一种基于连通性的聚类有效性评价指标

蔡昌许

(曲靖师范学院计算机科学与工程学院 云南 曲靖 655011)

摘 要 针对现有的聚类结果中类内紧致性差异对有效性指标的影响和不能很好地评价任意形状聚类的问题,提出一种基于连通性的聚类有效性指标并进行了仿真研究。首先,将对整个聚类结果的评价建立在对单个类评价的基础上,以便处理类内紧致性差异大的问题。其次,利用连通距离对形状和大小的不敏感性,处理对任意形状聚类的评价问题。仿真实验结果表明,该方法可以对各类的类内紧致性差异较大的任意形状的聚类结果进行评价。该指标是一种有效的聚类评价指标。

关键词 聚类分析 聚类有效性 连通性 仿真

中图分类号 TP391 **文献标识码** A DOI:10.3969/j.issn.1000-386x.2015.11.066

A CLUSTERING VALIDITY EVALUATION INDEX BASED ON CONNECTIVITY

Cai Changxu

(*College of Computer Science and Engineering, Qujing Normal University, Qujing 655011, Yunnan, China*)

Abstract For the problems of existing clustering results that the difference of intraclass compactness impacts on the effectiveness indicators and that they can't well rate the clustering of arbitrary shape, we proposed a connectivity-based clustering validity index and studied it by simulation. First, we set the evaluation of entire clustering results on the basis of evaluating a single class so as to deal with the problem of big difference in intraclass compactness. Secondly, we used the insensitivity of connectivity distance to the shape and the size to deal with the problem of arbitrary shape clustering evaluation. Simulation experimental results showed that the method could evaluate the clustering results of various classes with bigger difference in intraclass compactness and in arbitrary shape. The index is a kind of effective index for clustering evaluation.

Keywords Clustering analysis Clustering validity Connectivity Simulation

0 引 言

对未知类进行划分的聚类分析方法,是指将物理或抽象对象的集合分组成为由类似的对象组成的多个类的分析过程。从计算机人工智能化的角度来观察,具有探索性的聚类分析,是搜索簇的无监督学习过程,它不依赖预先定义的类或带类标记的训练实例,所以与分类不同。这种观察性学习的效果评价是聚类分析中的关键一环。事实证明,具有普遍适用性的聚类算法尚不存在[1]。所以,在使用聚类算法求解相关问题时,首先要对待处理数据本身具有的结构特征进行尝试性假定,假设此项假定有效与否尚不可知。退一步而言,如果假定有效的话,但选定的参数不合适,同样也得不到较好的聚类结果。因此需要对聚类的结果进行评价。这就是所谓的聚类有效性评价问题。目前,聚类处理效果的有效性衡量指标包括如下几类,分别是相对指标、内部指标与外部指标[2]。而硬聚类算法和模糊聚类算法也有不同的聚类有效性指标。这里仅讨论硬聚类算法的相对有效性指标。

聚类有效性是聚类分析中的一个重点和难点问题。研究者们对聚类有效性进行了深入的研究,提出了一些聚类有效性指标。如 Dunn's index[3], DB-index[4], RMSSTD 指标[5], I-index[6]、SD-index[7],以及最近提出的 CDbw 指标[8]和 Connect-index[9]等。当使用该算法进行运算时,均需对待处理的类进行适当的假定,通常需要对连通性或类原型等特征进行假定。其中,连通性假定的运算方法的运算优势是能够发现各种特征的聚类目标,可是此方法解决相切类(数据集中某些属于不同类的数据点间的距离小于同一类内数据点间的距离)的问题不易实现,但针对类原型假定的方法在一定程度上能够很好地解决该问题,却一般处理不好任意形状的相关问题。与此方法的预设假定相对应的算法涉及的有效性指标当中同样含有对应的假设成分,因此,使用上述指标进行结果评估仅对于相同假定的算法问题成立。比如大多数指标项目都是基于类原型假设的指标,所以难以做到对任意形状的运算结果具有正确全面的衡量功能[10]。最近提出的 Connect-index[9]是基于连通性的聚类有效性指标,针对形状多样化的问题运算结果具有较强的评估功能,缺点是对类内紧致差较强的运算效果无法进行有效的衡量。

