

聚类集成中的差异性度量研究

罗会兰^{1),2)} 孔繁胜¹⁾ 李一啸¹⁾

¹⁾(浙江大学人工智能研究所 杭州 310027)

²⁾(江西理工大学信息工程学院 江西 赣州 341000)

摘 要 集体的差异性被认为是影响集成学习的一个关键因素. 在分类器集成中有许多的差异性度量被提出, 但是在聚类集成中如何测量聚类集体的差异性, 目前研究得很少. 作者研究了7种聚类集体差异性度量方法, 并通过实验研究了这7种度量在不同的平均成员聚类准确度、不同的集体大小和不同的数据分布情况下与各种聚类集成算法性能之间的关系. 实验表明: 这些差异性度量与聚类集成性能间并没有单调关系, 但是在平均成员准确度较高、聚类集体大小适中和数据中有均匀簇分布的情况下, 它们与集成性能间的相关度还是比较高的. 最后给出了一些差异性度量用于指导聚类集体生成的可行性建议.

关键词 集成学习; 聚类集成; 差异性; 度量

中图法分类号 TP181

An Analysis of Diversity Measures in Clustering Ensembles

LUO Hui-Lan^{1),2)} KONG Fan-Sheng¹⁾ LI Yi-Xiao¹⁾

¹⁾(Institute of Artificial Intelligence, Zhejiang University, Hangzhou 310027)

²⁾(School of Information Engineering, Jiangxi University of Science and Technology, Ganzhou, Jiangxi 341000)

Abstract The diversity of an ensemble is known to be an important factor in determining its performance. There are a number of ways to quantify diversity in ensembles of classifiers, while little research has been done in clustering ensembles. This paper compares seven diversity measures of clustering ensembles with regard to their possible use in ensemble design. Five experiments have been designed to examine the relationships between the accuracy of the clustering ensembles and the measures of diversity under conditions of difference ensemble methods, different ensemble size and different data distributions respectively. Experiments show the relationships between these diversity measures and ensemble performances are not monotonous. However, when constructing ensembles with moderate ensemble size by suitable clustering algorithms for a given data set with uniform cluster distribution, the correlation coefficients between the diversity measures and ensemble performances are relatively high. Finally, the authors give some useful suggestions about the usefulness of diversity measures in building clustering ensembles.

Keywords ensemble learning; clustering ensemble; diversity; measure

1 引 言

在集成学习中, 只有当集体中的成员在一些输

出上不一致时, 结合它们才能有效提高学习效率^[1-2], 这种不一致性称为集体的差异性. 当集体的差异性增加, 也就是成员学习器之间的关联性越低时, 集成学习的优势就越明显. 就差异性应如何来严

格定义,如何来测量,在分类器集成环境中有一些比较深入的研究^[3-4],但在聚类集成环境中,这方面的研究还比较少。

在聚类集成环境中,Fern 等^[5]中使用规范化互信息(Normalized Mutual Information,NMI)来表示聚类集体差异程度,并且绘制出聚类对之间的平均准确度对聚类对间的 NMI 值的点图.实验表明当点对之间的 NMI 值分布较宽,也就是说集体内的差异性大,且集体的准确度也较高时,能得到较好的聚类性能.Tang 等^[6]中也使用了 NMI 值来选择聚类成员以提高聚类集成效果.Greene 等^[7]中使用了一个基于熵的方法来测量差异性,对 3 种不同的集体生成方法生成的集体进行了差异性对比,用实验说明了差异性在聚类集成时的关键作用.在文献^[8]中,作者基于 adjusted Rand index 提出了 4 种测量聚类集体差异性的方法,但作者在实验中发现,如果以这 4 种方法来衡量聚类集体差异性,则选择差异性为中间值的集体进行集成比选择最大差异性的集体更能保证集成性能。

我们提出了 4 种新的可用于测量聚类集体差异性的度量,然后对这 4 种差异性度量加上基于 NMI^[5-6],基于熵^[7]和基于 adjusted Rand index^[8]的聚类集体差异性度量,与 6 种聚类集成方法的准确度之间的关系进行了全面的实验研究.为了研究在不同情况下各种差异性度量与不同的集成方法性能之间的关系,我们在实验中采用了人工模拟生成具有各种特点的聚类集体的方法。

本文第 2 节介绍 7 类聚类集体差异性度量方法和它们的来源;第 3 节首先分别在不同的平均成员聚类准确度、聚类集体大小和不同的数据源分布条件下,研究各种差异性度量与同一集成方法性能之间的关系,然后研究在聚类成员准确度相同情况下各种聚类集体差异性度量与不同聚类集成方法间的关系;第 4 节总结实验结果。

2 聚类集成中差异性测量方法

假设有一个待聚类的势为 N 的数据集 $X = \{x_1, x_2, \dots, x_N\}$,在此数据集上产生了一个聚类集体(或称为集合) $\Pi = \{\pi_1, \pi_2, \dots, \pi_L\}$,它有 L 个聚类成员.针对这个聚类集体有如下差异性测量方法。

2.1 基于 NMI 的差异性度量 NMIBDM

互信息(Mutual Information, MI)是一个用来测量两个分布的共享统计信息的对称度量.如果用

X 和 Y 表示两个随机变量,则这两个随机变量间的互信息(Mutual Information, MI)定义为^[9]

$$I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \quad (1)$$

为了将这个度量值的范围限于 0 和 1 之间,在文献^[10]中作者引入了规范化互信息(Normalized Mutual Information, NMI):

$$NMI = \frac{I(X,Y)}{\sqrt{H(X)H(Y)}} \quad (2)$$

其中 $H(X)$ 和 $H(Y)$ 分别表示 X 和 Y 两个随机变量的熵值,其中熵值的定义如下^[9]

$$H(X) = \sum_x p(x) \log \frac{1}{p(x)} \quad (3)$$

所以如果有两个聚类 π_a, π_b ,则它们之间的 NMI 值定义为

$$\begin{aligned} NMI(\pi_a, \pi_b) &= \frac{\sum_{h=1}^{k_a} \sum_{l=1}^{k_b} \frac{n_{h,l}}{N} \log_2 \left(\frac{\frac{n_{h,l}}{N}}{\frac{n_h}{N} \cdot \frac{n_l}{N}} \right)}{\sqrt{\left(\sum_{h=1}^{k_a} \frac{n_h}{N} \log_2 \frac{n_h}{N} \right) \left(\sum_{l=1}^{k_b} \frac{n_l}{N} \log_2 \frac{n_l}{N} \right)}} \\ &= \frac{\sum_{h=1}^{k_a} \sum_{l=1}^{k_b} n_{h,l} \log_2 \left(\frac{N \cdot n_{h,l}}{n_h \cdot n_l} \right)}{\sqrt{\left(\sum_{h=1}^{k_a} n_h \log_2 \frac{n_h}{N} \right) \left(\sum_{l=1}^{k_b} n_l \log_2 \frac{n_l}{N} \right)}} \end{aligned} \quad (4)$$

