基于隶属比的聚类有效性指标①

时念云, 侯双双, 马 力

(中国石油大学(华东) 计算机与通信工程学院, 青岛 266580)

摘 要: 针对模糊聚类需要预知最佳聚类个数的问题, 提出了一种新的基于隶属比的聚类有效性指标V_{new}, 首先根据经典有效性指标的设计思路, 充分考虑数据集合的隶属度矩阵特征和几何空间分布, 通过重新定义类内距和类间距的方式, 推导出基本的有效性指标; 其次, 定义隶属比的概念, 放大基本有效性指标的计算值; 最后, 为了避免隶属比对有效性指标造成过分影响而失去意义, 引入分类个数进行抑制. 理论分析和仿真实验表明, 通过对相同数据集进行分析处理, 与经典的 XB 指标相比V_{xb}, 新指标V_{new}具有更高的准确率和可靠性, 在类间有重叠数据的情况下也能够做出正确的划分, 具有一定的推广价值.

关键词: 模糊聚类; 模糊 C 均值; 有效性指标; 类间距; 隶属度矩阵

Validity Index for Fuzzy Clustering Based on Belong Proportion

SHI Nian-Yun, HOU Shuang-Shuang, MA Li

(College of Computer and Communication Engineering, China University of Petroleum(East China), Qingdao 266580, China)

Abstract: Aiming at the problem that fuzzy clustering needs to know the best cluster numbers, a new validity index named new validity index for fuzzy clustering based on belong proportion is proposed in the reference of existing cluster validity indexes. Firstly, it proposes a basic validity index after full consideration of the given dataset's partition matrix and geometrical structure by redefining the separation distance between different clusters. Secondly, it defines the concept of belong proportion to enlarge the calculation value. Finally, it introduces cluster number to restrain the index because belong proportion may cause excessive influence. The new validity index is proved to be more reliable and have a higher accuracy compared to the classical indexes like XB index because it still makes right decision even when the given dataset is overlapping, so the new index has some value to popularize.

Key words: fuzzy clustering; fuzzy c-means (FCM); cluster validity index; separation distance; partition matrix

模糊聚类是一种被广泛研究并且应用于多个关键领域的聚类方法[1], 其通过隶属度函数来确定每个数据隶属于各个类的程度, 美中不足的是, 模糊聚类算法需要事先知道划分的类数[2], 由于大多数情况下, 人们对于数据集的几何结构和空间分布并不知情, 因此在不知道具体划分类数的前提下调用模糊聚类算法的效果并不乐观[3]. 为了能提前预知数据集的最佳划分类数, 人们引入了有效性指标的概念, 即通过一定算法对聚类结果进行评估.

目前, 在模糊聚类的领域已经提出了多种聚类有效性指标, 这些有效性指标大致可以划分为两类: 基于数据集模糊划分的有效性指标和基于数据集几何结构的有效性指标. 基于数据集模糊划分的有效性指标是最早被提出也是最简单的一类有效性指标. 例如, Bezdek 提出的划分系数(Partition Coefficient, V_{pc})和划分熵(Partition Entropy, V_{pe})有效性指标[4]以及随后提出改进的划分系数(Modified Partition Coefficient, V_{mpc})的有效性指标[5], 这三种有效性指标是典型的基于数据集模糊划分的有效性指标, 并未考虑数据集的几何构造[6], Bezdek 将这三种指标应用于国际通用的数据

① 基金项目:中央高校基本科研业务费专项基金(14CX02032A)

收稿时间:2015-11-24;收到修改稿时间:2016-01-11 [doi:10.15888/j.cnki.csa.005274]

© 中国科学院软件研究所 http://www.c-s-a.org.cn

集如 Iris, Butterfly 等, 得到了较为理想的结果. 然而当面临具有复杂几何结构的数据集如Wine时, 这三种指标的效果并不理想. 因此这三种有效性指标仅适用于几何结构并不复杂的低维数据. 基于数据集几何结构的有效性指标是模糊聚类有效性研究的主要分支之一, 研究者相继提出了一系列聚类有效性指标并取得了较好的效果. 其中最著名的是由 Xie 和 Beni 定义的 XB 指标(Xie Beni, V_{xb})指标和 Fukuyama 和 Sugeno 提出的 FS 指标(Fukuyama Sugeno, V_{fs})指标. 上述两种有效性指标同时兼顾了数据成员的隶属度矩阵和几何结构[7], 因此可以处理几何结构相对复杂的高维数据集, 这两个有效性指标在全世界范围内得到了广泛的应用并且取得了理想的效果. 然而大量实验证明, 随着聚类个数的增加, 指标会因为其过分单调而丧失评价的价值, 尤其对于聚类个数较大的情况并不适用[8].

