 210044)

摘 要: 在先前指标的基础上提出了一种改进的聚类有效性指标, 该指标以紧密性和分离性之比定义, 不仅体现了数据的隶属度和几何结构的相关信息, 同时也反映了数据的分布情况. 实验表明, 该指标可以使数据集实现正确划分, 具有较高的可靠性.

关键词: FCM; 模糊聚类; 聚类有效性; 评价指标

中图分类号: TP391

文献标识码: A

文章编号: 1000-7180(2014)04-0068-03

An Improved Validity Index for Fuzzy Clustering

WANG Li-na, MA Xiao-xiao

(School of Electronic and Information Engineering,
Nanjing University of Information Science & Technology, Nanjing 210044, China)

Abstract: An improved validity index is proposed in this paper, which is based on previous indices. This improved index, defined by the ratio of separation measure and compactness measure, not only contains the information of data membership degree and data structure, but also shows the character of data distribution. Experiments results prove that the improved index can make a correct partition of datasets. Compared with other indexes, it is relatively more effective.

Key words: Fuzzy C-Means (FCM); Fuzzy clustering; Clustering validity; Evaluating indicator

1 引言

聚类分析是一种无监督的学习方法, 由于它对数据对象进行分类时, 并没有依据任何的先验知识, 因此需要对聚类结果进行有效性的判定. 人们将对聚类结果的评估称为聚类有效性问题, 利用评价指标, 并基于迭代的过程从而确定最佳聚类数或者最佳聚类结果. 解决聚类问题有效性的途径是构造有效性函数, 即评价指标, 并利用评价指标对聚类结果进行定量研究^[1-3].

目前, 在模糊聚类的领域里已提出很多评价指标^[4-6], 这些评价价值指标从定义上大致可分为三类, (1) 基于数据集的模糊划分的评价指标; (2) 基于数据集的几何结构的评价指标; (3) 基于数据集的统计信息的评价指标. 目前为止没有一种指标普遍适用

于所有领域, 因此寻找更加有效的评价指标仍然是重要研究课题之一.

本文提出了一个改进的模糊聚类评价指标. 通过实验表明, 该指标性能稳定, 可以有效处理类间有重叠的数据, 具有较高的可靠性.

2 FCM 算法

2.1 FCM 算法理论基础

随着模糊集理论的提出, 模糊 C 均值算法是一种运用逐步迭代原理的基于目标函数的聚类算法, 此算法的目标是通过将目标函数 J 最小化, 从而将给定的数据集 X 划分成 c 类.

设给定的数据集 $X = \{x \mid x_1, x_2, \dots, x_n\}$, n 为数据集中的样本个数. 样本 x_j 具有 s 个属性, 将 X 划分为 c 类, 即有 c 个聚类中心 $V = \{v \mid v_1, v_2, \dots, v_c\}$,

收稿日期: 2013-06-03; 修回日期: 2013-07-30

基金项目: 国家自然科学基金项目(61203273)

其中 $2 \leq c \leq n$, 将样本 x_j 与聚类中心 v_i 的欧氏距离定义为 d_{ij} , 第 j 个样本在第 i 类中的隶属度定义为 u_{ij} . FCM 函数的目标函数如式(1).

$$\min J(X, U, V) = \sum_{j=1}^n \sum_{i=1}^c u_{ij}^m (d_{ij})^2 \quad (1)$$

式中, $0 \leq u_{ij} \leq 1$, $\sum_{i=1}^c u_{ij} = 1$, m 是模糊划分系数.

FCM 算法具体步骤入如下:

(1) 初始化: 聚类数 c ($2 \leq c \leq n$), 模糊划分系数 m (不做特殊要求通常取 2), 迭代终止阈值 ϵ (通常在 0.001 和 0.01 之间) 以及聚类中心 V , 迭代次数 $l = 1$.

(2) 按式(2), 基于最优化方法计算划分矩阵.

$$u_{ij}^{(l)} = 1 / \left(\sum_{k=1}^c \left(\frac{d_{ij}^{(l)}}{d_{kj}^{(l)}} \right)^{\frac{1}{m-1}} \right) \quad (2)$$

(3) 按式(1), 基于最优化方法计算聚类中心.

$$v_i^{(l+1)} = \frac{\sum_{j=1}^n (u_{ij}^{(l)})^m \times x_j}{\sum_{j=1}^n (u_{ij}^{(l)})^m} \quad (3)$$

$i = 1, 2, \dots, c, \forall j$

(4) 如果 $\|V^{(l+1)} - V^{(l)}\| < \epsilon$ 则停止迭代, 否则 $l = l + 1$, 将转向步骤 2.