为了解决该问题,本文给出了一种基于连通性的聚类有效性指标。实验结果表明,该指标可以更好地对任意形状的聚类进行评价。

收稿日期:2014-10-28。曲靖师范学院校级科研项目(2011MS011)。**蔡昌许**,讲师,主研领域:计算机信息管理。

1 基于连通性的聚类有效性指标

1.1 一种聚类有效性指标的框架

已有的评价指标项目均不能实现对形状任意的聚类问题实施有效评价。由学者 Halkidi 研究的 CDbw[8] 衡量指标对于上述聚类问题的解决有了实质性突破，但正如其文章中提到的，该指标适用于对同一聚类数下不同的聚类结果的评价，不适合用于指出正确的聚类数。尽管 connect-index[9] 能够对形状复杂的聚类问题实施妥当的处理，却在解决内紧致性差异较大的情况时效果不甚理想。对于这些问题的解决，本文首先对设定指标实施符合要求的若干假设，在此基础上，结合连通性原理给出 new-index 新的评价指标。经过相关操作和对比可知：本文论述的新的衡量指标可以对各类的类内紧致性差异较大的任意形状的聚类结果进行评价。

如图 1 所示，数据集 Datasetl 内包含以二维形式存在的数据点 80 个，以三个相对分散的类的形式存在。图示显示各个类对应的密度大小不一样，这就说明类内紧致性存在很大差异，此种情况将会造成相关的评价结果不准确。通常，对运算结果的评价拆分成类间分离性和类内紧致性两个度量部分，以提高评价和衡量的准确性和可行性。其中后者是用于评价类内数据点相互之间的近似情况，若结果中各类之间类内相似度差异很大，一般的聚类衡量指标将无法实施评价。究其原因，是在假定时忽略对类之间具有的紧致性不一致的现象，导致假定不合适[11,12]。比如 Dunn's index 用聚类结果中最大的类直径的倒数来定义聚类结果的紧致性，忽略了各类的直径可能差异很大的情况。一般情况下，另外的评价指标也会出现同种情况。

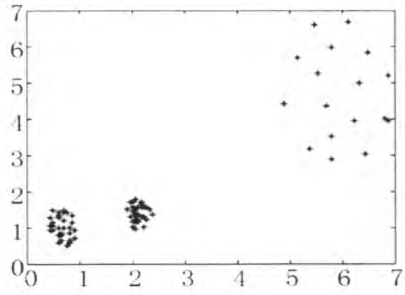

图 1 数据集 Datasetl

本文针对该问题提出了一种新的对聚类结果的评价思路。众所周知，适合数据的聚类算法可以获得紧致性与分离性理想的处理效果，因此可以把上述两类衡量紧密关联起来设计出聚类有效性指标。依据该原则，本文提出以下三条关于聚类有效性指标的假定条件：

（1）待评价的各个类能够各自对紧致性与有效性实施度量和评价。

（2）某个类的评价指标的大小与自身特性、类间距离和别的类的紧致性密切相关。在对某个类的指标进行评价时，还应顾及周围类对该类有效性的影响。这种影响体现在两个方面：第一，该类与其他类的类间距离对该类有效性的影响；第二，别的类的结构特性的干扰。

（3）整个聚类结果的有效性指标大小取自各个类对应的最小值，这也就是所说的“木桶理论”。

在上述假定的基础上，本文提出一种聚类有效性指标新方法。对聚类结果 $C = \{c_1, c_2, \cdots, c_K\}$（其中 K 为聚类数），定义有效性指标如下：

$$index(C) = \min_{i=1}^{K} \min_{j \neq i} \left(\left(\sum_{k=1}^{K} w_k compact(c_k) \right) \times dist(c_i, c_j) \right) \quad (1)$$

其中 $compact(c)$ 表示类 c 的类内紧致性，$dist(c_i, c_j)$ 表示类 c_i 和 c_j 的类间距离，w_k 表示权重因子，显示出 c_k 类对 c_i 类进行有效性衡量时的干扰程度。此时构建的框架能够充分体现并符合上述三个假设的规定。经分析可知，Dunn's index 与 connect-index 的本质就是此框架对应的特殊情况，只是将紧致性最差的类所对应的权重因子设为 1，其余类的权重因子设为 0，从这也可以看出这两个衡量指标的定义忽略了类与类之间紧致性相差悬殊的情形。

