doi:10.11959/j.issn.1000-436x.2015146

基于证据空间有效性指标的聚类选择性集成

毕凯，王晓丹，邢雅琼
(空军工程大学 防空反导学院，陕西 西安 710051)

摘 要：首先针对距离空间在描述数据复杂结构信息方面的不足给出证据空间的概念。然后基于证据空间扩展有效性指标 Davies-Bouldin，同时利用聚类成员的类别相关矩阵度量差异性。最后以较高有效性和较大差异性为目标选择聚类成员并用于集成。实验结果显示所提方法能够有效提高聚类集成算法的有效性。
关键词：Davies-Bouldin 指标；证据空间；聚类选择性集成；互相关矩阵
中图分类号：TP391 **文献标识码**：A

Cluster ensemble selection based on validity index in evidence space

BI Kai, WANG Xiao-dan, XING Ya-qiong
(School of Air and Missile Defense, Air Force Engineering University, Xi'an 710051, China)

Abstract: At first, the concept of evidence space was proposed to overcome the weakness of distance space for describing the complex structure of data sets. And then, the Davies-Bouldin index was extended based on the evidence space proposed. Meanwhile the label-correlation matrix was used to measure the difference of clusters members. At last, the cluster members with better effectiveness and bigger differences were selected for cluster ensemble. The experimental results show that the proposed method is able to improve the effectiveness of cluster ensemble.
Key words: Davies-Bouldin index; evidence space; cluster ensemble selection; co-association matrix

1 引言

聚类过程可描述为基于一定准则将数据集合划分为不同的类别，使同类数据间相似度较高而异类数据间相似度较低。由于聚类算法不需要先验知识，更为客观灵活，符合人们对客观事物认识的一般规律，因而数十年来聚类一直作为模式识别的重要研究领域且应用广泛[1~4]。2002 年，Alexander[5]等首次提出了聚类集成的概念。与集成学习不同，由于不受先验知识的束缚，聚类集成通常具有更好的泛化性能，但也正是由于缺乏必要的先验知识，大多集成学习方法难以直接应用于聚类集成[5~7]。

近年来选择性集成已成为聚类集成研究的重要方向，相关学者通过构造不同的有效性和差异性指标，对聚类的选择性集成进行了较为全面的探讨。如文献[8]分析了多种差异性度量标准与集成结果间的关系，指出差异性与集成结果间的关系是复杂的，不同的成员有效性、不同的数据分布、不同的集成方法都显著影响二者间的关系。文献[9]基于规范化互信息给出了 5 种成对差异性度量方法，实验结果表明中等差异性的聚类成员更有利于形成良好的集成结果。文献[10]综合考虑聚类成员的有效性和聚类成员间的差异性，将选择性集成问题转化为有效性和差异性的组合优化问题，聚类成员与随机子集的平均一致性可以反映聚类算法在数据集上的聚类有效性，但是当数据集空间分布严重不均匀时，失衡的随机子集无法有效反映数据的空间结构。文献[11]提出了一种基于 bagging 的聚类选择性集成方法，选择与全集成结果具有较高相似性的聚类成员进行集成，全集成结果虽然能够反映集成

收稿日期：2014-07-07；**修回日期**：2014-10-09
基金项目：国家自然科学基金资助项目（60975026, 61273275）
Foundation Item: The National Natural Science Foundation of China (60975026, 61273275)

的总体趋势，但是由于聚类成员分类性能差异会导致全集成结果具有一定的波动性，若以较差的全集成结果作为标准则难以保证选择出聚类成员的有效性。文献[12]则通过大量的实验分析了多种聚类有效性指标与选择性集成结果间的关系，已有有效性标准通过分析聚类结果的类内相似性和类间差异性来作为有效性评价标准，此类指标多基于一定数据结构假设，当假设与实际不符时，有效性评价可能失效。

基于上述分析可见，合理的有效性指标是聚类选择性集成的重点和难点。本文提出一种基于证据空间有效性的聚类选择性集成方法。首先给出证据空间的概念，而后基于证据空间扩展 Davies-Bouldin 指标，以其作为聚类成员的有效性描述，并利用聚类成员间类别相关矩阵的不同来度量差异性，最后利用前述有效性和差异性进行聚类的选择性集成。

2 聚类的选择性集成

设数据集合为 $X=\{x_1,x_2,\cdots,x_N\}$，N 为数据规模，$\Pi=\{\pi_1,\pi_2,\cdots,\pi_M\}$ 为数据集 X 上 M 个独立差异性聚类结果，称之为聚类全体，其中 $\pi_i=\{c_{i1},c_{i2},\cdots,c_{ik^i}\}$ 为聚类成员，k^i 为类别数。对于任意数据 x，若 $x\in c_{ij}$ 则表示该数据点在第 i 次聚类中类别为 j。聚类集成本质上是利用合理的方法对 M 个独立聚类结果的一致性分析和优化组合。常用的组合方法包括：CSPA、HGPA 和 MCLA[5]、Co-Association Matrix 方法[6]，以及类别特征空间方法[7]等。

与选择性集成学习类似，有效性和差异性是评价聚类成员是否有利于集成的主要标准。有效性指标描述聚类成员自身的准确性，依据是否需要参考对象，可分为外部方法和内部方法。外部方法将每个聚类结果与标准划分进行对比，一致性越高表示该聚类结果有效性越高。全集成结果(LCE, labels of cluster ensemble)或聚类成员的全部结果(LEC, labels of every cluster)通常被作为标准划分，表示集成的总趋势。归一化互信息(NMI, normalized mutual information)、Rand 统计、Jaccard 系数、Fowlkes-Mallows 指标等均被用于描述聚类成员与总趋势间的一致性[8]。内部方法则通过分析聚类结果与数据集合的空间结构相似性确定其有效性。常用的评价标准有轮廓指标(silhouette index)、Dunn 指标、Davies-Bouldin 指标、Gap 统计等[12]。

差异性指标当前研究较为深入，文献[9]基于 AR(adjusted rand index)提出了 5 种差异性度量方法，并通过大量实验分析发现中等差异程度的聚类成员更有利于获取较好的集成结果，文献[8]对聚类成员差异性与集成结果间关系进行了更为广泛的研究，将文献[9]中 5 种差异性指标以及基于 NMI、信息熵(entropy)、条件熵(conditional entropy)、double fault measure、coincident failure diversity、measurement of inter-rater agreement 构造差异性指标，大量的实验研究表明各种差异性度量与集成准确度之间并没有严格的单调关系，聚类成员有效性、不同的集成规模、不同的数据结构、不同的集成方法等都可能影响选择有效性。