其中 k_a 和 k_b 分别表示聚类 π_a, π_b 中簇(cluster)的个数; $n_{h,l}$ 表示同时位于 π_a 的 h 簇和 π_b 的 l 簇中点的个数; n_h 表示 π_a 的 h 簇中点的个数; n_l 表示 π_b 的 l 簇中点的个数. NMI 值是一个对称的用于测量两个聚类之间的共享信息的一种度量,为了测量一个聚类集体的差异性,在文献^[5]中使用聚类集体中所有聚类对之间的平均 NMI 值来表示集体的差异性:

$$NMI_{\text{平均}} = \frac{2}{L(L-1)} \left(\sum_{i=1}^{L-1} \sum_{j=i+1}^L NMI(\pi_i, \pi_j) \right) \quad (5)$$

由于这个值越大表示差异性越小,所以我们稍微做了一点改动,使用如下式子来表示聚类集体的差异性,我们将其命名为 NMIBDM(NMI Based Diversity Measure):

$$NMIBDM = 1 - NMI_{\text{平均}} \quad (6)$$

这样当 NMIBDM 值越大时,也表示聚类集体的差异性越大.它的值位于 0 和 1 之间,当两个聚类完全等同时,它的值为 0。

2.2 基于熵的差异性度量 EBDM

在文献^[7]中使用了一个基于熵的方法来测量聚

类集体的差异性,我们把它命名为 EBDM(Entropy Based Diversity Measure):

$$EBDM = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^N - (p_{ij} \log_2 p_{ij} + (1-p_{ij}) \log_2 (1-p_{ij})) \quad (7)$$

p_{ij} 表示点对 x_i 和 x_j 在聚类集体中被分在同一个簇的概率: $p_{ij} = \frac{1}{L} \sum_{k=1}^L \delta(\pi_k(x_i), \pi_k(x_j))$, 其中 $\pi_k(x_i), \pi_k(x_j)$ 分别是在聚类 π_k 中点 x_i 和 x_j 的簇标签, 如果它们相等则 δ 函数值为 1, 否则为 0.

EBDM 值位于 0 和 1 之间, 越大表示差异性也越大.

2.3 基于 CE(Conditional Entropy) 的差异性度量 CEBDM

在文献[5]中用 CE 来评估一个聚类的性能, CE(Conditional Entropy)定义如下

$$CE(\pi_a, \pi_b) = \sum_{l=1}^{k_b} \frac{n_l \cdot (-\sum_{h=1}^{k_a} p_{hl} \log(p_{hl}))}{N} \quad (8)$$

其中 p_{hl} 表示 π_b 的 l 簇中的点属于 π_a 的 h 簇的概率.

我们基于这个概念定义了一个用来测量聚类集体的差异性度量, 称其为 CEBDM(CE Based Diversity Measure):

$$CEBDM = \frac{2}{L(L-1)} \left(\sum_{i=1}^{L-1} \sum_{j=i+1}^L (CE(\pi_i, \pi_j) + CE(\pi_j, \pi_i)) / 2 \right) \quad (9)$$

这个值越大表示聚类集体差异性越大.

2.4 基于 Adjusted Rand Index 的度量

文献[8]的作者基于 adjusted Rand index^[11] 定义了用于测量两个聚类之间差异性的度量. 其中 adjusted Rand index 定义为

$$ar(\pi_a, \pi_b) = \frac{\sum_{h=1}^{k_a} \sum_{l=1}^{k_b} \binom{n_{h,l}}{2} - t_3}{\frac{1}{2}(t_1 + t_2) - t_3} \quad (10)$$

$$t_1 = \sum_{h=1}^{k_a} \binom{n_h}{2}, \quad t_2 = \sum_{l=1}^{k_b} \binom{n_l}{2}, \quad t_3 = \frac{2t_1 t_2}{N(N-1)},$$

其中 $k_a, k_b, n_{h,l}, n_l$ 和 n_h 与式(4)中的意思一样. 当两个聚类完全独立时, adjusted Rand index 的值应该为 0, 当接近于 0 时, 说明知道一个聚类对预测另一个聚类没有一点作用. 所以当为 0 时并不意味着两个聚类没有一点共同之处. 而当两个聚类同时, 它的值为 1. 利用 adjusted Rand index 定义了以下 5 个聚类集体的差异性度量 adRBDM(adjust Rand

index Based Diversity Measure):

$$adRBDM_1 = \frac{2}{L(L-1)} \sum_{i=1}^{L-1} \sum_{j=i+1}^L (1 - ar(\pi_i, \pi_j)) \quad (11)$$

$$adRBDM_2 = \frac{1}{L} \sum_{i=1}^L (1 - ar(\pi_i, \pi^*)) \quad (12)$$

$$adRBDM_3 = \sqrt{\frac{1}{(L-1)} \sum_{i=1}^L (1 - ar(\pi_i, \pi^*) - adRBDM_2)^2} \quad (13)$$