1.2 基于连通性的聚类有效性指标

目前，一般采用图连通距离来表示数据类内两点之间的差异程度。使用该物理量来衡量相异度的大小时，能够有效避免欧氏空间相关因素的干扰，提高了评价结果的有效性。Saha 等人[8]利用类似思路提出了适用于评价任意形状聚类的有效性指标 connect-index 。

在无向图 $G(V,E,W)$ 中，顶点集为 $V = \{x_1, x_2, \cdots, x_n\}$，边的集合为 $E = \{e_{ij} \mid$ 顶点 x_i 和 x_j 之间存在边$\}$，E 的权重集合为 $W = \{w_{ij} \mid e_{ij} \in E\}$，设 G 上的两个顶点 x_i 和 x_j 之间路径的集合为 $path(x_i, x_j) = \{path_1, path_2, \cdots, path_k, \cdots, path_p\}$，$p$ 为 x_i 和 x_j 间的路径数，其中一条路径 $path_k$ 上的边记为 $e_1^k, e_2^k, \cdots, e_{nk}^k$，而将对应的权值记为 $w_1^k, w_2^k, \cdots, w_{nk}^k$，则 x_i 和 x_j 间的连通距离定义如下：

$$d_{connect}(x_i, x_j) = \min_{k=1}^{p} \max_{m=1}^{n_k} w_m^k \quad (2)$$

其中，n_k 表示 x_i 和 x_j 之间的路径 $path_k$ 所包含的边数。

将每个类看成一个无向完全图，顶点集定义为数据点的集合，而顶点之间边的权重定义为数据点间的距离（该距离要与待评价聚类算法采用的距离或相似度一致，本文实验选取欧氏距离）。结合连通距离的概念及上一节给出的聚类有效性指标框架，可以定义一个适用于评价任意形状聚类的有效性指标。首先利用连通距离定义单个类的类内紧致性，然后根据第（1）和第（2）两条假设定义单个类的有效性指标，并按照假定式（3）计算出所求指标大小。具体的定义如下：

根据 c 中两点间连通距离的最大值的倒数定义聚类 c 的类内紧致性：

$$compact(c) = \frac{1}{\max_{x,y \in c} \{ d_{connect}(x,y) \}} \quad (3)$$

用两个类之间最近两点间的欧氏距离来定义两个类的类间距离（ $d()$ 表示欧氏距离）：

$$dist(c_i, c_j) = \min_{x \in c_i, y \in c_j} d(x,y) \quad (4)$$

单个类 c_i 的有效性指标 $index(c)$ 定义如下：

$$\min\left(dist(c_i, c_j) \times \left(\frac{|c_i| \times compact(c_i) + |c_j| \times compact(c_j)}{(|c_i| + |c_j|)} \right) \right) \quad (5)$$

其中 $|c|$ 表示类 c 中数据点的个数。该式是用 c_i 和 c_j 类紧致性平均数值与类间距离相乘所得的数值当成 c_i 较 c_j 的有效性指数；取所求的 c_i 较其他类的最小值表示 c_i 类对应的有效性指标数值大小。这样的定义符合上一节提出的第（2）条假设。定义单个类的有效性指标后，就可以依据第（3）条假设对整个聚类结果 $C = \{c_1, c_2, \cdots, c_k\}$ 的有效性指标进行定义：

$$new - index(C) = \min_{i=1}^{K} index(c_i) \quad (6)$$

通过该定义求得的指标数值能够满足三个假设的具体规定，能够准确地解决聚类结果中各类的类内紧致性差异较大的情况。

Ackerman 和 David 提出了聚类有效性指标应满足的基本公理,其中包括同构不变性要求,局部一致性要求和 co-final richness 要求。可以证明,本文提出的聚类有效性指标满足了聚类有效性指标的这三条基本公理。