需要指出的是，单独考虑有效性或差异性都很难得到较为理想的实验结果，大量研究表明[10~12]，二者的组合考虑更能获得稳定、良好的选择结果。

3 证据空间有效性

3.1 距离空间

距离空间作为一种重要的拓扑空间，其一般化定义可描述如下。

定义 1 （距离空间）对于非空集合 $X=\{x_i\}$，定义 X 上的二元实值函数 D，若满足下列 3 个条件：

1) 非负性，$D(x_i,x_j)\geqslant 0$，当且仅当 $x_i=x_j$ 时取 0；

2) 对称性，$D(x_i,x_j)=D(x_j,x_i)$；

3) 三角不等式，$D(x_i,x_j)\leqslant D(x_i,x_k)+D(x_j,x_k)$。

则称 $D(x_i,x_j)$ 为 x_i 和 x_j 的距离，X 为以 D 为距离的距离空间。常用的距离描述包括欧氏距离、马氏距离、巴氏距离等，其中欧氏距离利用多维空间内任意两点间的直线测距表示，理解直观、计算简单且具有较好的普遍适用性，当前多数聚类方法和有效性评价准则多是基于欧氏距离空间提出，本文后续的相关讨论与实验均以欧氏距离空间作为对比。

3.2 证据空间

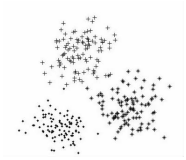
互相关矩阵[6](CAM, co-association matrix)是聚类集成中应用较为广泛的一致性描述方法，利用证据积累的方法组合差异性聚类成员。其一般描述如式(1)所示。

$$\boldsymbol{S}=\begin{bmatrix} s_{1,1} & \cdots & s_{1,N} \\ \cdots & \ddots & \cdots \\ s_{N,1} & \cdots & s_{N,N} \end{bmatrix}_{N\times N} \tag{1}$$

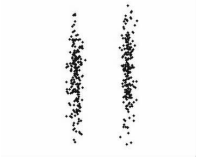
其中，$s_{i,j}=\dfrac{n_{i,j}}{M}$，$n_{i,j}$ 为数据点 x_i 与数据点 x_j 在 M 个聚类成员中归为同类的次数，$s_{i,j}\in[0,1]$，其本质上是各聚类成员对数据点 x_i 与数据点 x_j 同类的证据积累，$s_{i,j}$ 越接近于 1 表明两数据点同类的支持度越高，$s_{i,j}$ 越接近于 0 则两数据点同类的支持度越低。可以看到互相关矩阵本质上是各聚类成员类别相关矩阵的累积。文献[6]在较宽的类别变化范围内利用 *k*-means 聚类算法获取聚类成员，通过样本点间类别关系的证据积累构造互相关矩阵。

大量的研究表明[6,13,14]，利用合理差异性聚类成员构造出的互相关矩阵能够提供较欧氏距离空间更为合理的数据结构描述。

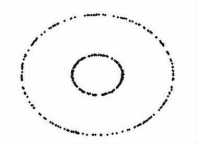
例 1　如图 1 所示，3 个二维人工数据集，图 1(a)为 3 个高斯分布数据集，各类规模均为 100；图 1(b)为 2 个链式分布数据集，各类规模均为 200；图 1(c)为 2 个环形分布数据集，各类规模均为 200。分别利用欧氏距离和互相关矩阵描述图中各数据集中的点的相关性，如图 2 所示，其中，图 2(b)、图 2(d)、图 2(f)利用 100 次 *k*-means 聚类算法生成聚类成员，通过类别参数扰动(在区间[10, 60]内随机取值)和初始中心点扰动保证差异性，各图为式(1)所示互相关矩阵的图形化表示，亮度越高表示横、纵坐标所对应的两样本点相似性越高，亮度越低则表示相似度越低。由于聚类成员的类别取值区间([10, 60])远高于实际数据集的类别数，能够提供数据集更细致的结构描述，通过一定规模的证据积累，互相关矩阵往往具有更好的空间结构描述性。



(a) 3 个高斯数据集

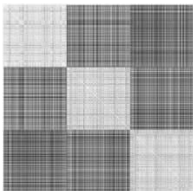


(b) 双链式数据集

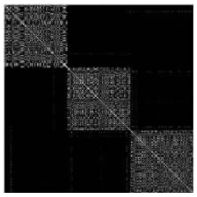


(c) 双环形数据集

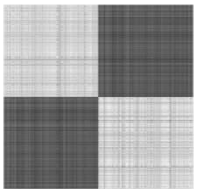
图 1　人工数据集



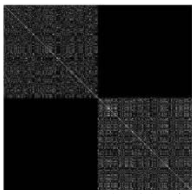
(a)图 1(a)欧氏距离相似性



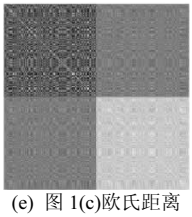
(b) 图 1(a)互相关矩阵相似性



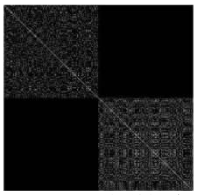
(c) 图 1(b)欧氏距离相似性



(d)图 1(b)互相关矩阵相似性



(e) 图 1(c)欧氏距离相似性



(f)图 1(c)互相关矩阵相似性

图 2　人工数据集欧氏距离空间与互相关矩阵相似性描述对比

合理的空间结构描述对于数据集的分类和分类结果的有效性评价具有重要意义，考虑互相关矩阵在数据空间结构描述的合理性，这里基于互相关矩阵给出证据空间的定义。

定义 2　(证据) 聚类成员划分结果的类别相关矩阵称之为证据。

定义 3　(证据空间) 对于非空集合 $X=\{x_i\}$，可利用多个证据的积累来描述其空间结构，其中任意两样本点间的相似性描述表示为

$$S(x_i,x_j)=\frac{n_{i,j}}{M} \tag{2}$$

其中，M 为证据总个数，$n_{i,j}$ 为 M 个证据中将 x_i 和 x_j 划为同类的证据个数。

3.3　基于证据空间的 Davies-Bouldin 指标

DB(Davies-Bouldin)指标由 Davies 和 Bouldin 提出[15]，是一种基于类内相似性和类间差异性的有效性评价。

对于聚类划分 π，C_i 为其第 i 个类别划分，$i=1,\cdots,K$，K 为总类别数。设 s_i 为 C_i 的分散程度，$d_{i,j}$ 为类别 C_i 和 C_j 的不相似程度，定义 C_i 和 C_j 间的相似性指标 $R_{i,j}$ 满足下述 5 个条件[15]。

