

基于噪点抑制的聚类有效性评价函数构建^{*}

翟丽丽, 张 雪, 彭定洪, 李艳来

(哈尔滨理工大学 高新技术产业发展研究中心, 哈尔滨 150040)

摘 要: 针对传统聚类有效性评价函数中没有利用到数据集结构信息和噪点删除过量等问题, 提出一种新的聚类有效性评价函数。该函数由紧密性度量与分离性度量组成, 在紧密性度量中加入距离函数表示数据集几何结构, 避免单一理论给评价带来的不全面性; 在分离性度量中, 设定距离临界值 L , 与原有的隶属度临界值 T 两者之间的相互约束, 减少删除噪点的数量, 避免因数据信息丢失对评价结果造成的不准确性。最后, 将新构建的评价函数与原函数进行对比实验, 结果表明该方法具有更好的适用性。

关键词: 聚类分析; 有效性评价函数; 紧密性度量; 分离性度量; 噪点抑制

中图分类号: TP311 **文献标志码:** A **文章编号:** 1001-3695(2014)01-0037-03

doi: 10.3969/j.issn.1001-3695.2014.01.007

Construction of clustering validity evaluation function based on noise suppression

ZHAI Li-li, ZHANG Xue, PENG Ding-hong, LI Yan-lai

(High-tech Industrial Development Research Center, Harbin University of Science & Technology, Harbin 150040, China)

Abstract: As traditional clustering validity evaluation function did not take advantage of the structure information of the data set, and deleted excessive noise, this paper designed a new clustering validity evaluation function. The function was composed by the tightness measure and separability measure, and distance function was added to the tightness measure to represent the geometric structure of the data set, to avoid being not comprehensive if evaluated by a single theory. In separability measure, it set the distance threshold L and being mutual restraint on original membership threshold T to reduce the amount of noise deleted, to avoid inaccuracies that was caused by loss of data information on the evaluation results. Finally, the new building of the evaluation function compared with the original function of experimental results show that the proposed method has better applicability.

Key words: cluster analysis; effectiveness evaluation function; tightness measure; separability measure; noise

聚类有效性评价是对聚类结果优劣进行评价的方法, 根据有效性函数评价准则不同, 聚类有效性评价方法可分为内部评价法、相关评价法和外部评价法^[1]。根据评价理论的不同, 聚类有效性评价又可分为基于数据集模糊划分的方法、基于数据集几何结构的方法和基于数据统计信息的方法^[2]。聚类有效性评价主要采用类内紧密度与类间分离度作为评价指标。类内各元素相似度越大, 类内紧密度越好; 类间各元素相似度越小, 分离度越好。

在构建聚类有效性评价函数过程中, 所运用的理论知识不同, 则紧密度和分离度的表达形式不同, 得到的有效性评价函数亦不同, 如 Dunn 函数(DI)^[3]、I 函数^[4]、CH^[5] 函数等是根据数据集几何结构理论设计不同的距离来表达类内紧密度和类间分离度。基于数据集几何结构的聚类有效性评价函数能够很好地利用数据结构, 但计算量大, 表述复杂。随着聚类分析方法的增多, 模糊聚类的出现, 几何结构的评价方法无法全面地评价聚类结果的优劣。Bezdek^[6] 提出的划分系数(PC)和 Shannon 的划分熵(PE)^[7]、Windham^[8] 的比例系数等指标是基于模糊划分, 利用概率理论中的隶属度作为评价指标, PC 和 PE 指标具有很好的数学性质, 但也存在着随类数目 c 的增加而单调递增或递减的趋势, 仅利用到数据集本身, 缺乏与数据

几何结构的直观联系。目前, 多理论的结合已成为聚类有效性评价函数研究的热点。Zahid 等人^[9] 将模糊紧密度和模糊分离度的概念引入到传统的有效性指标中, 得到 SC 指标。Xie 等人^[10] 首先结合了数据集几何特征和模糊理论, 构建了 V_{xie} 指标。唐明会等人^[11] 在分析传统有效性评价函数基础上, 总结了基于数据集几何结构的模糊聚类有效性函数的研究进展。

根据 SVM 特性, 把对分类敏感的并且降低分类精度的样本点视为噪点, 它们位于异类最近点处^[12]。由于噪点的存在, 影响了聚类评价的效果, 本文对文献[13]提出的聚类有效性评价函数(下文简称为 MPO 函数)进行改进, 加入数据集几何结构特性, 在紧密型指标中增加距离临界值 L 并给出其约束条件, 构建新的聚类有效性评价指标。该函数有效地避免了噪点对聚类结果检验的影响, 减少了剔除噪点的个数, 避免数据丢失, 并能更加准确地得到类与类之间的噪点。