针对上述有效性指标存在的问题可知, 有效性指标应当充分考虑隶属度矩阵和数据成员的几何结构, 在此基础上, 本文提出了一个新的模糊聚类有效性指标, 该指标在继承前人研究的基础上, 创造性地定义了隶属比的概念, 同时引入了惩罚函数的定义来避免指标的过分单调, 因此适用于现实生活中大多数的数据集, 通过大量实验证明, 该有效性指标性能稳定, 可靠性强, 并且可以处理类间有重叠的数据集.

1　模糊C均值算法

模糊 C 均值(Fuzzy C-Means, FCM)算法是一种通过逐步迭代来确定每个数据点属于某个聚类概率的算法. 该聚类算法可以看作是对传统硬聚类算法的一种改进.

1.1 FCM 算法原理

设给定的数据集为 $X=\{x_1, x_2, ..., x_n\}$, 我们需要找到一个数 c 将数据集分为 c 类($X_1, X_2 \ldots, X_c$), 那么数据集的隶属度矩阵为 U(X), U(X) 为 c 行 n 列的矩阵, 其中的元素 u_{ij} 表示第 x_j 个元素隶属于第 X_i 类的概率, 其中 $i\in[1, c]$, $j\in[1, n]$, 我们规定 V_i 为第 X_i 类的聚类中心, 共有 c 个, FCM 算法满足以下关系:

$$0<u_{ij}<1 \tag{1}$$

$$0<\sum_{j=1}^{n}u_{ij}<n \quad i=1,2,3,...c, \tag{2}$$

$$\sum_{i=1}^{c}u_{ij}=1 \qquad j=1,2,3,...n, \tag{3}$$

$$\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}=n \tag{4}$$

$$J_m(\mathbf{U},\mathbf{V})=\sum_{j=1}^{n}\sum_{i=1}^{c}u_{ij}{}^m\left\|x_j-v_i\right\|^2 \quad \text{通常} m=2 \tag{5}$$

FCM 算法原理就是通过不断迭代使得准则函数 $J_m(\mathbf{U},\mathbf{V})$ 取得最小值[9].

1.2 FCM 算法具体步骤

FCM 算法可以通过以下几个步骤来实现:

1) 初始化隶属度矩阵 **U(X)**, 使其满足条件:

$$\sum_{i=1}^{c}u_{ij}=1 \qquad j=1,2,3,...n,$$

2) 计算各个聚类中心 V_i, 使用公式

$$V_i=\frac{\sum_{j=1}^{n}\left(u_{ij}\right)^m x_j}{\sum_{j=1}^{n}\left(u_{ij}\right)^m} \qquad 1\le i\le c$$

即可得到每个类的聚类中心.

计算新的 u_{ij}, 使用公式

$$u_{ij}=\left[\sum_{k=1}^{c}\left(\frac{\left\|x_j-v_i\right\|^2}{\left\|x_j-v_k\right\|^2}\right)^{1/(m-1)}\right]^{-1}, \quad 1\le i\le c,\ 1\le j\le n$$

即可更新隶属度矩阵 **U(X)**.

3) 测试准则函数 $J_m(\mathbf{U},\mathbf{V})$ 是否达到标准要求, 根据已给定的准则函数 $J_m(\mathbf{U},\mathbf{V})=\sum_{j=1}^{n}\sum_{i=1}^{c}u_{ij}{}^m\left\|x_j-v_i\right\|^2$ 来判定, 函数如果小于给定阈值或准则函数前后两次计算值不发生明显变化即可结束, 否则返回执行上述第(2)步, 进行迭代优化.