1) $R_{i,j}\geqslant 0$;

2) $R_{i,j}=R_{j,i}$;

3) 若 $s_i=0$ 且 $s_j=0$，$R_{i,j}=0$;

4) 若 $s_j>s_k$ 且 $d_{i,j}=d_{i,k}$，$R_{i,j}>R_{i,k}$;

5) 若 $s_j=s_k$ 且 $d_{i,j}<d_{i,k}$，$R_{i,j}>R_{i,k}$;

条件 1)、2)表明，$R_{i,j}$ 非负且对称，条件 3)表明，当 2 类别均重叠于一点时，$R_{i,j}$=0，条件 4)表明，与

其他 2 类别 C_j 和 C_k 差异性相同的类别 C_i，与具有较大分散度的类别更相似，条件 5)表明，两分散度相同的类别 C_j 和 C_k，类别 C_i 更相似于差异性较小的类别。令

$$d_{i,j}=\left\|v_i-v_j\right\|_q \tag{3}$$

$$s_i=\left(\frac{1}{|C_i|}\sum_{u\in C_i}\left\|u-v_i\right\|_q\right)^{1/q} \tag{4}$$

则一种简单的 $R_{i,j}$ 可表示为

$$R_{i,j}=\frac{s_i+s_j}{d_{i,j}} \tag{5}$$

$$R_i=\max_{j=1,\cdots,K,j\neq i}R_{i,j} \tag{6}$$

则 *DB* 指标表示为

$$DB=\frac{1}{K}\sum_{l=1}^{K}R_l \tag{7}$$

上述各式中，v_i 为类别 C_i 的类别中心点，$|C_i|$ 为类别 C_i 的数据个数。较小的 *DB* 指标意味着聚类结果中各类别具有较好的类内相似性和类间差异性。

通常情况下，*DB* 指标利用距离空间给出相似性和差异性描述，当特征合理且数据结构致密时可以取得较好描述(如图 2(a)和图 2(b)所示)，但是当所提取特征难以描述数据空间结构(如图 2(c)所示)时，*DB* 指标失效。考虑证据空间对于数据结构良好的描述性，这里给出基于证据空间的 *DB* 指标(DB^{ES}, DB based on evidence space)。

对于定义 3 中所述证据空间，构造出形式如式(1)所述关系矩阵，矩阵元素为两数据点同类的归一化支持程度。基于该关系矩阵分散度和差异性可表示为

$$s_i^{ES}=\begin{cases}\dfrac{1}{|C_i|(|C_i|-1)}\displaystyle\sum_{x_1\in C_i}\sum_{x_2\in C_i,x_2\neq x_1}\left(1-S(x_1,x_2)\right), & |C_i|\neq 1\\ 0, & |C_i|=1\end{cases} \tag{8}$$

$$d_{i,j}^{ES}=\frac{1}{|C_i||C_j|}\sum_{x_1\in C_i}\sum_{x_2\in C_j}\left(1-S(x_1,x_2)\right) \tag{9}$$

其中，s_i^{ES} 表示类别 C_i 内数据点间不同类的平均支持度，支持度越高表明类别 C_i 的分散程度越高。$d_{i,j}^{ES}$ 表示类别 C_i 与类别 C_j 中数据点间的平均差异程度描述，其值越大表示 2 类别在证据空间内差异性越大，其值越小则表明差异性越小。

借鉴式(5)~式(7)的描述形式，则 DB^{ES} 可表示为

$$DB^{ES}=\frac{1}{K}\sum_{l=1}^{K}R_l^{ES} \tag{10}$$

DB^{ES} 取值越小表明聚类结果有效性越高，其中 $R_i^{ES}=\max_{j=1,\cdots,K,j\neq i}R_{i,j}^{ES}$，$R_{i,j}^{ES}=\dfrac{s_i^{ES}+s_j^{ES}}{d_{i,j}^{ES}}$。易于证明基于证据空间定义的 $R_{i,j}^{ES}$ 满足前边所述的 5 个条件。

4 聚类的选择性集成

本节首先给出一种基于类别相关矩阵(LAM, label association matrix)的差异性描述方法，而后进一步讨论利用有效性和差异性的聚类选择性集成。

4.1 基于类别相关矩阵的差异性描述

对于任意聚类成员 π，利用式(11)构造其类别相关性矩阵

$$LAM_\pi(i,j)=\begin{cases}1, & label(i)=label(j)\\ 0, & label(i)\neq label(j)\end{cases} \tag{11}$$

其中，$label(i)$ 为该聚类成员结果中第 i 个数据点的类标签。观察两聚类成员 π_1 和 π_2 所构造的类别相关性矩阵 $\boldsymbol{LAM}_{\pi_1}$ 和 $\boldsymbol{LAM}_{\pi_2}$，二者的差异性表示为

$$DLAM_{\pi_1,\pi_2}(i,j)=LAM_{\pi_1}(i,j)-LAM_{\pi_2}(i,j) \tag{12}$$

$\boldsymbol{DLAM}_{\pi_1,\pi_2}$ 为 $N\times N$ 矩阵，矩阵元素在 $\{-1,0,1\}$ 集合内取值，0 表示两聚类成员均认为数据点 i 与 j 属于同类(或异类)，−1 和 1 表示两聚类成员中一个认为两数据点属于同类，而另一个则认为两数据点不属于同类。可见 $\boldsymbol{DLAM}_{\pi_1,\pi_2}$ 中非 0 矩阵元素即为两聚类成员的差异性位置，对差异性的量化描述如式(13)所示。

$$SDLAM(\pi_1,\pi_2)=\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left|DLAM_{\pi_1,\pi_2}(i,j)\right| \tag{13}$$

其中，$\dfrac{1}{N^2}$ 为归一化因子，$SDLAM(\pi_1,\pi_2)$ 取值范围在 [0,1]，其值越大表示差异性越大，反之则差异性越小。

单个聚类成员的差异性可以通过与其余聚类成员间的差异性求和取得，因而每个聚类成员的差异性度量(DDM, difference degree of cluster member)可表示为

$$DDM(\pi_i)=\sum_{j=1,\cdots,M,j\neq i}SDLAM(\pi_i,\pi_j) \tag{14}$$

4.2　聚类的选择集成

好的聚类成员应该同时兼顾有效性和差异性，因而聚类成员的选择过程可以看作是式(15)所示有效性和差异性的两目标的最优化求解[10]

$$\begin{cases} \min\limits_{i=1,\cdots,M}(DB^{ES}(\pi_i)) \\ \max\limits_{i=1,\cdots,M}(DDM(\pi_i)) \end{cases} \tag{15}$$