2.2 基于 FCM 算法的聚类有效性

聚类有效性主要从两个方面来评价, 一是类内的对象尽可能相似, 即具有较高的紧密性, 二是类间具有较大的相异度, 即具有较高的分离性. 聚类有效性实际上是求解最佳聚类数 c 的问题. 对于给定的数据集 X 使用不同的参数运行特定的算法, 从而将给定的数据集划分成不同的类, 计算相应的划分下指标所对应的值. 观察分析各个指标值变化时, 找出在指标达到预定值状态下的聚类数 c , 此条件下的聚类数为寻找的最优划分. 利用 FCM 算法求解最佳聚类数的流程如图 1 所示.

3 新指标的提出

在当前各类模糊聚类评价指标中, 基于模糊划分的指标仅包含样本的隶属度, 而基于几何结构的评价指标包含数据的几何结构信息, 基于统计信息的评价指标反应数据的分布情况. 本文结合以上的指标提出了一个新的评价指标, 此项指标也是基紧密性和分离性的基础而定义的. 其定义如式(4)所示.

$$V_{\text{new}}(U, V, c) =$$

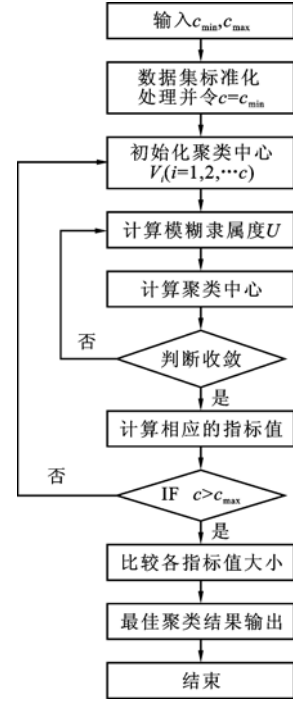


图 1 基于 FCM 算法求解最佳聚类数 c 的流程图

$$\frac{\sum_{i=1}^c \sum_{j=1}^n u_{ij}^m \|v_i - x_j\|^2 + \frac{1}{n} \sum_{j=1}^n \|x_j - \bar{v}\|^2}{\frac{1}{c} \sum_{i=1}^c \sum_{j=1}^n u_{ij}^m \|v_i - \bar{v}\|^2 + \frac{1}{c}} \quad (4)$$

由公式可以看出, 分母前半部分 $\sum_{i=1}^c \sum_{j=1}^n u_{ij}^m \|v_i - \bar{v}\|^2$ 表示类间总变差, 也是分离性度量. 对于一个好的模糊聚类而言, 分离性越大越好, 使得不同类间的样本相似性最小. 分子中 $\sum_{i=1}^c \sum_{j=1}^n u_{ij}^m \|v_i - x_j\|^2$ 是从 FCM 算法中目标函数借鉴而来, 从样本间的类内紧密性定义, 表示类内总变差, 紧密性越小分类效果越好. 类内的总变差和类间的平均距离表示了数据的分布情况. 因此此项指标从定义上既包含样本了隶属度, 几何结构, 也包含了数据分布情况的信息. 其值越小, 聚类效果越好.

在 XB 指标中, 当聚类数趋于 c 时候, 指标值减为 0, 因此需引入一个惩罚项来抑制指标失效的情况.

$$\lim_{c \rightarrow n} V_{\text{XB}}(U, V, c) = \frac{\frac{1}{n} \sum_{i=1}^c \sum_{j=1}^n u_{ij}^m \|v_i - x_j\|^2}{\min_{i \neq t} \|v_i - v_t\|^2} = 0$$

因此, 在文中所提出的新指标中也引入了惩罚项 $\frac{1}{n} \sum_{j=1}^n \|x_j - \bar{v}\|^2$, 防止当聚类数 c 趋近样本数 n 时, 指标值减为 0 从而失效. 所以本文提出的新指标

有效地弥补了 XB 的不足,证明如下.

$$\begin{aligned}
 & \lim_{c \rightarrow n} V_{\text{new}}(U, V, c) \\
 &= \lim_{c \rightarrow n} \frac{\sum_{i=1}^c \sum_{j=1}^n u_{ij}^m ||v_i - x_j||^2 + \frac{1}{n} \sum_{j=1}^n ||x_j - \bar{v}||^2}{\frac{1}{c} \sum_{i=1}^c \sum_{j=1}^n u_{ij}^m ||v_i - \bar{v}||^2 + \frac{1}{c}} \\
 &= \frac{c \sum_{j=1}^n ||x_j - \bar{v}||^2}{n \sum_{i=1}^c \sum_{j=1}^n u_{ij}^m ||v_i - \bar{v}||^2 + n} \\
 &\neq 0.
 \end{aligned}$$