$$adRBDM_4 = \frac{1}{2} (1 - adRBDM_2 + adRBDM_3) \quad (14)$$

$$adRBDM_5 = \frac{adRBDM_3}{adRBDM_2} \quad (15)$$

$adRBDM_1$ 先测量每对聚类的差异性, 因为 $ar(\pi_a, \pi_b)$ 表示两个聚类的相似性, 所以 $1 - ar(\pi_a, \pi_b)$ 表示两个聚类之间的差异性, 然后计算出平均值作为集体的差异性值. 文献[8]的作者先用大多数投票法得到一个集成聚类 π^* , 然后以它为基准, 计算集体中的成员聚类与它的差异值, 然后平均得到 $adRBDM_2$, 在第 3 节的实验中我们用 CSPA 算法^[10]对聚类集体进行集成得到的聚类结果来表示 π^* . $adRBDM_3$ 用各个聚类与 π^* 的差异性的标准偏差来表示集体的差异性. $adRBDM_4$ 和 $adRBDM_5$ 则是综合 $adRBDM_2$ 和 $adRBDM_3$ 来表示集体的差异性, 当把 π^* 当成正确聚类时, $adRBDM_2$ 越小时表示聚类集体平均聚类成员准确度越高, 所以在这种意义下 $adRBDM_4$ 和 $adRBDM_5$ 度量同时反映集体的准确度和差异性.

如果用 CSPA 算法^[10]对聚类集体进行集成得到的聚类结果 π^* 作为基准, 对其它聚类的簇标签进行匹配统一, 则可以利用分类集成环境中的差异性测量方法来度量聚类集体的差异性. 文献[3]的作者对分类器集体差异性的十种测量方法进行了聚类分析, 将它们分成了三个簇, 我们从每个簇中选择了一种方法, 基于它们(double fault, coincident failure diversity, interrater agreement)来定义聚类集体的差异性度量.

2.5 基于 Double Fault Measure 的差异性度量 DFBDM

文献[12]的作者使用 Double Fault Measure 来形成分类集体的成对差异性矩阵, 用它来指导选择最不相关的分类器. 我们基于 Double Fault Measure 定义

了一个聚类集体差异性度量方法 DFBDM(Double Fault Based Measure)

$$DFBDM = 1 - \frac{2}{NL(L-1)} \sum_{i=1}^{L-1} \sum_{j=i+1}^L n_{i,j}(-1, -1)$$

(16)

其中 $n_{i,j}(-1, -1)$ 表示聚类 π_i 和 π_j 同时与 π^* 意见不同的点的个数. $DFBDM$ 值为 0 时, 表示聚类集体的差异性最小, 它的值越大表示差异性越大.

2.6 基于 CFD(Coincident Failure Diversity) 的差异性测量 CFDBDM

我们基于 Partridge 和 Krzanowski 提出的一种度量 Coincident Failure Diversity^[13] 定义了 CFDBDM(Coincident Failure Diversity Based Diversity Measure)

$$CFDBDM = \begin{cases} 0, & p_0 = 1 \\ \frac{1}{1-p_0} \sum_{i=1}^L \frac{L-i}{L-1} p_i, & p_0 < 1 \end{cases}$$

(17)

其中 p_i 表示 L 个聚类中刚好有 i 个与 π^* 意见不相同的点在整个数据集中所占的比例, p_0 相应表示有 0 个聚类与 π^* 意见不相同的点的比例. 当所有的聚类等同时 $CFDBDM$ 的值为 0; 而当对于所有的数据点至多有一个聚类与 π^* 意见不同时, 它的值为 1. $CFDBDM$ 的值越大表示差异性越大.

2.7 基于 Measurement of Inter-Rater Agreement 的差异性度量 IRABDM

文献[14]的作者用 Measurement of Inter-Rater Agreement 来测量分类器间的可信性, 也称为 κ , 它可用来测量分类器集体中的一致性水平. 我们基于 Measurement of Inter-Rater Agreement 定义了一个聚类集体差异性度量方法 IRABDM(Inter-Rater Agreement Based Diversity Measure)

$$IRABDM = \frac{\sum_{i=1}^N (L - l_i) l_i}{NL(L-1)P(1-P)}$$

(18)

其中 l_i 表示就数据点 x_i , 聚类集体中与 π^* 意见不同的聚类成员个数. 而 P 表示以 π^* 作为标准, 聚类集体的平均聚类准确率. $IRABDM$ 的值越大表示聚类集体的差异性越大.

3 聚类集体差异性度量分析

3.1 聚类集体差异性与聚类集成准确度之间的散点图

为了直观地看出上述各种差异性度量与聚类集

成的准确度之间的关系, 我们采用模拟生成聚类集体的方法. 首先生成一个大小为 30 的一维矢量表示在数据规模为 30 的数据集上的一个真实聚类, 每个簇 10 个数据点. 例如, 用 10 个 1, 10 个 2 和 10 个 3 表示的一个真实聚类为 $[1, 1, \dots, 1, 2, \dots, 2, 3, \dots, 3]$. 然后在此基础上随机改变其中的 $30 \times (1-p)$ 个点的簇标签, 使之与原来的标签不同来模拟生成一个准确度为 p 的聚类, 例如改变第一个点的簇标签为 2, 则这个聚类成为 $[2, 1, \dots, 1, 2, \dots, 2, 3, \dots, 3]$. 使用此方法我们构造 300 个大小为 3、平均准确度为 0.6 的聚类集体且计算这 300 个集体的差异性度量值和使用 CSPA 算法^[10] 集成之后的聚类准确度. CSPA^[10] 集成算法将每个数据点表示成一个顶点, 两个点被分在同一个簇中的次数占聚类集体中所有簇的比例为相应两顶点间边的权重, 这样根据一个聚类集体生成一个图后, 再利用图形划分方法即得到最终聚类结果. 之所以选用 CSPA 算法^[10] 是因为在以往的聚类集成对比实验研究中它的性能比较稳定, 而且准确度较高. 而为了要计算集成准确度, 需要对簇标签进行匹配, 在这里我们选用的是 Hungarian 算法^[15], 它根据两个聚类的簇对间的相同数据点比例进行簇标签匹配.

针对不同的差异性度量方法我们得到了 11 组数据, 将其绘成了散点图(如 3.6 节图 1 所示), 图中每个点表示一个聚类集体. 在这种情况下造成集成性能各不相同的因素应该来说只有各个聚类集体的差异性, 因为各个聚类集体的平均成员准确度相同, 集体大小相同, 集成方法相同.