1 MPO 聚类有效性评价函数

在噪声环境进行有效性分析时, 考虑聚类的紧密性与分离性时也将噪声点包含进来, 使得有效性指标对噪声和野值敏感。MPO 聚类的有效性评价函数由紧密性度量与分离性度量

收稿日期: 2013-05-02; **修回日期:** 2013-06-18 **基金项目:** 国家自然科学基金资助项目(71072085, 71272191); 黑龙江省研究生创新基金重点资助项目(YJSCX2011-003HLJ); 黑龙江省高等学校哲学社会科学创新团队建设计划资助项目(TD201203)

作者简介: 翟丽丽(1963-), 女, 教授, 硕士, 主要研究方向为管理信息系统、电子商务、虚拟企业(zhailili@126.com); 张雪(1988-), 女, 硕士研究生, 主要研究方向为决策方法; 彭定洪(1982-), 男, 博士研究生, 主要研究方向为评价方法; 李艳来(1974-), 女, 副教授, 主要研究方向为综合评价方法、数据挖掘等。

组成,紧密性度量由模糊隶属度矩阵和聚类数目共同确定,代表类内的紧密程度;分离性度量定义为不同模糊集合间的距离,代表不同聚类之间的分离程度。通过对 PC 指标的改进,考虑 PC 指标中受聚类数目 c 的增加而呈现单调趋势的问题,引入了 u_M 和 $\left(\frac{c+1}{c-1}\right)^{1/2}$ ($2 \leq c \leq n$) 为调整指标,减少聚类数目变化对结果的影响,得到紧密性度量 $\text{com}(U, c)$ 。

$$\text{com}(U, c) = \left(\frac{c+1}{c-1}\right)^{1/2} \sum_{i=1}^c \sum_{j=1}^n u_{ij}^2 / u_M \quad (1)$$

其中: u_{ij} 为第 j 个元素属于第 i 个类的隶属度; $u_M = \min_{1 \leq i \leq c} \left(\sum_{j=1}^n u_{ij}^2 \right)$, 度量类 i 相对于最紧密类的紧密度; $\text{com}(U, c)$ 代表类内数据的紧密程度,其值越大,代表获得的模糊划分越好。

为了在含有噪点和野值的聚类环境下得到正确的划分,得到分离性度量方法。

$$\text{sep}(U, c) = \frac{1}{n} \sum_{j=1}^n \left(\sum_{a=1}^{c-1} \sum_{b=a+1}^c o_{abj}(U; c) \right) \quad (2)$$

其中: $o_{abj}(U; c) = \begin{cases} 1 - |u_{aj} - u_{bj}| & \text{如果 } |u_{aj} - u_{bj}| \geq T, a \neq b, \\ 0 & \text{其他} \end{cases}$,

u_{aj} 和 u_{bj} 表示第 j 个元素分别到 a 、 b 两类的隶属度值。

在 $o_{abj}(U; c)$ 中,应用阈值 T 来排除聚类边界中那些含糊不定的数据点^[14],而噪点正是这种情况。 $\text{sep}(U, c)$ 是通过隶属度矩阵计算得到的数据集中所有数据点分离度的总和, $\text{sep}(U, c)$ 值越小,表明聚类划分越分明。

MPO 函数为紧致性度量和分离性度量的差:

$$\text{MPO}(U, V) = \text{com}(U, c) - \text{sep}(U, c) \quad (3)$$

文献中通过与多种聚类有效性指标,包括 PC、FS、XB、OS、PACES、CO、W 指标对比得到, MPO 能很好地确定聚类个数,并能避免噪点对数据集的影响。但是, MPO 函数中没有利用到数据集的结构信息,对距离不敏感。本文主要在 MPO 基础上加入了数据集几何结构特性,避免了单一理论对检验结果的影响;在分离性度量中,删除噪点需要十分小心,删除的力度过大会造成数据的丢失。本文在 $\text{sep}(c)$ 中加入一个新的临界值 L , 与原有临界值 T 共同约束噪点范围,避免因数据丢失而造成的结果不准确性。