2　仿真研究

对实验结果判断和衡量,本文使用 single-linkage 算法理论来实施。该算法以连通性理论为基础,是一类理想算法。

首先将 new-index, connect-index 及 Dunn's index 在数据集 Datasetl(如图 1 所示)上进行实验,表明当聚类结果中类内紧致性差异较大时 new-index 具有优势。数据集 Datasetl 分成 3 个类,各类的类内紧致性有较大差异。用 new-index, Dunn's index 和 connect-index 对聚类数取 $K=2,3,4$ 时的聚类结果进行评价(如表 1 中所示,将各个指标取到最优聚类数时所对应的评价结果用红色斜体标出)。new-index 选择了正确的聚类数 $K=3$,其他两个指标则选择了 $K=2$ 作为最优聚类数。出现这种错误,是因为数据中包含的各个类,它们的紧致性特征相差甚远,上述两种算法对此类问题无效。

表 1　在数据集 Datasetl 上各指标对 single-linkage 算法的聚类结果的评价结果

聚类数	Dunn's index	connect-index	new-index
$K=2$	**0.9182**	**3.3660**	3.4734
$K=3$	0.1977	1.1501	**6.1711**
$K=4$	0.4083	1.1163	1.1686

Saha 和 Bandyopadhyay 的文献[8]中展示了在一些人工数据集上 connect-index 和 Dunn's index 的评价结果,并进行了对比,说明了 connect-index 对任意形状的聚类进行评价时具有优势。本文在相同的人工数据集 Dataset2, Dataset3, Dataset4(如图 2 所示)上选取能够满足 single-linkage 算法理论的相关指标与本文论述的 new-index 进行比较,具体结果如表 2 所示,黑体字表示该指标选择了正确的聚类数。从中可以看出 connect-index 和本文提出的 new-index 在对任意形状的聚类进行评价时效果明显好于其他的聚类有效性指标,在三个数据集上都得到了正确的聚类数。然而这些人工数据集几乎都是密度相等的,因此获得理想的评价结论非常方便。

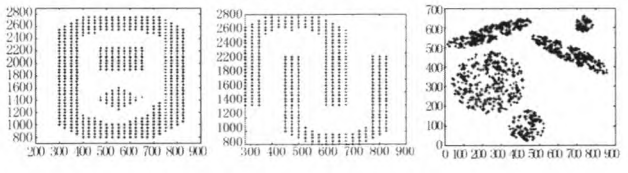

图 2　数据集 Dataset2-4

表 2　在数据集 Dataset2-4 上各指标选择的最优聚类数

数据集	RS	RMSSTD	I-index
Dataset2	2	2	5
Dataset3	2	3	2
Dataset4	3	3	5
数据集	**SD-index**	**connect-index**	**new-index**
Dataset2	4	3	3
Dataset3	2	2	2
Dataset4	2	5	5

为充分证明本文提出的指标参数比 connect-index 更加有效,选择一组聚类分析中常用的数据集将 new-index 和 connect-index 实施性能比较。集合 Dataset5-10 如图 3 所描绘,采取 single-linkage 算法,取 $K=2,3,4,5,6$ 对上述数据集实施聚类操作。由图可知,该算法在对应的五个数据集内能够获得理想的操作效果,但是很难在 Datasetl0 上面获得理想效果。接下来需要对 Datasetl0 正确的聚类结果进行评价,得到 $K=5$ 时的评价结果。表 3 中给出了在这些数据集上两个指标分别得到的最优聚类数,其中黑体字表示该指标选择了正确的聚类数。

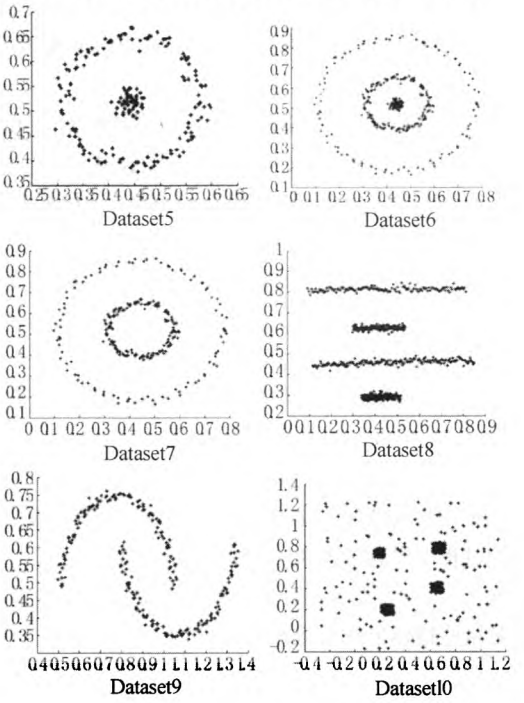

图 3　数据集 Dataset5-10

表 3　在数据集 Dataset5-10 上 connect-index 不 II new-index 选择的最优聚类数

数据集	connect-index	new-index
Dataset5	2	2
Dataset6	2	3
Dataset7	2	2
Dataset8	4	4
Datase9	2	2
Dataset10	3	5