从图 1 可以看出没有一种差异性度量与 CSPA^[10] 聚类集成准确度之间存在严格单调递增关系. 也就是说这些差异性度量值越大并不意味着用 CSPA^[10] 进行聚类集成时准确度就越高. 这其中的原因有可能是这些差异性度量并没有很好地把我们直觉上所认为的聚类成员间的独立性和不会犯同样错误的那种概念表达出来. 也有可能是 CSPA^[10] 这种集成方法不能很好地利用集体的差异性.

3.2 不同的平均聚类成员准确度情况下集体差异性度量与集成准确度之间的关系

为了验证集体的差异性度量与集成性能之间的关系是否受到集体平均成员准确度的影响, 我们利用 3.1 节描述的实验方法产生了具有不同平均成员准确度的集体, 对每一个准确度我们生成了 30 个大小为 3 的集体, 从而计算出由这 30 个集体产生的差异性与 CSPA^[10] 集成准确度之间的相关系数, 重复

20 次这样的过程,得到了这些相关系数的平均值.表 1 列出了各种聚类集体差异性度量与 CSPA^[10]集成准确度之间的相关系数,表头中, Ep 表示在各种平均成员准确度情况下,平均 CSPA^[10]集成准确度.在实验中我们使用了 Spearman's rank Correlation Coefficient(RCC)方法来估算差异性与集成准确度之间的相关程度.RCC 是由 Spearman 在 1904 年提出的非参数的,即与分布无关的相关性估算方法,它用来测量两个变量的关联强度^[16].之所以用它是认为差异性度量与集成准确度之间的关系是非

线性的^[3],它的值越大表示相关性越高. RCC 的定义如下

$$RCC(X,Y)=1-6\cdot\sum_{i=1}^N\frac{(Rank(x_i)-Rank(y_i))^2}{N(N^2-1)}$$

(19)

这里 $X=\{x_1,x_2,\cdots,x_N\}$ 和 $Y=\{y_1,y_2,\cdots,y_N\}$ 是实数的集合,计算这个 RCC 参数,我们使用了 matlab 自带的函数.它的值从 $-1\sim 1$,为 0 时表示两个变量是独立的.负数表示负相关,正数表示正相关,绝对值越大表示相关性越高.

表 1 不同平均成员准确度情况下各种差异性度量与 CSPA 集成准确度的相关系数
(其中的 p 表示集体中成员聚类的准确度, Ep 表示平均 CSPA 集成准确度)

	相关系数					
	$p=0.5$ $Ep=0.4996$	$p=0.55$ $Ep=0.5647$	$p=0.6$ $Ep=0.6049$	$p=0.7$ $Ep=0.7533$	$p=0.8$ $Ep=0.8960$	$p=0.9$ $Ep=0.9743$
NMIBDM	-0.0435	-0.0784	-0.0525	0.1252	0.4702	0.6016
EBDM	-0.0291	-0.0874	-0.0705	0.1116	0.5043	0.7389
CEBDM	-0.0442	-0.0491	-0.0291	0.1317	0.4507	0.5916
adRBDM_1	-0.0447	-0.0738	-0.0518	0.1316	0.5136	0.7421
adRBDM_2	-0.0678	-0.1572	-0.1880	-0.0725	0.2820	0.4096
adRBDM_3	0.0500	0.0627	-0.0071	-0.3661	-0.5429	-0.8303
adRBDM_4	0.0629	0.1012	0.0437	-0.3180	-0.5964	-0.8259
adRBDM_5	0.0549	0.0767	0.0088	-0.3559	-0.5633	-0.8347
DFBDM	0.0365	0.1030	0.1908	0.4553	0.4878	0.5714
CFDBDM	0.0316	0.0763	0.1562	0.4428	0.5650	0.5445
IRABDM	0.0206	-0.0049	0.0409	0.3479	0.5984	0.7713

从表 1 可以看出(进行横向比较时),随着聚类成员平均准确度的增加,各种度量与集成准确度之间的相关系数的绝对值也在增加.在成员聚类准确度 ≤ 0.6 时,各种差异性度量与集成性能之间的相关性很低,都接近于 0.而当高于 0.6 时, $adRBDM_3$, $adRBDM_4$ 和 $adRBDM_5$ 从相关程度很低变成了负相关,也就是说差异性值的增加反而意味着集成准确度降低,这可能是因为这三个度量将成员聚类准确度作为度量的一个因素的原故.其它的度量在平均成员准确度高于 0.6 时,与集成性能之间的相关性都呈现明显的正相关关系.并且发现平均成员准确度 0.6 像是一个分水岭,各种差异性度量在此点上由相关性几乎为 0 转变成相关性较高.这说明当一个集体中成员聚类的准确度较高时,用大多数的差异性度量来指导一个聚类集体的生成或评价一个聚类集体的好坏应该说还是有一定的可取性的.从表 1 也发现随着平均成员准确度的增加,集成性能相对于成员聚类的提高幅度也在增加.

3.3 不同的集体大小情况下各种差异性度量与 CSPA 集成准确度的相关性研究

为了研究集体的大小是否也会影响差异性与集成准确度间的相关性,我们使用 3.1 节描述的实验

方法分别生成在不同的聚类集体大小情况下,平均成员准确度 $p=0.65$ 的集体 30 个,从而计算在不同的集体大小情况下集体的差异性和集成准确度的相关系数,重复 20 次此过程计算得到的平均值如表 2 所示.表头中的 Ep 也如表 1 一样表示在相应情况下 CSPA 集成准确度的平均值,集体大小用 L 表示.

从表 2 可以看出集体大小确实影响着差异性与集成准确度间的相关性,随着集体大小的增大,虽然平均 CSPA 集成准确度在增加,但各种差异性度量与集成性能之间的相关性却不总是在增加,而是一开始随着集体大小的增大相关性在增加,但当集体大小增加到一定程度时,这种相关性却又开始有所下降.大多数的度量在集体大小位于 $15\sim 20$ 之间时,与集成性能的相关性最强. $adRBDM_3$ 和 $adRBDM_5$ 在集体大小 $L=20$ 时相关系数值分别达到了 -0.9076 和 -0.9031 ,呈现了相当强的负相关关系.而 $DFBDM$, $CFDBDM$ 和 $IRABDM$ 度量在 $L=20$ 时与 CSPA 集成性能的正相关程度也达到了较高的值.这也给了我们一点启示,当我们想用差异性度量来指导生成差异性集体时,最好生成大小为 $15\sim 20$ 的集体.