2 新的聚类有效性评价函数构建

2.1 紧密性度量

本文考虑 MPO 函数与数据集几何结构缺乏直接联系,不能全面评价聚类效果的缺点,将反映聚类内部紧密性程度的几何结构信息融入到紧密性度量 $\text{com}(U, c)$ 中,得到新的紧密度指标 $\text{com}'(U, c, d)$ 。

$$\text{com}'(U, c, d) = \sum_{i=1}^c \frac{\sum_{j=1}^n u_{ij}^2}{i u_M d_M} \left(\frac{c+1}{c-1} \right)^{1/2} \quad (4)$$

其中: d_M 为欧式距离,改进后的函数既包含了 x_j 对象到聚类中心 v_i 的距离,也包含了 x_j 对类 i 的隶属函数值; $\text{com}'(U, c, d)$ 值反映了各聚类内数据点的紧密性程度,其结果数值越大表明聚类内各元素越紧密,划分效果越好。

2.2 基于噪点双重抑制的分离性度量

聚类过程中,噪点和孤立点对聚类结果的影响尤为严重,文献^[13]中提到了 MPO 函数中用隶属度临界值 T 能够在概率条件下初步分离出噪点,然而,仅应用概率约束来界定噪点

的范围可能会造成数据的丢失。为了更加准确地确定噪点,本文在隶属度临界值 T 基础上增加了几何结构的临界值 L ,两个临界值共同确定噪点,有效地避免了单一原理下数据的丢失。结合概率条件和几何结构新构建的分离性度量如下:

$$\text{sep}'(U, c, d) = \frac{1}{n \sum_{j=1}^n \sqrt{\frac{1}{c} \sum_{i=1}^c d_{ij}^2}} \sum_{j=1}^n \left(\sum_{a=1}^{c-1} \sum_{b=a+1}^c W(U; c, d) \right) \quad (5)$$

其中:

$$W(U; c, d) = \begin{cases} 0 & |u_{aj} - u_{bj}| \leq T \text{ 且 } |d_{aj} - d_{bj}| \leq L_j, a \neq b \\ 1 - |u_{aj} - u_{bj}| & \text{其他} \end{cases} \quad (6)$$

设定两个临界值 T 和 L ,用来排除聚类边界中那些含糊不定的点,根据聚类数据分布的结构不同, T 和 L 取值可由专家给出,也可根据函数自行定义。本文通过大量数据实验,初步给出取值为 $T=0.01$, $L_j = \sqrt{\frac{1}{c} \sum_{i=1}^c d_{ij}^2}$ (d_{ij} 表示第 j 个元素属于第 i 个类的距离), $W(U; c, d)$ 为对给定数据点的分离度。 $\text{sep}'(c)$ 值越小,得到聚类划分效果越明显。

2.3 聚类有效性函数构建

将紧密性度量与分离性度量进行整合,得到新的聚类有效性评价函数为

$$\text{VS}(U, V) = \frac{\text{com}'(U, c, d)}{\text{sep}'(U, c, d)} \quad (7)$$

一个好的聚类要求较大的 $\text{com}'(U, c, d)$ 和较小的 $\text{sep}'(U, c, d)$ 。 $\text{VS}(U, V)$ 定义有效性函数, $\text{VS}(U, V)$ 值越大表明聚类划分结果越好。

3 算例及结果分析

3.1 算例 1

运用文中新建的聚类有效性函数,对我国 31 个省、市、自治区和直辖市 12 个月的平均气温的模糊聚类结果^[15]进行有效性评价。利用 MATLAB 求得类中心坐标矩阵 c_j , 隶属度矩阵 u_{ij} , x_i 对象到聚类中心 c_i 的欧式距离 d_M , 分别计算 MPO 聚类有效性评价函数和 VS 聚类有效性函数,得到结果如表 1、2 所示。

表 1 VS(U, V)算例结果

c	$\text{com}'(c)$	$\text{sep}'(c)$	$\text{VS}'(U, V)$
3	1.543 3	2.36 2	0.65 3
4	1.502 1	5.25 5	0.28 6
5	1.363	4.269	0.319

表 2 MPO(U, V)算例结果

c	$\text{com}(c)$	$\text{sep}(c)$	$\text{VS}(U, V)$
3	4.674 1	2.06	2.614
4	6.076	5.082	0.994
5	6.886 1	4.438	2.448

当 $c=3$ 时,北京属于第一类和第三类的隶属度差值为 0.000 8,太原属于第二类和第三类的隶属度差值为 0.01,合肥属于第一类和第三类的隶属度差值为 0.001 5,均小于等于 $T=0.01$,但是,只有北京到第一类和第三类距离的差值小于 L_j 。因此,在计算 $\text{sep}'(c)$ 过程中,当 $j=1, a=1, b=3$ 时, $W(U; c, d)=0$,其他均取 $1 - |u_{aj} - u_{bj}|$ 的值,代入公式最终得到 $\text{sep}'(3)$ 的值为 2.362。