从上述图形中可知,使用 connect-index 算法共有两组选择了错误的最优聚类数,但是 new-index 算法求得的结果十分理想。通过对比可知,connect-index 算法运算错误是因为其没有考虑类内紧致性差距甚大的情形。

而此时论述的 new-index 能够在保证准确评估任意形状的聚类问题的前提下,对上面所述问题的出现进行了实质性的防范,确保了所得结果的优质高效性。

最后给出在 IRIS 数据集上的实验结果。IRIS 数据集可以拆分成 Virginica、Versicolour 和 Setosa 三种。任意一个类中均包括元素 50 个。它们当中,前两个不易拆分。

利用 new-index 理论对经过 single-linkage 方法处理的相关结果实施评估,结果如表 4 中所示。

表 4 在 IRIS 数据集上 connect-index 和 new-index 的评价结果

聚类数	connect-index	new-index
$K=2$	1.592	2.2113
$K=3$	1.1139	1.2175
$K=4$	1.1339	1.1800

与 connect-index 相同，本文定义的 new-index 在 IRIS 范围内更容易选择 $K=2$ 当成最优聚类数来进行相关控制。在对聚类算法进行测试时，最优聚类数取 $K=2$ 是研究者们公认的正确结果。所以本文提出的有效性指标在 IRIS 数据集上是有效的。

经过上述分析可知，本文论述的有效性指标对于求解任意形状聚类评估相关问题，在效果上与传统方法相比，更具有高效准确性。进一步分析可以看出，connect-index 等指标在定义聚类结果的类内紧致性的过程中，选取的是类内紧致性的最小数值，采用这种方式，将致使该指标对于处理类内紧致性相距甚远情形下的问题时，很大程度上会出现符合 RamakriShna 与 Kim 提出的设计聚类有效性指标的要求。

他们设计的有效性指标的前提条件是符合类内紧致性的度量值在 $K_{opt}-1$ 至 K_{opt} 的变化过程中应该迅速上升，在 K_{opt} 至 $K_{opt}+1$ 的范围内迅速跌落的变化走势。指标 connect-index 对聚类结果类内紧致性的定义采用各类紧致性的最小数值，但在取此值的过程中没有考虑类内紧致性的差别情况，这将促使聚类数由 $K_{opt}-1$ 上升至 K_{opt} 时聚类结果的类内紧致性变化不明显。这是这些指标出现失效结果的原因所在。借鉴上述教训，指标 new-index 使用一一针对所有类实施评价的方式，有效避免了上述问题的发生。

New-index 的关键任务是求得全部类对应的类内连通距离最大值，与 Dunn's index 相比，它的计算复杂度相对较高。但是，指标 Dunn'sindex 对于求解任意形状的聚类评价问题无效。经分析还可以知道，所有的类内最大连通距离能够通过 single-linkage 方法求解获得。因此，在对由 single-linkage 产生的相关结果实施评估过程中，new-index 的运算复杂系数是 $O(K^2)$。因此，采用该指标对经由 single-linkage 求得的结果实施评估的情况下，将出现计算复杂度十分微小的理想情况。

本文提出的指标也有不足之处。首先，基于连通性的评价指标在实施对涉及相切类的聚类结果评估时是无效的，指标 new-index 同样会面临这种情况。此外，new-index 沿用了 single-linkage 算法对类间距离的定义，而这样的选择将致使相关指标参数对离群点的反应出奇的迅速。处理此类问题的通常做法是重新选取合适的类间距离来弱化此类现象的发生。此时，本文设定的相关框架继续有效。