表 2 不同的集体大小情况下各种差异性度量与 CSPA 集成准确度的相关系数

	相关系数					
	$L=3$	$L=5$	$L=8$	$L=15$	$L=20$	$L=30$
	$Ep=0.6469$	$Ep=0.7146$	$Ep=0.8022$	$Ep=0.9211$	$Ep=0.9553$	$Ep=0.9858$
NMIBDM	-0.0458	0.1042	0.2393	0.2866	0.2305	0.1603
EBDM	-0.0656	0.0917	0.2206	0.2292	0.2089	0.1290
CEBDM	-0.0188	0.0836	0.2075	0.1923	0.1491	0.0855
adRBDM_1	-0.0428	0.1042	0.2469	0.2887	0.2173	0.1871
adRBDM_2	-0.2034	-0.1435	-0.0351	-0.1160	-0.1876	-0.2598
adRBDM_3	-0.0905	-0.3241	-0.5962	-0.8782	-0.9076	-0.7770
adRBDM_4	-0.0346	-0.2425	-0.4711	-0.7684	-0.7945	-0.6695
adRBDM_5	-0.0725	-0.3083	-0.5822	-0.8728	-0.9031	-0.7740
DFBDM	0.2832	0.4386	0.4608	0.6981	0.7519	0.6733
CFDBDM	0.2367	0.3854	0.4018	0.6071	0.7208	0.7041
IRABDM	0.1331	0.2825	0.3814	0.5750	0.6499	0.5581

3.4 数据分布对差异性度量与聚类集成的准确度之间的关系的分析

为了研究数据的分布是否会影响聚类集成效果与差异性度量之间的相关程度,我们使用 3.1 节描述的实验方法人工生成了不同数据分布上的真实聚类,其中模拟的数据分布如表 3 所示.在此基础上进行随机扰动,产生不同数据分布上、大小为 3、平均成员准确度为 0.7 的聚类集体 30 个,分别计算它们的差异性度量值和用 CSPA 算法进行集成的准确度,然后分别计算各种差异性度量与集成性能之间

的相关系数.此过程重复 20 次,得到的各种差异性度量与集成性能之间在不同的数据分布上的相关系数如表 4 所示.

表 3 6 个不同的数据分布

	数据量	簇的个数	每簇中点的个数
Dataset1	60	3	20;20;20
Dataset2	60	3	40;10;10
Dataset3	210	3	70;70;70
Dataset4	210	3	125;50;35
Dataset5	1200	3	400;400;400
Dataset6	1200	3	100;1000;100

表 4 不同数据分布类型上的差异性度量与 CSPA 集成准确度之间的相关系数

	相关系数					
	Dataset1	Dataset2	Dataset3	Dataset4	Dataset5	Dataset6
	$Ep=0.7728$	$Ep=0.5436$	$Ep=0.7997$	$Ep=0.5371$	$Ep=0.8018$	$Ep=0.4134$
NMIBDM	0.1442	-0.0153	0.2620	-0.1262	0.3732	-0.1321
EBDM	0.1346	-0.1163	0.3007	-0.1294	0.3602	-0.1953
CEBDM	0.1582	0.0743	0.2596	-0.1272	0.3682	0.2102
adRBDM_1	0.1387	-0.0294	0.2965	-0.1602	0.3677	-0.0933
adRBDM_2	0.0918	-0.3718	0.0572	-0.6303	-0.0924	-0.3014
adRBDM_3	-0.5813	0.0032	-0.5681	-0.0423	-0.4516	0.0332
adRBDM_4	-0.6265	0.1955	-0.5799	0.2017	-0.4070	0.1295
adRBDM_5	-0.5863	0.0052	-0.5797	-0.0208	-0.4545	0.0405
DFBDM	0.4799	-0.0005	0.5614	0.1095	0.5092	0.3882
CFDBDM	0.5459	-0.0613	0.6230	0.1120	0.5255	0.1624
IRABDM	0.4291	-0.1909	0.7656	-0.1224	0.6053	-0.1210

从表 4 可以看到,数据量、数据中簇的分布等等不同的数据源情况确实影响着差异性度量与 CSPA 集成性能的相关性.在同样数据量情况下,有均匀的簇分布时,除了 adRBDM_2 度量,其它度量与 CSPA 集成性能的相关性比非均匀的簇分布时强许多,而且相应的 CSPA 集成性能也比在非均匀的簇分布时高得多.而对于 adRBDM_2 度量,情况与其它度量相反,它是在数据具有非均匀的簇分布时,与 CSPA 集成性能相关性更强.这可能与它测量集体的差异性值的方法有关.在同样的均匀簇分布情况下,数据量不同时,各种差异性度量与集成准确度之间的相关度各不相同,对于 NMIBDM,EBDM,CEBDM 和 adRBDM_1,随着数据量的增加,相关

性也在增加.但对于其它的度量,数据量对它们虽然也有影响,但这种影响是不确定的.

3.5 不同聚类集成算法与差异性度量之间的关系

为了研究不同的聚类集成算法利用集体差异性的能力是否不同,或者说不同的聚类集成方法与差异性之间的相关性不同,我们产生 30 个成员聚类准确度 $p=0.75$,大小为 10 的聚类集体,然后计算它们的各种差异性度量值,并用不同的聚类集成方法对其进行集成,从而计算出不同的集成准确度,在此基础上计算它们之间的相关系数.表 5 是采用这种方法运行 20 次得到的平均值.表头中的 Ep 表示在相应情况下各种集成方法的平均准确度.