文献[10]给出了一种平衡有效性和差异性的选择方法，聚类成员的集成适应度(FE, fitness for ensemble)通过对其有效性和差异性的加权组合来获取

$$FE(\pi_i)=(1-\lambda)\frac{DDM(\pi_i)}{\max(DDM)}+\lambda\left(1-\frac{DB^{ES}(\pi_i)}{\max(DB^{ES})}\right) \tag{16}$$

其中，λ 取值在[0,1]之间，为有效性和差异性的平衡因子，$\frac{DDM(\pi_i)}{\max(DDM)}$ 为归一化后的差异性指标，$1-\frac{DB^{ES}(\pi_i)}{\max(DB^{ES})}$ 为归一化后的有效性指标，根据 DB^{ES} 定义，其值越小表示聚类结果有效性越高，略作调整，此时 FE 取值越大表明聚类成员越适合于参与聚类集成。下面给出一种利用 FE 的聚类选择性集成算法。

算法 1　(聚类的选择性集成)

输入：数据集合 X，集成规模 M，选择规模 m，平衡因子 λ，聚类算法 F，基于 CAM 的二次聚类算法 G

输出：数据集合 X 的选择性聚类集成结果

Step1　对聚类算法 F 通过参数扰动、状态扰动等方法[6]生成 M 个差异性聚类成员 $\Pi=\{\pi_1,\pi_2,\cdots,\pi_M\}$；

Step2　利用式(2)和聚类成员集合 Π 生成证据空间 S；

Step3　基于证据空间 S，利用式(8)~式(10)计算每一聚类成员的有效性 $DB^{ES}(\pi_i)$；

Step4　利用式(12)~式(14)计算每一聚类成员的差异性度量 $DDM(\pi_i)$；

Step5　利用式(16)计算每一聚类成员的集成适应度 $FE(\pi_i)$ 并排序，选择出前 m 个具有最大适应度的聚类成员构造 CAM；

Step6　利用聚类算法 G 对 CAM 进行分类，获取数据集合 X 的集成分类结果。

4.3　复杂度分析

由于聚类成员的生成过程(算法 1 中的 Step1)和 CAM 的聚类过程(算法 1 中的 Step6)可以选择多种不同的方法，时间复杂性可根据实际情况进行调整，这里仅分析聚类成员的有效性评价、差异性评价以及选择过程的时间复杂度。

假设证据空间已经求得，聚类成员的有效性指标 DB^{ES} 的计算时间主要包括类别分散度(式(8))、类间相似度(式(9))以及式(10)3 个计算过程，对于类别数为 K，各类别数据个数为 $|C_i|$ 的聚类划分，取单位运算时间为 1，则 DB^{ES} 的计算时间为

$$T(DB^{ES})=\sum_{i=1}^{K}|C_i|(|C_i|-1)+\sum_{i=1}^{K-1}\sum_{j=i+1}^{K}|C_i||C_j|+\frac{K(K-1)}{2}+K \tag{17}$$

其中，$\sum\limits_{i=1}^{K}|C_i|(|C_i|-1)$ 为分散度的计算时间，$\sum\limits_{i=1}^{K-1}\sum\limits_{j=i+1}^{K}|C_i||C_j|$ 为类别相似度的计算时间，$\frac{K(K-1)}{2}$ 为 $R_{i,j}^{ES}=\frac{s_i^{ES}+s_j^{ES}}{d_{i,j}^{ES}}$ 的计算时间，K 为 $R_i^{ES}=\max\limits_{j=1,\cdots,K,j\neq i}R_{i,j}$ 的计算时间，综上 DB^{ES} 的时间复杂度不高于 $O(KN^2)$。

分析式(12)~式(14)可知任意聚类成员间差异性的计算时间可表示为

$$T(DDM(\pi_i))=\sum_{j=1,j\neq i}^{K}N^2 \tag{18}$$

其中，N^2 为任意两 LAM 间差异性的计算时间，可见聚类成员差异性的时间复杂度为 $O(KN^2)$。

聚类成员的选择过程实际上是式(16)计算出的集成适应度的排序和选择，因而选择过程时间复杂度与排序过程有关，时间复杂度不高于 $O(M^2)$，M 为集成规模。

通常 $M<<N$，因而选择过程的时间复杂度为 $O(KN^2)$，与聚类成员的类别数成正比，与数据规模的平方成正比。

此外，由于 DB^{ES} 指标基于互相关矩阵给出，且在计算过程中未明显增加存储需求，因而空间复杂度为 $O(N^2)$。差异性描述 DDM 度量任意两类别相关矩阵间差异性，因而需要存储各聚类成员的类别相关矩阵，其空间复杂度为 $O(MN^2)$。

5　实验与验证

5.1　实验数据与设计

实验数据包括 2 部分：8 组 UCI 标准数据集以

及 4 组人工数据集：2-ring、2-line、3-gauss(如图 1(c)、图 1(b)、图 1(a)所示)、half-ring[6]，上述数据集的详细描述如表 1 所示。

表 1　　实验数据描述

数据集	样本数	维数	类别
iris	150	4	3
wine	178	13	3
soybean	306	35	4
sonar	208	60	2
ionosphere	351	33	2
vehicle	846	18	4
segment	2 310	18	7
yeast1	1 484	8	10
2-ring	400	2	2
2-line	400	2	2
3-gauss	300	2	3
half-ring	400	2	2

以集成结果与标准类标签的 FM(fowlkes mallows)指标作为评价标准

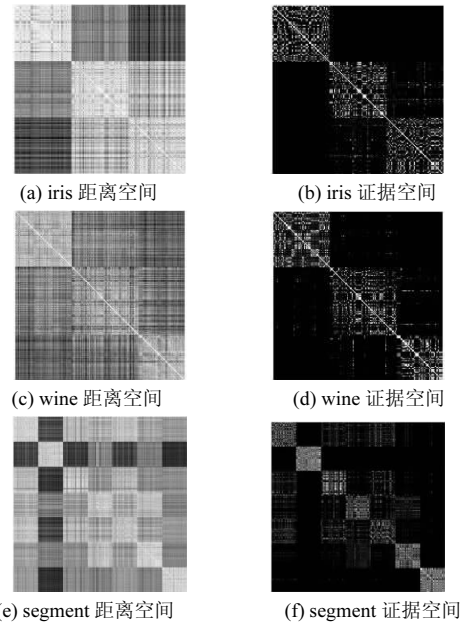
$$FM=\sqrt{\frac{a}{a+b}\frac{a}{a+c}} \tag{19}$$