3 结 语

Connect-index 作为一种基于连通性的聚类有效性指标，针对形状多样化的问题运算结果具有较强的评估功能，但对类内紧致差较强的运算效果无法进行有效的衡量。本文提出了一中适用于基于连通性聚类算法的聚类有效性指标解决了上述缺陷。首先，对整个聚类结果的评价建立在对单个类评价的基础上，以便处理类内紧致性差异大的问题。其次，利用连通距离对形状和大小的不敏感性，处理对任意形状聚类的评价问题。仿真研究表明，该指标针对于任意形状聚类评估的相关问题都效果非常理想，同时还能够有效弱化因类内紧致性的差异所形成的各种干扰。后续工作将对聚类结果的评价方法做更深入的探讨。从本文的三项假定出发，不断完善相关参量的设定作用，从而形成更具可行性的评价指标。

参 考 文 献

[1] 孙吉贵，刘杰，赵连宇. 聚类算法研究[J]. 软件学报，2008，19(1)：48－61.

[2] Ozden I，Lee H M，Sullivan M R，et al. Identification and clustering of event patterns from in vivo multiphoton optical recordings of neuronal ensembles[J]. Journal of neurophysiology，2012，100(1)：495－503.

[3] 刘明术，方宏彬，张建，等. 属性相似度在聚类算法中的有效性研究[J]. 计算机应用与软件，2012，29(9)：146－147，174.

[4] 应文豪，许敏，王士同，等. 在大规模数据集上进行快速自适应同步聚类[J]. 计算机研究与发展，2014，51(4)：36－40.

[5] 王丽娜，马晓晓. 一种改进的模糊聚类有效性指标[J]. 微电子学与计算机，2014，11(4)：11－15.

[6] Maulik U，Bandyopadhyay S. Performance evaluation of some clustering algorithms and validity indices[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2012，24(12)：1650－1654.

[7] Halkidi M，Vazirgiannis M，Batistakis Y. Quality scheme assessment in the clustering process[J]. Principles of Data Mining and Knowledge Discovery，2013，5(6)：265－276.

[8] Halkidi M，Vazirgiannis M. A density-based cluster validity approach using multi-representatives[J]. Pattern Recognition Letters，2011，29(6)：773－786.

[9] Saha S，Bandyopadhyay S. A validity index based on connectivity[J]. Seventh International Conference on Advances in Pattern Recognition，2012，3(5)：91－94.

[10] 肖满生，汪新凡，朱永平. 非均衡原型结构模式模糊聚类方法研究[J]. 小型微型计算机系统，2013，34(4)：19－23.

[11] Žalik K R，Žalik B. Validity index for clusters of different sizes and densities[J]. Pattern Recognition Letters，2011，32(2)：221－234.

[12] 何云斌，肖宇鹏，万静，等. 基于密度期望和有效性指标的 K-均值算法[J]. 计算机工程与应用，2013，49(24)：118－123.

(上接第 254 页)

[4] 方薇，何留进，宋良图. 因特网舆情传播的协同元胞自动机模型[J]. 计算机应用，2012，32(2)：399－402.

[5] 聂恩伦，陈黎，王亚强，等. 基于 K 近邻的新话题热度预测算法[J]. 计算机科学，2012，39(6A)：257－260.

[6] 杨频，李涛，赵奎. 一种网络舆情的定量分析方法[J]. 计算机应用研究，2009，26(3)：1066－1069.

[7] 方薇，何留进，宋良图. 因特网上舆情传播的预测建模和仿真研究[J]. 计算机科学，2012，39(2)：203－207.

[8] 钱爱玲，瞿彬彬，卢炎生，等. 多时间序列关联规则分析的论坛舆情趋势预测[J]. 南京航空航天大学学报，2012，44(6)：904－910.

[9] 聂恩伦，陈黎，王亚强，等. 基于 K 近邻的新话题热度预测算法[J]. 计算机科学，2012，39(6A)：257－260.

[10] 王巍，杨武，齐海凤. 基于多中心模型的网络热点话题发现算法[J]. 南京理工大学学报：自然科学版，2009，33(4)：422－426.

[11] 史志伟，韩敏. ESN 岭回归学习算法及混沌时间序列预测[J]. 控制与决策，2007，22(3)：258－267.