在这里采用的集成方法有在文献[10]中的三种

集成方法 MCLA, HGPA 和 CSPA, 这三种算法的代码也是文献[10]作者 Strehl 个人网站上的代码; 还有 Majority Vote 方法, 也就是将集体中的聚类先进行标签的匹配和统一后用投票方法进行聚类; 还有就是将聚类集体看成概念型数据, 用概念型数据聚类算法 Kmodes^[17]进行聚类集成^[18]; 还有 ALBE 方法, 它表示将聚类集体先转化成数据点对之间的相似性矩阵^[19], 然后在此基础上用层次聚类算法 average-link 进行聚类集成。

从表 5 可以看出, 所有的集成算法的平均准确度都高于集体的平均成员准确度, 这也说明了聚类集成的优势. 我们注意到对于 NMIBDM, EBDM, CEBDM, adRBDM_1 和 adRBDM_2 度量, 与它们相关性最高的集成方法是 MCLA. 其它的差异性度量都与 CSPA 集成性能的相关性最强, 特别是 adRBDM_3, adRBDM_4 和 adRBDM_5 度量与 CSPA 集成性能的相关性达到了 -0.8, 非常强的负相关关系. 而 DFBDM 和 IRABDM 与 CSPA 集成

性能的相关性也达到了 0.6 以上, 较强的正相关关系. 同时我们注意到 MCLA, CSPA 和 Majority Vote 算法不但与各种差异性度量间有相对来说更高的相关性, 而且它们的平均集成准确度也较高, 而 HGPA, Kmodes 和 ALBE 与各种差异性度量的相关系数与它们相比更低, 相应的集成准确度也较低. 这说明能正确地捕捉到聚类集体差异性的集成算法的性能也相应更好. 同时我们从表中也观察到集成方法不同, 和它相关性最强的差异性度量也不同, 与 MCLA 相关性最高的差异性度量是 adRBDM_1, HGPA 是 adRBDM_2, CSPA 负相关最强的差异性度量是 adRBDM_3, 正相关是 DFBDM. Majority-Vote 和 Kmodes 负相关最强的差异性度量是 adRBDM_4. ALBE 相关性最强的差异性度量是 IRABDM. 这说明每种集成算法有自己适用的差异性度量来指导聚类集体的生成, 也许定义一个对所有的集成算法都适用的差异性度量是不切实际的.

表 5 各种差异性度量与不同聚类集成算法集成准确度之间的相关性

	相关系数					
	MCLA <i>Ep</i> =0.9818	HGPA <i>Ep</i> =0.7527	CSPA <i>Ep</i> =0.9728	MajorityVote <i>Ep</i> =0.9827	Kmodes <i>Ep</i> =0.9411	ALBE <i>Ep</i> =0.9518
NMIBDM	0.4010	-0.1241	0.3084	0.3565	0.2215	0.3579
EBDM	0.4152	0.0425	0.3404	0.3571	0.2189	0.3160
CEBDM	0.3347	-0.1540	0.2425	0.2946	0.2059	0.3064
adRBDM_1	0.4767	-0.0524	0.3822	0.4314	0.2722	0.3999
adRBDM_2	0.2086	-0.1703	-0.0315	0.1815	0.0968	-0.0077
adRBDM_3	-0.4306	0.0564	-0.8716	-0.4640	-0.3185	-0.3255
adRBDM_4	-0.4478	0.0836	-0.8373	-0.4912	-0.3380	-0.3076
adRBDM_5	-0.4334	0.0613	-0.8693	-0.4686	-0.3202	-0.3242
DFBDM	0.3495	-0.0284	0.6296	0.3500	0.2789	0.4221
CFDBDM	0.1036	-0.0203	0.3571	0.0679	0.1168	0.1885
IRABDM	0.4306	-0.0171	0.6028	0.4143	0.3208	0.4725

3.6 分 析

从图 1 所示的实验结果中我们发现, 这 7 种差异性度量都没有出现我们所期望的与集成性能间的单调单值关系. 当差异性度量值增加, 也就是意味着它们所度量出来的差异性增加时, 集成准确度并不能保证增加; 当具有同样的集体平均成员准确度, 且具有同样的差异性度量值时, 用同一种集成方法产生了多个不同的集成准确度. 这也许是所有这些度量虽然在一定程度上反映了人们直觉上所认为的聚类成员不犯同样错误, 错误互补, 对于任何的样本总有一部分成员聚类对其进行正确聚类等等差异性所包含的意义, 但是由于聚类集体的差异性并没有一个严格和统一的定义, 所以对它的度量也就相对比较困难. 图 1 的实验结果也许让人失望, 但是从后面

的几个实验中, 我们还是看到了可喜的一面. 从表 1 可以看出, 在 3.2 节的实验环境下, 当集体的大小为 3 时我们提出的 DFBDM, CFDBDM 和 IRABDM 度量在成员聚类准确度 ≥ 0.7 以后, 它们与集成性能的相关系数比较大. 而从表 2 同样可以看出, 当成员聚类准确度为 0.65 时, DFBDM, CFDBDM 和 IRABDM 度量在各种集体大小情况下, 它们与集成性能的相关系数比较大, 在集体大小设为 20 时, 它们的相关系数达到了相当强的正相关关系. 但从表 4 我们观察到 DFBDM, CFDBDM 和 IRABDM 与其它大多数度量方法一样, 当数据中具有均匀大小的簇分布时, 它们与集成性能的相关性高, 而当数据中簇是非均匀分布时, 它们与集成性能的相关系数非常低, 几乎可以认为没有任何的指导意义. 在对不

同的集成方法与集体差异性度量关系的实验研究中我们发现,7 类差异性度量都在一定程度上反映了

差异性,所以与它们相关系数相对较大的集成方法的性能也相对更好.