对于两划分 π_1、π_2，a 表示既属于 π_1 中同一聚类，又属于 π_2 中同一聚类的样本对个数，b 表示属于 π_1 中同一聚类，但不属于 π_2 中同一聚类的样本对个数，c 表示不属于 π_1 中同一聚类，但属于 π_2 中同一聚类的样本对个数。FM 指标越大，表明两划分结果越接近，反之亦然。

实验共包括 3 部分：1）对比验证证据空间能够更好地描述数据空间结构；2）分析算法中涉及的选择规模 m 和平衡因子 λ 对集成结果的影响；3）将与现有选择方法对比，证明本文方法的优越性。

5.2　证据空间分析

本节首先在 iris、wine 和 segment3 个数据集上图形化分析证据空间与欧氏距离空间在表达数据结构时的性能差异，而后在表 1 所示的全体数据集上量化对比二者的数据结构描述性能。证据空间由 100 次差异性聚类成员通过证据积累获得，差异性聚类成员基于 k-means 聚类通过扰动类别参数[10, 60]和初始聚类中心点求得。3 个数据集在欧氏距离空间和证据空间的相似性度量矩阵如图 3 所示。



(a) iris 距离空间　(b) iris 证据空间
(c) wine 距离空间　(d) wine 证据空间
(e) segment 距离空间　(f) segment 证据空间

图 3　欧氏距离空间与证据空间图形化对比

对比图 3 中(a)和(b)、(c)和(d)、(e)和(f)可以看到，证据空间更有利于消除不同类数据间的相似性，能够较好描述数据集的空间结构。为进一步量化对比二者的优劣，基于真实类别分配，分别计算各数据集中同类数据的平均相关性 SS 和异类数据的平均相关性 SD，并计算二者比值 $\frac{SS}{SD}$，比值越大表明同类数据相关性高，而异类数据相关性低，从整体上表明该空间相似性描述更为合理，反之亦然，更多数据集的量化对比如表 2 所示。分析表 2 可以看到，证据空间中 $\frac{SS}{DD}$ 具有更高的比值，即证据空间能够更好地描述数据集的真实结构信息。

表 2　　欧氏距离空间与证据空间量化对比

数据集	距离空间			证据空间		
	SS	SD	$\frac{SS}{SD}$	SS	SD	$\frac{SS}{SD}$
iris	0.828 8	0.503 9	1.644 8	0.189 5	0.005 1	36.973
wine	0.644 3	0.440 0	1.464 3	0.186 1	0.004 2	43.991
soybean	0.527 2	0.305 8	1.723 7	0.095 4	0.013 6	6.998 9
sonar	0.486 2	0.463 1	1.049 8	0.082 5	0.040 4	2.043 8
ionosphere	0.637 1	0.539 8	1.180 3	0.154 0	0.039 3	3.920 0
vehicle	0.656 2	0.614 0	1.068 8	0.101 9	0.035 3	2.885 7
segment	0.778 3	0.536 9	1.450 0	0.305 8	0.0215	14.234
yeast1	0.798 3	0.743 1	1.074 3	0.120 8	0.048 8	2.478 0
2-ring	0.580 0	0.490 1	1.183 6	0.098 2	0	+∞
2-line	0.818 2	0.358 6	2.281 9	0.112 9	0	+∞
3-gauss	0.829 2	0.493 4	1.680 6	0.163 8	0.001 5	112.773
half-ring	0.767 9	0.282	2.720	0.100 6	0.000	5 030

5.3　参数讨论

如前所述，本文选择性集成中涉及的主要参数有选择规模 m 和平衡因子 λ，本节在表 1 中前 8 个数据集上讨论在不同选择规模和不同平衡因子对选择性集成的影响。全体聚类成员利用 k-means 算法通过初始聚类中心扰动和类别参数扰动[10,60]求得，规模为 100，利用规范化切割方法(NCUT, normalized CUT)对 CAM 进行二次分类。

首先令平衡因子 $\lambda=0.5$，即将有效性和差异性视为同等重要，观察选择规模 $m=10,20,\cdots,100$ 情况下选择集成的结果，实验结果如图 4 所示，图中横坐标为选择规模，纵坐标为集成结果与标准类标签