1.3 利用 FCM 算法确定聚类个数

利用 FCM 确定一个数据集的最佳聚类个数应当包含以下几步[10]:

1) 确定最佳聚类个数的范围[$c_{\min}, c_{\max}$];

2) 令自变量 $i=c_{\min}$, 将 i 带入 FCM 聚类算法得到划分矩阵 **U(*i*)**;

3) 利用 **U(*i*)** 获得有效性指标对应的值 V_i;

4) 如果 $i=c_{\max}$, 获取 V_i 中的最大值(或最小值), 其所对应的 i 为最佳聚类个数, 否则 $i=i+1$, 返回步骤(2).

2　模糊聚类的有效性指标

针对模糊聚类的有效性指标, Bezdek 最早提出了

© 中国科学院软件研究所 http://www.c-s-a.org.cn

划分系数V_{pc}的聚类有效性指标, 该聚类有效性指标仅通过隶属度矩阵就可以生成. 定义为:

$$V_{pc}=\frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^{2}$$

V_{pc}指标表示了数据成员的平均隶属度, V_{pc}的取值范围为[1/c,1], 当V_{pc}等于 1 时表示当前为硬聚类. 当分类越模糊时隶属度矩阵每一列的值就越趋向于平均值 1/c , 即V_{pc}的值越小. 相反分类越分明时隶属度矩阵每一列值的差距越大, 所以通过V_{pc}计算的指标值就越大. 因此我们规定 c 的取值范围为$c_{\min}$到$c_{\max}$, 取V_{pc}最大值所对应的 c 为最佳聚类个数.

随后 Bezdek 提出了划分熵V_{pe}的聚类有效性指标, 该聚类有效性指标通过对u_{ij}取对数的方式来定义:

$$V_{pe}=-\frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}\log_{a}u_{ij}$$

在该公式中a表示对数的底, 取值大于 1, 在本文中测试用值a=10, 根据分析可知V_{pe}的取值范围为$[0,\log_{a}c]$, 取V_{pe}最小值对应的 c 为最佳聚类个数.

V_{pc}和V_{pe}表示了分类信息划分矩阵的累积和, 具有计算简单, 运算速度快的优良特性, 因此得到了广泛的应用. 然而通过分析两个指标的公式可知, 两个指标存在一个明显的缺点, 即在运行过程中会随着 c 的增加而单调性增加或者减小, 在数据集分类个数较大时, 导致指标的失效[11].

针对上述问题Dave改进了V_{pc}提出了V_{mpc}的有效性指标. 该指标抑制了V_{pc}和V_{pe}过分单调的特性, 实验效果良好. V_{mpc}定义如下:

$$V_{mpc}=1-\frac{c}{c-1}\left(1-V_{pc}\right)$$

Dave 的改进降低了V_{pc}和V_{pe}随着类数增加而单调递减的趋势, 可以看出V_{mpc}并没有过多的增加算法的复杂程度, 保留了V_{pc}简单轻便的特点.

上述 3 种算法仅考虑了数据集的隶属度矩阵, 缺少对数据成员的直接引用, 更未考虑数据成员之间的几何关系, 因此适用于几何结构并不复杂的低维数据集合[12]. 为克服上述问题, Fukuyama 和 Sugeno 通过对数据集直接引用, 提出V_{fs}的有效性指标, 定义如下:

$$V_{fs}=\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^{m}\left\|x_{j}-v_{i}\right\|^{2}-\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^{m}\left\|v_{i}-\bar{v}\right\|^{2}$$

在此指标中第一部分为聚类的准则函数, 将隶属度矩阵与数据集 X 的几何紧密度结合了起来, 第二部分定义了平均质心与每个聚类中心点的距离, 然后将其与隶属度矩阵结合起来. 分析可知, 使V_{fs}取得最小值所对应的 c 为最佳聚类个数. 实验证明, V_{fs}的效果并不稳定, 这主要是因为指标的结构是通过减法实现, 相邻的 c 值对应的指标值相差较小, 变化相对平稳导致.