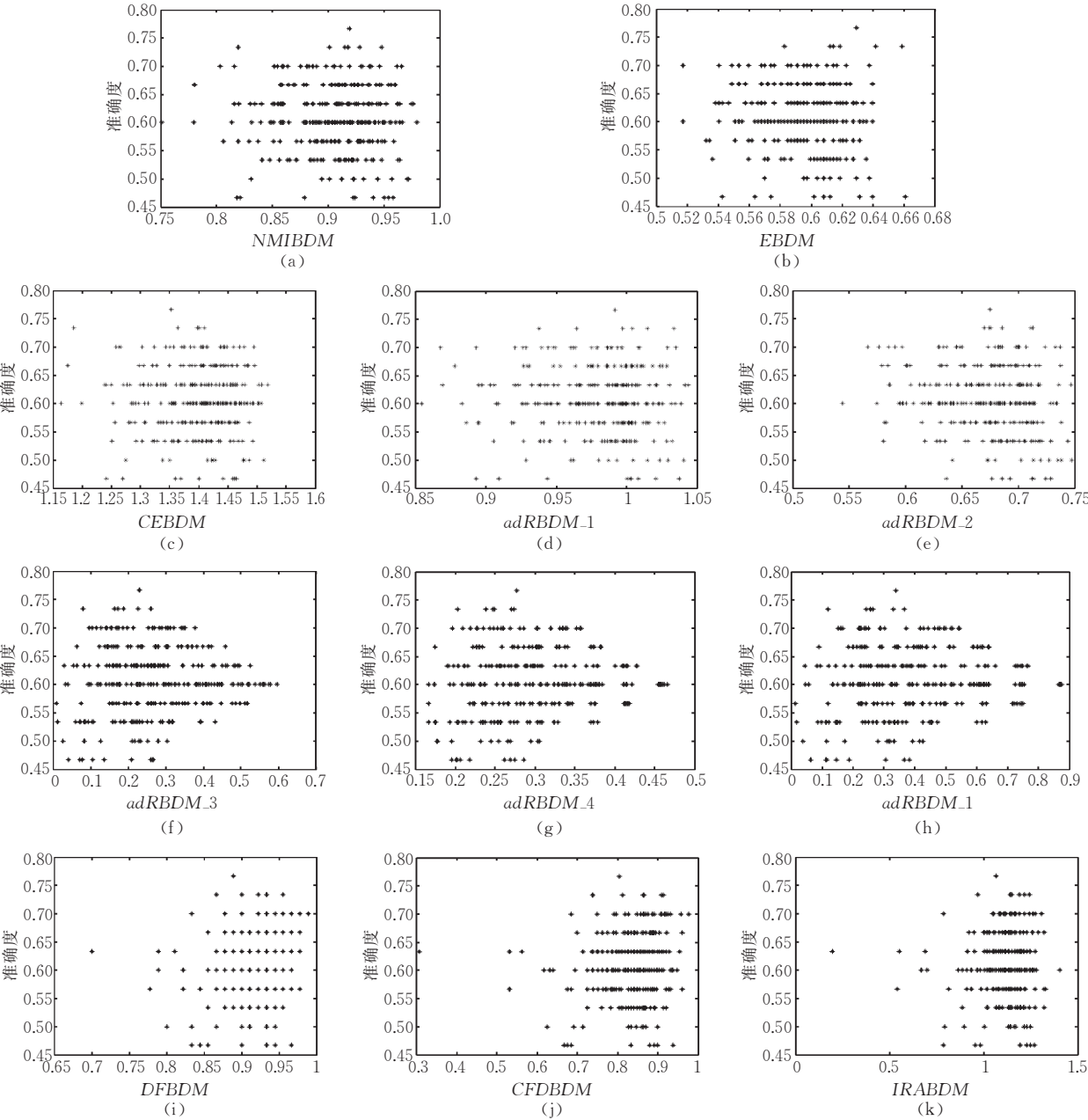


图 1 差异性度量值对 CSPA 集成准确度散点图(成员聚类的平均准确度为 0.6,集体大小为 3)

在对这 7 类差异性度量与聚类集成性能间相关系数的实验研究中,我们观察到一个不符合常理的现象.在文献[8]中基于 adjusted Rand index 定义的几个差异性度量在很多情况下呈现了与集成性能间相当强的负相关关系,特别是 *adRBDM_3*, *adRBDM_4* 和 *adRBDM_5*,这与它们定义的差异性概念好像有点相背离. *NMIBDM*, *EBDM* 和 *CEBDM* 在上述这些实验中与集成性能的相关性表现得很相似,但它们都低于集成性能与 *DFBDM*, *CFDBDM* 和 *IRABDM* 度量间的相关性.基于以上

这一点,当我们使用较强的基聚类算法且待聚类数据有基本均匀的簇分布时, *DFBDM*, *CFDBDM* 与 *IRABDM* 度量可以用于指导生成大小为 15~20 之间的聚类集体,或用来测量聚类集体的差异性,并进一步用来选择聚类成员以提高聚类集成的效果,但这一定要非常谨慎.

4 结 论

为了分析各种差异性度量与集成准确度之间的

关系, 也为了探讨用差异性度量来指导生成好的聚类集体的可行性, 我们进行了 5 个实验, 在这些实验中我们发现各种差异性度量与集成准确度之间并没有严格的单调正相关关系. 影响这种相关性的因素很多, 在不同的平均成员准确度情况下, 不同的集体大小情况下, 不同的数据分布情况下和不同的集成方法情况下, 这种相关性都不同. 实验中我们发现集体大小在 15~20 之间时, 各种差异性度量与集成准确度之间的相关性还是较好的. 我们还发现随着平均聚类成员准确度的增加, 各种差异性度量与集成准确度之间的相关度也在增加. 大多数的差异性度量均匀簇分布数据源情况下比非均匀簇分布数据源情况下与集成性能间的相关性更强. 同时我们的实验也证明了之所以在同样的聚类集体上不同的集成方法的性能不同, 原因是它们利用聚类集体差异性的能力不同, 反应在我们的实验中是各种集成方法的性能与差异性度量间的相关程度不同, 相关性更强的集成方法集成性能也更好. 所以如果我们试图使用某种差异性度量来指导一个聚类集体的生成, 我们不但要考虑待聚类数据的分布特点, 而且要考虑将使用的聚类集成方法.