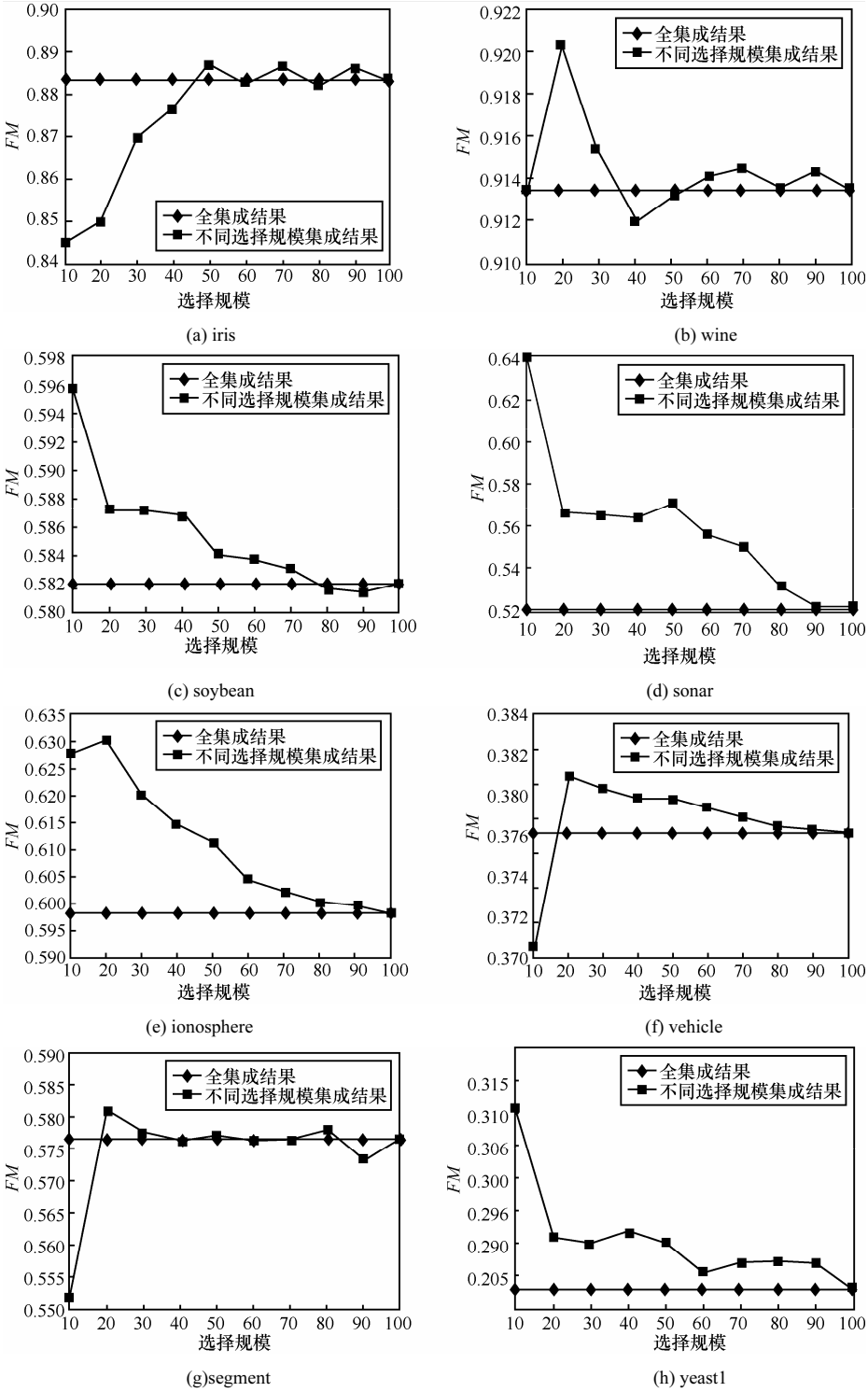


图 4　选择规模与集成结果间关系

间的 FM 指标。图中直线为全集成结果，曲线为不同选择规模情况下集成结果，每一点均为 20 次独立重复实验的平均值。分析图 4 可以看到对于不同的数据集，规模的影响不尽相同。按照曲线变化趋势可将数据集分为 3 类，第 1 类如 iris 数据集，选择效果并不理想，仅在选择规模较大(大于 50)时，选择性集成的分类效果与全集成接近或略有提高；第 2 类如 wine、vehicle、segment 数据集，当选择规模为 10 时，3 个数据集上的分类结果均不理想，但是当选择规模为 20 时，均取得了最优分类结果，而后随着选择规模的增大，选择性集成的分类性能逐渐降低并趋近于全集成；第 3 类如 soybean、sonar、ionosphere、yeast1 数据集，选择规模与分类结果间变化趋势较为明显，随着选择规模的增大，分类性能逐渐减小并趋近于全集成。

基于上述分析可以得到以下结论。

1) 与全集成方法相比，本文所提选择性集成可以获得更好的分类结果，但是对于不同的数据集，选择规模与集成结果间的关系不尽相同；

2) 较大的选择规模普遍能够获得较为稳定的分类结果，其分类性能通常略优于全集成的分类结果；

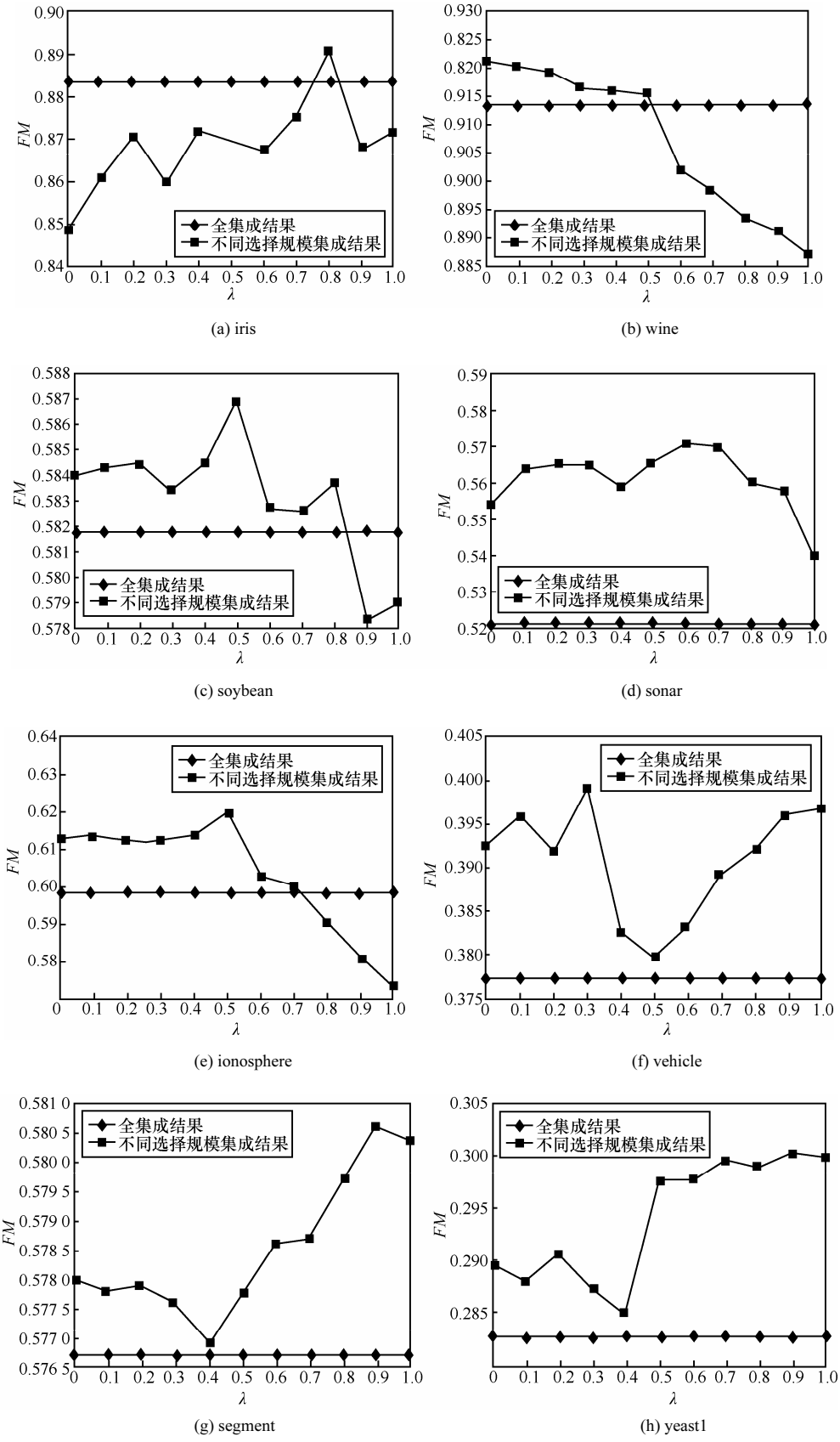
3) 较小的选择规模可以获得更高的分类性能，但是对于不同的数据集最佳选择规模并不相同。