当 $c=4$ 时,北京的 $u_{11} - u_{14} = 0.0041$,天津的 $u_{21} - u_{24} = 0.0055$,济南的 $u_{15,1} - u_{15,3} = 0.01$,昆明的 $u_{26,2} - u_{26,3} = 0.0006$,均小于等于 $T=0.01$;只有济南满足 L_j 约束,得到 $j=15, a=1, b=3$ 时, $W(U; c, d)=0$,其他均取 $1 - |u_{aj} - u_{bj}|$ 的值,代入公式,最终得到 $\text{sep}'(4)$ 的值为 5.5255。

当 $c=5$ 时,各地区属于第二类和第五类的隶属度差值最小,大多数值都小于 $T=0.01$,同时考虑临界值 T 和 L ,仅有 15

种 $W(U; c, d) = 0$ 的情况,避免了因数据丢失给结果带来的不准确性。

根据计算出的结果,当 $c = 3$ 时 $VS(U, V)$ 的值最大,划分效果最明显。与原 $MPO(U, V)$ 指标进行对比,表明加入距离临界值 L 的 $VS(U, V)$ 有效性函数不仅与原 $MPO(U, V)$ 函数一样,都能确定最优聚类数为 3,划分效果更明显,而且能避免因噪点的过量删除带来的数据信息丢失,并能更加准确地得到类与类之间的模糊点。

3.2 算例 2

对全国 31 个省、市、自治区和直辖市的科研课题能力的模糊聚类结果进行有效性评价。仿照算例 1 步骤,得到结果如表 3、4 所示。

表 3 $VS(U, V)$ 算例结果

c	$com'(c)$	$sep'(c)$	$VS'(U, V)$
3	3.916 4	0.440 6	8.888 8
4	2.610 7	1.485 4	1.757 6
5	2.577 8	2.348	1.097 9

表 4 $MPO(U, V)$ 算例结果

c	$com(c)$	$sep(c)$	$VS(U, V)$
3	15.050 6	1.706 6	8.019
4	13.230 4	3.953 5	3.346 5
5	16.165 7	7.310 6	2.211 3

当 $c = 3$ 时,仅有上海第一类和第二类的隶属度差值为 0.01,小于等于 $T = 0.01$,但是第一类和第二类的距离的差值小于 L_j 。因此,在 $c = 3$ 时没有被删除的噪点,所有值均取 $1 - |u_{aj} - u_{bj}|$,代入公式,最终得到 $sep'(c)$ 的值为 0.4406。

当 $c = 4$ 时,四川的 $u_{23,1} - u_{23,2} = 0.0079$ 同时满足约束, $W(U; c, d)$ 取值为 0,其他值均取 $1 - |u_{aj} - u_{bj}|$,最终得到 $sep'(4)$ 的值为 1.4854。

当 $c = 5$ 时,满足 $|u_{aj} - u_{bj}| \leq T$ 的点有 19 个,加入 $|d_{aj} - d_{bj}| \leq L_j$ 约束后,最终 $W(U; c, d)$ 取值为 0 的点为 4 个。

通过计算得 $VS(U, V)$ 和 $MPO(U, V)$ 的值,确定最优聚类数为 3,再次证明了函数的有效性。

4 结束语

本文在分析了各种聚类有效性评价函数优缺点的基础上对其中的 MPO 评价函数进行改进,得到了新的聚类有效性函数。该函数通过距离和隶属度结合了几何结构域模糊划分两

种理论,构建紧密性度量和分离性度量,在原有隶属度临界值 T 的基础上引入距离临界值 L ,降低了类间噪点对聚类评价结果的影响。通过算例说明了该方法的合理性与有效性。

参考文献:

- [1] 胡雅婷. 可能性聚类方法研究及应用[D]. 吉林: 吉林大学, 2012.
- [2] 彭勇, 吴友情. 一种新的聚类有效性函数[J]. 计算机工程与科学, 2010, 46(6): 124-132.
- [3] DUNN J C. Well-separated clusters and the optimal fuzzy partitions[J]. *Journal of Cybernetics*, 1974, 4(1): 95-104.
- [4] MAULIK U, BANDYOPADHYA S. Performance evaluation of some clustering algorithms and validity indices[J]. *Pattern Analysis and Machine Intelligence*, 2002, 24(12): 1650-1654.
- [5] CALINSIK T, HARABASZ J. A dendrite method for cluster analysis[J]. *Communications in Statistics*, 1974, 3(1): 1-27.
- [6] BEZDEK J C. Numerical taxonomy with fuzzy sets[J]. *Journal of Mathematical Biology*, 1974, 1(1): 57-71.
- [7] BEZDEK J C. Cluster validity with fuzzy sets[J]. *Journal of Cybernetics*, 1973, 3(3): 58-73.
- [8] WINDHAM M P. Cluster validity for fuzzy clustering algorithms[J]. *Fuzzy Sets and Systems*, 1981, 5(2): 177-185.
- [9] ZAHID N, LIMOURI M, ESSAID A. A new cluster-validity for fuzzy clustering[J]. *Pattern Recognition*, 1999, 32(7): 1089-1097.
- [10] XIE Xuan-li, BENI G. A validity measure for fuzzy clustering[J]. *IEEE Trans on Pattern Analysis and Machine Intelligence*, 1991, 13(8): 841-847.
- [11] 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): 112-114.
- [12] 刘万里, 刘三阳, 薛贞霞. 基于距离核函数的除噪和减样方法[J]. 系统工程理论与实践, 2008, 7(7): 160-163.
- [13] HU Ya-ting, ZUO Chun-chang, YANG Yang, et al. A cluster validity index for fuzzy C-means clustering[C]//Proc of the 2nd International Conference on System Science, Engineering Design and Manufacturing Informatization. 2011: 263-266.
- [14] ZALIK K R. Cluster validity index for estimation of fuzzy clusters of different sizes and densities[J]. *Pattern Recognition*, 2010, 43(10): 3374-3390.
- [15] 谢中华. MATLAB 统计分析与应用: 40 个案例分析[M]. 北京: 北京航空航天大学出版社, 2010.

(上接第 36 页)

- [8] HART P E. The condensed nearest neighbor rule[J]. *IEEE Trans on Information Theory*, 1968, 14(3): 515-516.
- [9] WILSON D L. Asymptotic properties of nearest neighbor rules using edited data[J]. *IEEE Trans on Systems, Man and Cybernetics*, 1972, SMC-2(3): 408-421.
- [10] VIOLA P, JONES M. Rapid object detection using a boosted cascade of simple features[C]//Proc of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Society, 2001: 511-518.
- [11] WEISS G M. Learning with rare cases and small disjuncts[C]//Proc of the 12th International Conference on Machine Learning. San Francisco: Morgan Kaufmann, 1995: 558-565.
- [12] LAURIKKALA J. Improving identification of difficult small classes by balancing class distribution[C]//Proc of the 8th Conference on AI in Medicine in Europe. London: Springer-Verlag, 2001: 63-66.
- [13] TOMEK I. Two modifications of CNN [J]. *IEEE Trans on Systems, Man and Cybernetics*, 1976, SMC-6(11): 769-772.
- [14] RICHARDSON M, DOMINOWSKA E, RAGNO R. Predicting clicks: estimating the click-through rate for new ads[C]//Proc of the 16th International Conference on World Wide Web. New York: ACM

Press, 2007: 521-530.

- [15] ANDREW G, GAO Jian-feng. Scalable training of L^1 -regularized log-linear models[C]//Proc of the 24th International Conference on Machine Learning. New York: ACM Press, 2007: 33-40.
- [16] BYRD R H, NOCEDAL J, SCHNABEL R B. Representations of quasi-Newton matrices and their use in limited memory methods[J]. *Mathematical Programming*, 1994, 63(1-3): 129-156.
- [17] LING C X, LI Cheng-hui. Data mining for direct marketing: problems and solutions [C]//Proc of the 4th International Conference on Knowledge Discovery and Data Mining. 1998: 217-225.
- [18] KUBAT M, MATWIN S. Addressing the curse of imbalanced training sets: one-sided selection[C]//Proc of the 14th International Conference on Machine Learning. 1997: 179-186.
- [19] CHEN Chao, LIAW A, BREIMAN L. Using random forest to learn imbalanced data, 666[R]. Berkeley: Statistics Department, University of California, 2004.
- [20] LIU Xu-ying, WU Jian-xin, ZHOU Zhi-hua. Exploratory under-sampling for class-imbalance learning[C]//Proc of the 6th International Conference on Data Mining. 2006: 965-969.
- [21] CHU C, KIM S K, LIN Y A, et al. Map-Reduce for machine learning on multicore [C]//Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2007: 281-288.