随后, Xie 和 Beni 直接引用数据集合, 考虑数据成员的几何结构和内在联系, 给出了类内聚和类间距的定义, 提出了V_{xb}的有效性指标, 定义为:

$$V_{xb}=\frac{\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^{m}\left\|x_{j}-v_{i}\right\|^{2}}{n*\min_{i,j}\left\|v_{i}-v_{j}\right\|^{2}}$$

该指标中分子表示模糊划分的紧凑程度, 分母表示类间的分离程度. 理论分析可知, 我们应该尽可能地使聚类效果产生一个较小类内聚和相对较大的类间聚, 因此V_{xb}的最小值对应的 c 为最佳的聚类个数. 该指标得到了广泛的应用, 成为以后众多改进算法的基础[13]. 然而由于指标的分子部分随着 c 的增大而逐渐减小, 当 c 逐渐逼近 n 时, 分子趋向于 0, 分母则逐渐增大, 指标同样会因为单调而失去判断能力.

3 新的有效性指标

根据上述分析可知, 一个好的有效性指标应该满足以下几个条件[14]: (1) 聚类的类内紧凑度和类间分离度, 并且保证在最佳聚类数出现时具有最小的类内紧凑度和最大的类间分离度; (2)具有稳定性, 不应过分单调而导致评价结果失效; (3)应当具有处理重叠数据的能力, 在数据集出现重复的时候仍然保证聚类结果的正确性.

基于上述原则, 我们做了以下处理:

1) 为了充分考虑数据的几何结构和空间内在联系, 新指标应当定义类内距和类间距的概念, 同时直接引用数据成员增加指标的说服性. 通过分析现有指标可知, 准则函数在多数有效性指标中被广泛提及, 可以较好的表示类内的紧凑程度, 因此我们直接引用准则函数作为类内的紧凑型指标, 定义如下:

$$J=\sum_{j=1}^{n}\sum_{i=1}^{c}u_{ij}^{m}\left\|x_{j}-v_{i}\right\|^{2}$$

对于类间分离度, 首先使用聚类的中心点代替整

© 中国科学院软件研究所 http://www.c-s-a.org.cn

个聚类, 然后计算各聚类中心点之间的相对距离, 类间分离度可以使用各聚类中心的平均相对距离来实现, 因此, 定义类间分离度如下:

$$sep=\frac{\sum_{i\neq k}^{c}\|v_i-v_k\|}{c(c-1)/2}$$

2) 为了更加充分地利用隶属度矩阵的相关信息, 我们定义隶属比的概念 E. 对于一个隶属度矩阵, 每一列代表的是某个数据成员隶属于 c 个类的概率, 当分类效果恰好的时候, 每一个成员变量必定会出现一个较大的概率隶属于某一类, 因此, 将隶属度矩阵中的每一列中的最大值获取并求和形成分子, 每一列中的最小值获取并求和生成分母, 得到了隶属比 E 如下:

$$E=\frac{\sum_{j=1}^{n}\max_{1}^{c}u_{ij}}{\sum_{j=1}^{n}\min_{1}^{c}u_{ij}}$$

3) 为了保证指标的稳定性, 不会因为 c 的增大而导致指标过分单调从而失去判断价值, 我们应当引入一定的惩罚措施, 减缓指标的单调行为[15]. 因此, 我们在分母中引入 c, 当 c 较小时, 其他变量起主要作用, c 并不能影响指标的生成, 当 c 逐渐增大至 n 时, 由于 c 处于分母上, 一定程度抑制了指标的过分单调[16].

4) 通过上述分析, 得到新的有效性指标

$$V_{new}=\frac{sep}{cJ}*E$$

分析可知, 当分类越分明时, sep 应该具有较大的值, J 随着分类的个数增加而趋向于 0, E 表示隶属度矩阵的分离程度, 当分类越明显的时候, E 的分子应当具有较大的值, 分母具有较小值, 因此, 整个 E 在聚类分明时得到的是较大值. c 的引入是降低指标过分偏大而丧失有效性, 当聚类个数偏小的时候, V_{new} 中除了 c 以外的变量起作用, 当聚类个数偏大时, c 起到了抑制单调的作用. 实验证明这种组合的指标是有效的, 因此 V_{new} 的最大值对应的 c 为最佳的聚类个数.