参 考 文 献

- [1] Hansen L K, Salamon P. Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1990, 12(10): 993-1001
- [2] Tumer K, Ghosh J. Error correlation and error reduction in ensemble classifiers. *Connection Science*, 1996, 8(3/4): 385-403
- [3] Kuncheva L, Whitaker C. Measures of diversity in classifier ensembles and their relationship with ensemble accuracy. *Machine Learning*, 2003, 51(2): 181-207
- [4] Tang E K, Suganthan P N, Yao X. An analysis of diversity measures. *Machine Learning*, 2006, 65(1): 247-271
- [5] Fern X Z, Brodley C E. Random projection for high dimensional data clustering: A cluster ensemble approach//*Proceedings of the 20th International Conference on Machine Learning (ICML 2003)*. Washington DC, USA, 2003: 186-193
- [6] Tang Wei, Zhou Zhi-Hua. Bagging-based selective clusterer ensemble. *Journal of Software*, 2005, 16(4): 496-502 (in Chinese)
(唐伟, 周志华. 基于 bagging 的选择性聚类集成. *软件学报*, 2005, 16(4): 496-502)
- [7] Greene D, Tsymbal A, Bolshakova N et al. Ensemble clustering in medical diagnostics//*Proceedings of the 17th IEEE Symposium on Computer-Based Medical Systems CBMS'04*. Bethesda, MD, USA, 2004: 576-581
- [8] Hadjitodorov S T, Kuncheva L I, Todorova L P. Moderate diversity for better cluster ensembles. *Information Fusion*, 2006, 7(3): 264-275
- [9] MacKay D J C. *Information Theory, Inference and Learning Algorithms*. Cambridge: Cambridge University Press, 2003
- [10] Strehl A, Ghosh J. Cluster ensembles — A knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 2002, 3(3): 583-617
- [11] Hubert L, Arabie P. Comparing partitions. *Journal of Classification*, 1985, 2(1): 193-218
- [12] Giacinto G, Roli F. Design of effective neural network ensembles for image classification processes. *Image Vision and Computing Journal*, 2001, 19(9/10): 699-707
- [13] Partridge D, Krzanowski W J. Software diversity: Practical statistics for its measurement and exploitation. *Information and Software Technology*, 1997, 39(10): 707-717
- [14] Fleiss J. *Statistical methods for rates and proportions*. Hoboken, NJ, USA: John Wiley & Sons, 1981
- [15] Kuhn H W. The hungarian method for the assignment problem. *Naval Research Logistics*, 1955, 2(1): 83-97
- [16] Tsymbal A, Pechenizkiy M, Cunningham P. Diversity in ensemble feature selection. *The University of Dublin: Technical Report TCD-CS-2003-44*, 2003
- [17] Huang Z. Extensions to the k -means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery*, 1998, 2(3): 283-304
- [18] Luo H, Kong F, Li Y. Combining multiple clusterings via k -modes algorithm//*Proceedings of the Conference on Advanced Data Mining and Applications*. Xi'an, China, 2006: 308-315
- [19] Fred A L N, Jain A K. Data clustering using evidence accumulation//*Proceedings of the 16th International Conference on Pattern Recognition ICPR 2002*. Quebec, Canada. Los Alamitos, CA: IEEE Computer Society Press, 2002: 276-280



LUO Hui-Lan, born in 1974, Ph. D. candidate. Her current research interests include machine learning and data mining.

KONG Fan-Sheng, born in 1946, professor. His current research interests include artificial intelligence and knowledge discovery.

LI Yi-Xiao, born in 1982, Ph. D. candidate. His current research interests include data mining and image processing.

Background

Data clustering is a difficult inverse problem, and as such is ill-posed when prior information about the underlying data distributions is not well defined. Numerous clustering algorithms are capable of producing different partitions of the same data that capture various distinct aspects of the data. The exploratory nature of clustering tasks demands efficient methods that would benefit from combining the strengths of many individual clustering algorithms. This is the focus of research on clustering ensembles, seeking a combination of multiple partitions that provides improved overall clustering of the given data.

One challenging issue of the problem of combining mul-

tle clusterings is the choice of the generation method of the component partitions for the ensemble. Diversity among the member clusterings is deemed to be important when constructing a clustering ensemble. Numerous algorithms have been proposed to construct a good clustering ensemble by seeking the diversity among them. However, there is no generally accepted definition of diversity, and measuring the diversity explicitly is very difficult. While a number of ways are known to quantify diversity in ensembles of classifiers, little research has been done in clustering ensembles.

This paper focuses on the research of the diversity measures of clustering ensembles.

第三届中国高性能计算研讨会

会 议 通 知

(The 3rd Workshop on Chinese High Performance Computing)

<http://www.atip.org/node/94>

2007 年 11 月 11 日 美国 内华达州·里诺市

由亚洲科技资讯公司(ATIP)和超级计算大会组委会(SC07)共同主办、美国自然科学基金会(NSF)协办的第三届中国高性能计算研讨会将于 2007 年 11 月 11 日在 2007 超级计算大会(SC07)之前举行. 本次研讨会还得到了许多企业和机构的赞助, 如 Sun、Microsoft、SGI、Clearspeed、IDC、曙光、上海超算中心等. 中国高性能计算研讨会旨在加强中西方在该研究领域的理解, 促进合作, 已分别于 2003 年和 2004 年超级计算大会期间成功举办了两届. 欢迎中国从事高性能计算及相关研究和应用的专家、学者、用户、及厂商踊跃参加并投稿.

研讨会主要内容:

- A. 政府计划及项目
- B. 超算相关基础设施(超算中心, 网络等)
- C. 高性能计算相关研究(算法, 体系结构等)
- D. 高性能计算应用(科学计算, 工程计算, 企业应用等)
- E. 高性能计算产业(市场情况, 中外厂商)
- F. 小组座谈: 中国高性能计算现状
- G. 海报

有意参会者可提出自己愿意在哪一方面做口头报告或海报展示. 所有被本次研讨会采用的口头报告和海报都将被 ACM 数字图书馆收录.

对于使用高性能计算机或对高性能计算感兴趣的人来说, SC07(<http://sc07.supercomp.org>)是全球最大和最重要的活动. 今年估计将有超过 5000 人参会. 参会者不仅有机会与国际高性能计算领域的重要人物进行详细交流, 还可以参观来自世界各地的上百个展位(包括厂商, 工业用户及研究中心等). SC07 的学术活动包括与高性能计算有关的研究论文, 小组座谈和讨论会等.

详细情况, 请咨询 ATIP 北京代表处 陈道碧 女士

Tel: 010-62136752, 13601192280

E-mail: dchen@atip.org.cn 或 debbiechen@vip.sina.com.