同 Kwon 指标引入的惩罚项比较,当聚类数逐渐增大时,我们取极限如下所示:

$$\begin{aligned}
 & \lim_{c \rightarrow n} V_{\text{Kwon}}(U, V, c) \\
 &= \lim_{c \rightarrow n} \frac{\sum_{i=1}^c \sum_{j=1}^n u_{ij}^m ||v_i - x_j||^2 + \frac{1}{c} \sum_{i=1}^c ||v_i - \bar{v}||^2}{\min_{i \neq t} ||v_i - v_t||^2} \\
 &= \frac{\frac{1}{c} \sum_{i=1}^c ||v_i - \bar{v}||^2}{\min_{i \neq t} ||v_i - v_t||^2}.
 \end{aligned}$$

随着聚类数的增加,分母在减小,而分子无法判断,因此 Kwon 指标无法判定其是增大还是减小,所以不能很好的抑制指标值减为 0. 理论上讲,我们希望指标值是递增的,从而抑制指标值减小为 0 的趋势. 但是通过 Kwon 指标发现引入的惩罚项并不能很好的抑制指标值减为 0,所以我们考虑引入常数项. 文中新指标的惩罚项,表示所有样本到平均聚类中心的距离,所以对于同一数据集而言,此项为一个常量. 对于 XB 和 Kwon 指标而言,当 m 趋于无穷时,指标值趋于无穷,从而导致不稳定. 新指标分母中的第二项是一项惩罚项,目的在于防止模糊因子 m 趋于无穷时,指标值不稳定.

$$\begin{aligned}
 & \lim_{m \rightarrow \infty} u_{ij} = 1/c \\
 & \lim_{m \rightarrow \infty} v_i = \sum_{j=1}^n x_j / n = \bar{v} \\
 & \lim_{m \rightarrow \infty} V_{\text{new}}(U, V, c) \\
 &= \lim_{m \rightarrow \infty} \frac{\sum_{i=1}^c \sum_{j=1}^n u_{ij}^m ||v_i - x_j||^2 + \frac{1}{n} \sum_{j=1}^n ||x_j - \bar{v}||^2}{\frac{1}{c} \sum_{i=1}^c \sum_{j=1}^n u_{ij}^m ||v_i - \bar{v}||^2 + \frac{1}{c}} \\
 &= \frac{c \sum_{j=1}^n ||x_j - \bar{v}||^2}{n \sum_{j=1}^n ||x_j - \bar{v}||^2}
 \end{aligned}$$

但实际应用中,样本数往往远大于聚类数,所以在实验选取聚类数的范围往往限定于一个接近聚类

数的邻近范围.

4 实验

为了说明各项列举指标以及验证新指标的可靠性,利用 MATLAB 通过仿真实验在一些公众的数据集给出实验测试结果. 首先设定 $c_{\min} = 2, c_{\max} = \min(10, \sqrt{n})$ 中的最小值以保证样本数和聚类数之间的平衡, $\epsilon = 0.001$, 最大迭代次数为 100. 分别用四个不同的数据集对新指标进行验证,并与 PC, PE, XB, Kwon, FPF 指标值进行比较.

实验结果表明本文提出的新指标的最小值所取得最佳聚类数与数据集的真实最佳聚类数一致,因此该指标可以将数据集实现正确划分,从而证明了新指标的可靠性.

表 1 所示为本文列举的五种指标与新指标的结果分析,对于 IRIS 数据集和 Seeds 数据集而言,只有新指标可以将其分成三类,其余指标都只能将其分成两类,对于 Ionosphere 数据集,所有指标均可以实现正确划分,而在 Connectionist Bench 数据集上,只有 Kwon 指标不能将其最优正确划分. 从实验数据看出,本文提出的新指标相对于其他指标,具有较高的可靠性,可以将四个数据集实现最优划分.