4　实验结果对比与分析

为了验证新指标对划分结果的有效性, 并且说明优于其他结果, 我们使用 matlab 进行了一系列的仿真实验, 实验用到了一系列的人造数据集和公共数据集, 包括加利福尼亚大学欧文分校(University of California Irvine, UCI)发布的真实数据集. 对比算法是参考文献[6][8][10]中提到的 V_{pc}, V_{pe}, V_{mpc}, V_{xb}, V_{fs} 5 种有效性指标, 实验证明新指标在准确率, 稳定性能方面均超过了上述指标.

数据集

为了加强说服力, 实验数据集采用了人造数据集和通用数据集搭配的策略. 数据集共有 8 个, 其中包括 4 个人造数据集和 4 个通用数据集.

4 个人造数据集全部来源于参考文献 10 中引用的数据集合, 采用 MATLAB 产生高斯分布数据的方式生成, 目的是为了保证数据的权威性, 在文献 10 中, 作者使用这几个基本的数据集进行性能评估, 得到了良好的效果. 数据集共包含 4 个, 分别是 Data_5_2, Data_3_3, Example_1, Example_2. 以 Data_5_2 为例, 第一个数字 5 表示数据集的标准聚类为 5 类, 第二个数字 2 表示这是 2 维数据. 同理可知其它数据集. Example_1 的标准聚类为 3 类, Example_2 的标准聚类为 4 类. 下图以图像的形式展示了数据集的空间分布.

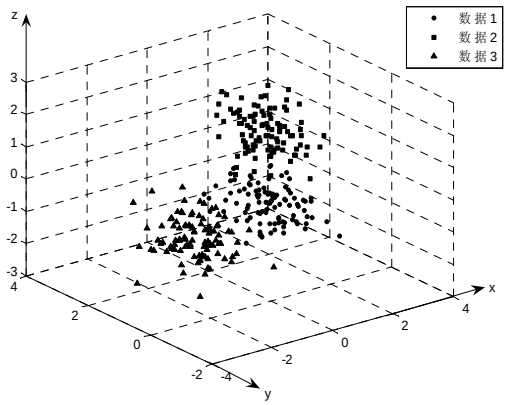

图 1　Data_3_3 数据集

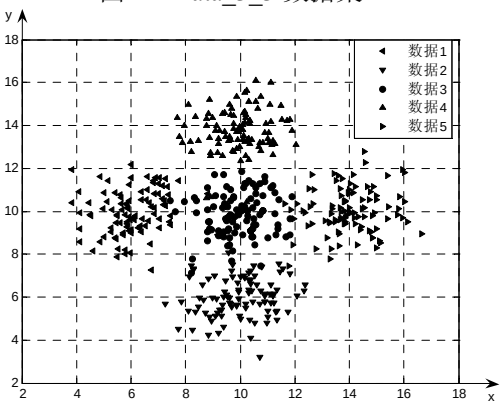

图 2　Data_5_2 数据集

© 中国科学院软件研究所 http://www.c-s-a.org.cn

4 个通用数据集分别是 UCI 公布的通用数据集, Iris, Liver disorder , Wine 和 Butterfly. 其中 Iris 包含 150 个 4 维样本, 用以表示 3 种不同的鸢尾花, 在数据表示上, 这三种品种里面有两种重度重叠, 第三种与前两种分离较好, 因此我们把它的最佳聚类效果分为 2 或 3 类. Liver disorder 是采集于医学领域的真实数据, 包括 345 个 6 维数据, 最佳聚类个数为 2 类. Wine 数据集包含 138 个 13 维样本, 最佳聚类个数为 3 类. Butterfly 数据集相对简单, 包括 15 个 2 维样本, 最佳聚类数为 2 类.