其次考虑平衡因子 λ 对于选择性集成的影响，基于前述选择规模对集成结果的影响，这里取选择规模 $m=30$(具有较为普遍的有效性)，令 $\lambda=0,0.1,0.2,\cdots,1$，实验结果如图 5 所示，横坐标为 λ 的取值，其他描述与图 4 相同。图中 $\lambda=0$ 时表示仅利用差异性选择聚类成员，$\lambda=1$ 时表示仅利用有效性选择聚类成员。观察图 5 可以看到，在不同的数据集中，有效性和差异性对于选择集成的意义并不相同，如 iris、vehicle、segment、yeast1 这 4 个数据集中，仅利用有效性进行选择集成的结果要优于仅利用差异性进行选择集成，而 wine、soybean、sonar、ionosphere 这 4 个数据集中，有效性和差异性的表现相反。进一步分析上述子图可以看到，虽然各数据集中有效性和差异性对于选择性集成的重要性各不相同，但是当二者重要程度相对接近时更有利于获得相对稳定、有效的选择结果，如 wine、soybean、sonar、ionosphere、yeast1 这 5 个数据集中，$\lambda=0.5$ 获得相对较好的选择结果，对于 vehicle、segment 数据集，虽然 λ 在 0.5 附近选择结果相对较差，但是选择集成结果仍高于全集成结果。基于上述分析可以得到如下结论，选择性集成中令有效性和差异性的权重相近更有利于获得较好的选择结果。

5.4 对比实验

为进一步证明本文方法的优越性，进行如下实验：取平衡因子 $\lambda=0.5$，在 10、20、30 这 3 个选择规模上对比本文所提有效性指标 DB^{ES} 与原始 DB 指标以及 NMI 指标在聚类选择性集成中的有效性，其中 NMI 指标以各聚类成员结果与全集成结果间的归一化互信息来评价有效性。考虑到共识函数对集成结果的可能影响，分别利用 NCUT、AL(average link)2 种聚类方法对 CAM 进行聚类，实验结果如表 3 和表 4 所示，表中数据均为 20 次实验取均值，不同选择规模情况下最优选择集成结果均用黑体加粗。

观察表 3，当选择规模为 10、20、30 时，DB^{ES} 指标均在 7 个数据集上取得最佳选择集成结果。进一步在 3 种选择规模上对比 DB 指标和 DB^{ES} 指标可以看到，3 种选择规模情况下，DB^{ES} 指标均在 10 个数据集上取得不小于 DB 指标的选择集成结果。观察表 4，当选择规模为 10、20、30 时，DB^{ES} 指标分别在 7、7、8 个数据集上取得最佳选择集成结果。进一步在 3 种选择规模上对比 DB 指标和 DB^{ES} 指标可以看到，3 种选择规模情况下，DB^{ES} 指标均在 9 个数据集上取得不小于 DB 指标的选择集成结果。通过上述对比可以得到以下结论：1）DB^{ES} 指标与 DB 指标和 NMI 指标相比更有利于选择出较好的聚类成员；2）选择规模和基于 CAM 的聚类方法都对选择性集成的结果具有一定影响；3）表 2 中证据空间与欧氏距离空间的量化对比显示，证据空间更有利于保留数据集合的结构信息，但是将其与表 3 和表 4 的实验结果进行对比可以看到，在部分数据集中，这些合理的结构描述并没有带来更好的选择结果，究其原因，在证据空间中虽然类内相似性和类间相似性的比值更高，但是利用证据积累可能造成部分类内相关性信息的丢失(如图 2 和图 3 所示证据空间中，同类数据间相关信息的空缺)，这些信息的丢失不论对于全集成或选择性集成都是不利的。



(a) iris　(b) wine

(c) soybean　(d) sonar

(e) ionosphere　(f) vehicle

(g) segment　(h) yeast1

图 5　平衡因子与集成结果间关系

表 3　　NCUT 方法

数据集	选择规模：10			选择规模：20			选择规模：30		
	DB	*NMI*	DB^{ES}	*DB*	*NMI*	DB^{ES}	*DB*	*NMI*	DB^{ES}
iris	0.823 8	0.831 3	**0.844 7**	0.843 6	0.838 7	**0.850 3**	0.866 8	0.856 6	**0.869 9**
wine	0.899 4	**0.916 3**	0.912 8	0.909 9	0.917 9	**0.920 2**	0.912 9	0.913 7	**0.915 4**
soybean	0.588 3	0.585 6	**0.595 9**	**0.590 8**	0.585 0	0.587 3	0.586 5	0.585 2	**0.586 9**
sonar	0.540 9	0.616 0	**0.639 9**	0.563 7	**0.601 1**	0.567 1	**0.580 0**	0.571 4	0.565 8
ionosphere	**0.633 0**	0.628 2	0.627 4	0.615 8	0.624 3	**0.630 2**	0.607 0	**0.625 5**	0.620 2
vehicle	0.373 2	**0.377 5**	0.370 3	0.375 7	0.379 0	**0.380 4**	0.378 3	0.379 2	**0.379 7**
segment	0.550 5	**0.621 3**	0.551 2	0.581 0	**0.633 6**	0.591 5	0.577 7	**0.629 6**	0.587 1
yeast1	0.307 5	0.294 1	**0.310 8**	**0.308 2**	0.292 3	0.291 1	0.285 2	**0.292 5**	0.290 1
2-ring	0.693 7	**1.000 0**	**1.000 0**	0.919 2	**1.000 0**	**1.000 0**	**1.000 0**	**1.000 0**	**1.000 0**
2-line	**1.000 0**	**1.000 0**	**1.000 0**	**1.000 0**	**1.000 0**	**1.000 0**	**1.000 0**	**1.000 0**	**1.000 0**
3-gauss	0.949 6	**0.968 9**	0.966 1	0.966 2	**0.971 3**	0.968 0	**0.971 7**	0.971 3	0.970 9
half-ring	0.879 5	0.950 2	**1.000 0**	0.856 4	**1.000 0**	**1.000 0**	**1.000 0**	**1.000 0**	**1.000 0**