表 1 不同数据集上的不同指标值的结果比较

数据集	IRIS	Ionosphere	Connectionist Bench	Seeds
最佳聚类数	3	2	2	3
PC	2	2	2	2
PE	2	2	2	2
XB	2	2	2	2
Kwon	2	3	4	2
FPF	2	2	2	2
new	3	2	2	3

5 结束语

本文提出的新指标实现了数据集正确的划分,具有较高的可靠性,与其他指标相比,该指标包含了更多的信息,不仅考虑了数据隶属度以及数据的几何结构,同时也反映了数据的分布情况,对于类间有重叠的数据集可以实现最佳划分. 对于本文所提出的新指标,在理论分析和实验方面需要开展深入的工作,从而使得新指标适用于更多的领域研究.

参考文献:

- [1] 唐明会,杨燕. 模糊聚类有效性的研究进展[J]. 计算机工程与科学, 2009, 31(9): 122-124.

(下转第 74 页)

于密云计算更加完备的服务,如信道保护、信息溯源、大数据中信息的归集等.在密云计算环境中,传输信道自身也被授予唯一的密钥标识,可根据需求用自身私钥对当前传输数据签名,实现对数据真实性的认证,避免了信道劫持等不安全因素.对于重要信息,发送者和经手者均需要签名.通过对传输数据签名的分析,可实现对信息的溯源.此外,网络中任意的信息都可以密钥为依据连接起来,用户可在复杂的网络大数据中方便地找到与自己相关的数据,即以密钥实现信息的归集.另一方面由于有标识密钥的保护,需要保护的信息不会被外界所破解.这一特性在网络大数据时代尤为重要.

4 结束语

文中介绍了一种新型的安全云计算框架,即为基于自认证密钥的密云计算.这一框架可实现原生的云计算环境中数据的安全服务,即“密即服务”,为云计算数据安全共享传输提供了有力地保证.本文的密云计算构建于自认证非对称密钥技术.这样密服务的实现,包括用户与服务器云端、用户与用户间的操作可真正以点对点的方式建立互信.这一过程中,公钥的获取不需要类似 CA 认证中心这样的机构,而是通过本地数据库查询或者自动生成等方式获得,从而在提高网络资源应用效率的同时,更是降低了非认证用户的入侵或第三方劫持的可能性.基于自认证的密云计算除了网络安全的优势,其基于标识密钥的数据标定方式还可以实现在大数据中准

确的获取有效信息,方便实现信息归集和溯源,实现网络大数据的高效处理.

参考文献:

- [1] 张建勋,古志民,郑超.云计算研究进展综述[J]. 计算机应用研究,2010,27(2):429-433.
- [2] Armbrust M, Fox A, Griffith R, et al. A view of cloud computing[J]. Communications of the ACM, 2010, 53(4):50-58.
- [3] Brodtkin J. Gartner: Seven cloud- computing security risks[N]. Net world, 2008-6-2.
- [4] 张健. 全球云计算安全研究综述[J]. 电信网技术, 2010,9:15-18.
- [5] Carlin S, Curran K. Cloud Computing Security[J]. International Journal of Ambient Computing and Intelligence, 2011, 3(1):14-19.
- [6] 李勇,吴作顺,刘秋彩. 基于 CPK 的融合网络密钥建立与认证[J]. 计算机应用, 2009,29(S2):72-75.
- [7] Tan WZ, Yang WS, Yang MJ, et al. A modification on public key infrastructure application[C]//E-Business and Information System Security. Wuhan,2009.

作者简介:

刘立男,(1975—),男,博士,副教授,硕士生导师.研究方向为信息安全、物联网技术、计算机视觉.

尹召芳女,(1988—),硕士研究生.研究方向为物联网技术、信息安全.

王建男,(1976—),博士,讲师.研究方向为彩色图像处理、计算机视觉.

(上接第 70 页)

- [2] 洪志令,姜青山,董槐林,等.模糊聚类中判别聚类有效性的新指标[J]. 计算机科学,2004, 31(10): 121-125.
- [3] 孔攀.模糊聚类分析及其有效性研究[D]. 重庆:西南大学,2009.
- [4] Krista Rizman Žalik, Borut Žalik. Validity index for clusters of different sizes and densities[J]. Pattern Recognition Letters,2011, 32(2): 221-234.
- [5] Witold Pedrycz. A dynamic data granulation through

adjustable fuzzy clustering [J]. Pattern Recognition Letters, 2008, 29(16): 2059-2066.

- [6] Zhang Yunjie, Wang Weina, Zhang Xiaona, et al. A cluster validity index for fuzzy clustering [J]. Information Sciences,2008, 178(4): 1205-1218.

作者简介:

王丽娜女,(1979—),博士研究生,讲师.研究方向有数据挖掘、机器学习、人工智能.