在 matlab 环境下, 使用新指标对指定测试数据集进行测试, 根据第四部分的分析可知, 当新指标 V_{new} 取得最大值时, 其所对应的 c 为最佳的聚类个数. 例如论文中选取了 Data_3_3, Data_5_2 和 Iris 数据集下的测试结果作为代表. 图 3 表示新指标 V_{new} 对数据集 Data_3_3 的测试结果, V_{new} 取得的最大值为 1.0305, 其所对应的分类个数为 3, 即新指标表明此数据集的最佳聚类个数为 3, 而其标准聚类个数也为 3 类, 所以新指标对数据集 Data_3_3 的测试结果是正确的, 如下图 3 所示. 图 4 和图 5 分别是新指标 V_{new} 对数据集 Data_5_2, Iris 的测试结果, 测试结果都与其标准聚类个数相同. 新指标 V_{new} 的测试结果如图 3 所示:

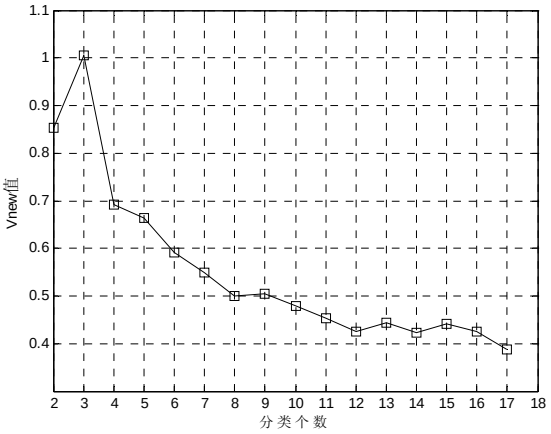

图 3 新指标对 Data_3_3 测试结果

实验的整体对比结果如表 1 所示:

由表 1 的数据结果可以看出, V_{pc} 和 V_{pe} 的聚类效果相差不大, 在测试中出现了 3 个错误的结果, 这与它们仅考虑隶属度矩阵而忽略数据集的空间结构有关系, V_{xb} 和 V_{fs} 的聚类效果类似, 均出现了 2 个判断失误的情况. 新指标与 V_{mpc} 指标的聚类效果相当, 在充分考虑数据集的复杂程度与多维情况下, 只有 wine 数据集出现了错误, 但是值得注意的是, 在 Iris 数据集中, 有两种数据存在重度重叠, 而第三种与前两种分离较好, 因此, 将数据集划分为 2 种或者 3 种均是正确结果. 在对比的指标中, 包括错误率相等的 V_{mpc} 指标, 聚类个数都是 2, 只有新指标聚类个数为 3, 这说明新指标具有一定处理重叠数据的能力. 之所以可以对重叠数据分类正确是因为在新指标在整体上添加了隶属比 E 的缘故, 因为 E 是一个放大系数, 在数据集中存在重叠数据时, 非重叠部分的效果将会被放大, 而指标又是取最大值对应的分类个数, 所以其分类结果倾向于偏大. 这也就是为什么 wine 数据集最佳聚类为 3 类, 而本指标却得出 10 的原因. 正如 Bezdek 所言, 任何一个有效性指标都不可能对所有的数据集都有效, 但是从整体的判断效果来看, 新指标与其他指标相比, 具有更高的准确性, 性能也更加稳定, 总体效果要超过了 V_{pc}, V_{xb} 等经典指标.