表 4　　AL 方法

数据集	选择规模：10			选择规模：20			选择规模：30		
	DB	*NMI*	DB^{ES}	*DB*	*NMI*	DB^{ES}	*DB*	*NMI*	DB^{ES}
iris	0.904 1	0.902 8	**0.912 7**	0.904 1	**0.913 9**	0.908 2	0.867 9	0.865 6	**0.899 3**
wine	0.853 6	0.883 6	**0.903 3**	0.856 8	0.900 3	**0.904 6**	0.834 9	**0.892 6**	0.879 3
soybean	**0.710 4**	0.590 2	0.599 4	**0.825 8**	0.595 2	0.587 0	**0.825 8**	0.619 7	0.579 7
sonar	**0.646 5**	0.591 7	0.627 3	**0.668 4**	0.615 5	0.626 8	**0.668 4**	0.629 7	0.632 2
ionosphere	**0.649 9**	0.667 9	0.646 3	0.632 5	0.662 5	**0.664 4**	0.695 0	0.658 5	**0.699 8**
vehicle	0.382 2	**0.385 7**	0.384 2	0.379 7	0.356 2	**0.382 2**	**0.374 4**	0.347 7	0.344 3
segment	0.613 5	**0.621 3**	0.620 2	0.600 1	**0.612 8**	0.610 7	0.610 4	0.614 3	**0.620 0**
yeast1	0.303 1	0.307 3	**0.343 2**	**0.343 1**	0.274 0	0.291 9	0.311 1	0.288 9	**0.315 7**
2-ring	**1.000 0**	**1.000 0**	**1.000 0**	**1.000 0**	**1.000 0**	**1.000 0**	**1.000 0**	**1.000 0**	**1.000 0**
2-line	**1.000 0**	**1.000 0**	**1.000 0**	**1.000 0**	**1.000 0**	**1.000 0**	**1.000 0**	**1.000 0**	**1.000 0**
3-gauss	0.961 2	0.974 7	**0.978 6**	0.953 4	0.966 1	**0.977 3**	0.853 9	0.961 3	**0.975 9**
half-ring	**1.000 0**	**1.000 0**	**1.000 0**	**1.000 0**	**1.000 0**	**1.000 0**	**1.000 0**	**1.000 0**	**1.000 0**

6　结束语

基于证据空间对于数据结构信息更为合理的描述，本文提出一种聚类的证据空间有效性评价指标 DB^{ES}，并将其应用于聚类的选择性集成中。实验结果显示，DB^{ES} 指标与原始的 *DB* 指标以及 *NMI* 有效性评价方法相比，具有更好的选择效果。需要指出的是，一方面本文仅对 *DB* 指标进行了证据空间扩展和研究，而实际中存在大量聚类的有效性评价方法，如何对它们进行合理的扩展具有广泛的意义；另一方面，证据空间虽然能够得到较为合理的数据空间结构描述，但是部分类内相关信息的丢失仍是不希望看到的，如何尽可能保留类内相关信息对于证据空间的稳定性和有效性具有重要意义，也必将成为后续研究的主要方向。

参考文献：

[1] ZHUANG W W, YE Y F, CHEN Y, *et al*. Ensemble clustering for internet security applications [J]. IEEE Transactions on Systems, Man, and Cybernetics-Part C: Applications and Reviews, 2012, 42(6): 1784-1796.

[2] GLLAVATA J, QELI E, FREISLEBEN B. Detecting text in videos using fuzzy clustering ensembles[A]. Proc of 8th IEEE International Symposium on Multimedia[C]. 2006.283-290.

[3] 卢志茂, 李纯, 张琦. 近邻传播的文本聚类集成谱算法[J]. 哈尔滨工程大学学报, 2012, 33(7): 899-905.

LU Z M, LI C, ZHANG Q. A document cluster ensemble spectral algorithm based on affinity propagation [J]. Journal of Harbin Engineering University, 2012, 33(7): 899-905.

[4] 邓晓政, 焦李成, 卢山. 基于非负矩阵分解的谱聚类集成SAR图像分割[J]. 电子学报, 2011, 39(12): 899-905.
DENG X Z, JIAO L C, LU S. Spectral clustering ensemble applied to SAR image segmentation using nonnegative matrix factorization[J]. Acta electronica sinica, 2011, 39(12): 899-905.

[5] ALEXANDER S, JOYDEEP G. Cluster ensembles-a knowledge reuse framework for combining partitionings[A]. AAAI[C]. 2002. 93-98.

[6] FRED A L N, JAIN A K. Combining multiple clusterings using evidence accumulation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(6): 835-850.

[7] TOPCHY A, JAIN A K, WILLIAM P. Clustering ensembles: models of consensus and weak partitions[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(12): 1866-1881.

[8] 罗会兰, 孔繁胜, 李一啸. 聚类集成中的差异性度量研究[J]. 计算机学报, 2007, 30(8): 1315-1324.
LUO H L, KONG F S, LI Y X. An analysis of diversity measures in clustering ensembles[J]. Chinese Journal of Computers, 2007, 30(8): 1315-1324.

[9] HADJITODOROV S T, KUNCHEVA L I, TODOROVA L P. Moderate diversity for better cluster ensembles[J]. Information Fusion, 2006, 7: 264-275.

[10] HONG Y, SAM K, WANG H L, *et al*. Resampling-based selective clustering ensembles [J]. Pattern Recognition Letters, 2009, 30: 298-305.

[11] JIA J H, XIAO X, LIU B X, *et al.* Bagging-based spectral clustering ensemble selection[J]. Pattern Recognition Letters, 2011, 32: 1456-1467.

[12] NALDI M C, CARVALHO A C P L F, CAMPELLO R J G B. Cluster ensemble selection based on relative validity indexes[J]. Data Min Knowl Disc, 2013, 27: 259-289.

[13] 周林, 平西建, 徐森等. 基于谱聚类的聚类集成算法[J]. 自动化学报, 2012, 38(8): 1335-1342.
ZHOU L, PING X J, XU S, *et al*. Cluster ensemble based on spectral clustering[J]. Acta Automatica sinica, 2012, 38(8): 1335-1342.

[14] LOURENGCO A, BULO S R, FRED A, *et al.* Consensus clustering with robust evidence accumulation[A]. EMMCVPR[C]. LNCS, 2013. 307-320.

[15] DAVIES D L, BOULDIN D W. A cluster separation measure[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1979, 1(2): 224-227.

[16] ZHOU Z H, WU J X, TANG W. Ensembling neural networks: Many could be better than all[J]. Artificial Intelligence, 2002, 137(1-2): 239-263.

作者简介：

毕凯（1985-），男，河南南阳人，空军工程大学博士生，主要研究方向为模式识别与智能信息处理。

王晓丹（1966-），女，陕西汉中人，空军工程大学教授、博士生导师，主要研究方向为智能信息处理和机器学习等。

邢雅琼（1986-），女，陕西渭南人，空军工程大学博士生，主要研究方向为智能信息处理和机器学习等。