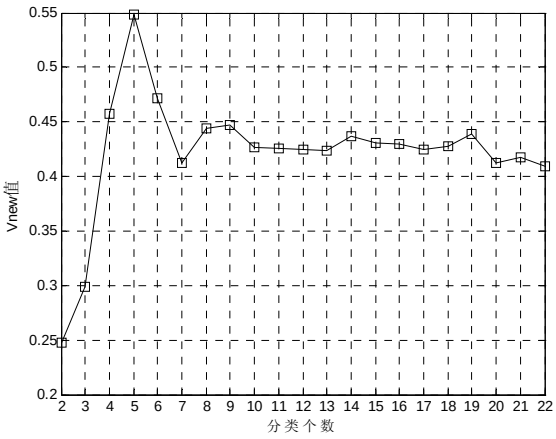

图 4 新指标对 Data_5_2 测试结果

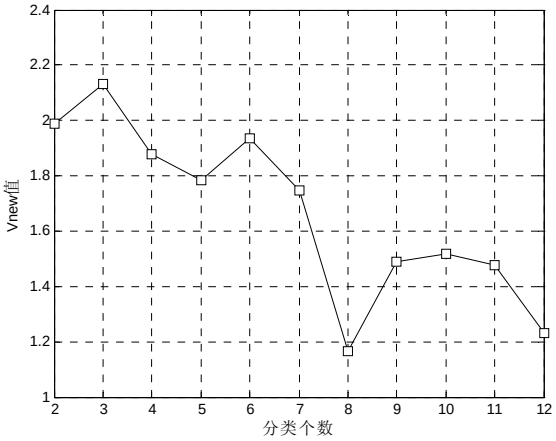

图 5 新指标对 Iris 测试结果

© 中国科学院软件研究所 http://www.c-s-a.org.cn

表 1 实验对比结果

数据集	C^*	V_{pc}	V_{pe}	V_{mpc}	V_{fs}	V_{xb}	V_{new}
Data_3_3	3	2	2	3	3	3	3
Data_5_2	5	4	2	5	5	4	5
Example_1	3	3	3	3	3	3	3
Example_2	4	4	4	4	4	4	4
Iris	2/3	2	2	2	2	2	3
Liver	2	2	2	2	19	2	2
Butterfly	2	2	2	2	2	2	2
Wine	3	2	2	2	4	2	10
错误数	0	3	3	1	2	2	1

5 结语

本文在总结对比前人提出的有效性指标的基础上, 充分考虑了数据集的分布和内在联系, 提出了一种新的有效性指标, 该指标定义了隶属度比 E 的概念, 能够对类内有重叠的数据进行正确判断, 但是对 wine 等数据集处理效果不佳, 这是因为指标内部隶属比 E 放大导致的, 为了能在得到准确分类结果的同时保持对数据集重叠情况的判断能力, 尝试引入分类个数 c 来部分抑制其效果, 但是这两点是矛盾的, 怎么能在 c 和 E 之间取得一个平衡点, 是作者正在研究的内容.

参考文献

1 Bezdek JC. Fuzzy mathematics in pattern classification[Ph.D. Dissertation]. Ithaca, NY: Cornell University, 1973.

2 白素琴,吴小俊.基于模糊聚类算法的有效性指标.江南大学学报(自然科学版),2007,6(6).

3 Bouguessa M, Wang SR. A new efficient validity index for fuzzy clustering. Proc. Third Internat. Conf. on Machine Learning and Cybernetics, Shanghai, 26-29 August 2004.

4 Trauwaert E. On the meaning of Dunn's partition coefficient for fuzzy clusters. Fuzzy Sets and Systems, 1988, 25: 217–242.

5 Dave RN. Validating fuzzy partition obtained through c-shells clustering, Pattern Recognition Lett. , 1996, 17: 613–623.

6 朱文捷,吴楠,胡学钢.一个改进的模糊聚类有效性指标.计算机工程与应用,2011,5:206–208

7 邱学芹.模糊聚类算法及其聚类有效性的研究[硕士学位论文].青岛:青岛理工大学,2010.

8 朱文婕.模糊聚类有效性指标研究[硕士学位论文].合肥:合肥工业大学, 2009.

9 刘梦娇,吴成茂.一种改进的局部模糊 C-均值聚类分割算法研究.计算机科学,2015,42(6):190–191.

10 Wang W, Zhang Y. On fuzzy cluster validity indices. Fuzzy sets and systems, 2007, 158(19): 2095–2117.

11 贺玲,吴玲达,蔡益朝.数据挖掘中的聚类算法综述.计算机应用研究,2007,24(1):10–13.

12 鲍正益.模糊聚类算法及其有效性研究[硕士学位论文].厦门:厦门大学,2006.

13 普运伟,金炜东,朱明,等.核模糊 C 均值算法的聚类有效性研究.计算机科学,2007,34(2):207–210.

14 Shim Y, Chung J, Choi IC. A comparision study of cluster validity indices using a nonhierarchical clustering algorithm. CIMCA-IAWTIC. Washington: IEEE Computer Society, 2005: 538–543.

15 王园园,倪志伟.基于决策树的模糊聚类评价算法及其应用.计算机技术与发展,2009,(9):32–37.

16 周涛,陆惠玲.数据挖掘中聚类算法研究进展.计算机工程与应用,2012,48(12).

© 中国科学院软件研究所 http://www.c-s-a.org.cn