层次聚类的簇集成方法研究

李 凯,王 兰
LI Kai,WANG Lan

河北大学 数学与计算机学院,河北省机器学习与计算智能实验室,河北 保定 071002
School of Mathematic and Computer,HeBei University,Key Lab in Machine Learning and Computational Intelligence of Hebei Province,Baoding,Hebei 071002,China
E-mail:likai@hbu.edu.cn

LI Kai,WANG Lan.Research on cluster ensembles methods based on hierarchical clustering.Computer Engineering and Applications,2010,46(27):120-123.

Abstract: Cluster ensembles method is considered as a robust and accurate alternative to single clustering runs.It mainly consists of both generation of individual member and fusion methods.In this paper,the cluster ensembles are studied where individual members are obtained based on *k*-means clustering algorithm and fusion method of hierarchical clustering is used. Three consensus functions,which are single linkage,complete linkage and average linkage,respectively,is studied and discussed in hierarchical clustering fusion.For evaluating performance of cluster ensembles,Adjusted Rand Index is considered. Experimental results show that performance of cluster ensembles with the average linkage is superior to one with single linkage and complete linkage.Moreover,the relationship between accuracy and ensemble size of the three fusion methods is also studied and discussed.

Key words: cluster ensembles;consensus functions;clustering;Adjusted Rand Index(ARI)

摘 要: 聚类集成比单个聚类方法具有更高的鲁棒性和精确性,它主要由两部分组成,即个体成员的产生和结果的融合。针对聚类集成,首先用*k*-means聚类算法得到个体成员,然后使用层次聚类中的单连接法、全连接法与平均连接法进行融合。为了评价聚类集成方法的性能,实验中使用了ARI(Adjusted Rand Index)。实验结果表明,平均连接法的聚类集成性能优于单连接法和全连接法。研究并讨论了融合方法的聚类正确率和集成规模的关系。

关键词: 聚类集成;融合函数;聚类;ARI
DOI:10.3778/j.issn.1002-8331.2010.27.033 **文章编号:** 1002-8331(2010)27-0120-04 **文献标识码:** A **中图分类号:** TP18

1 引言

聚类是数据挖掘中的一项重要技术,它能有效地分析数据并从中发现有用的信息。聚类将数据对象分为若干个类或簇,使得在同一个簇中的对象之间具有较高的相似度,而不同簇中的对象具有较大的差异。在知识发现和数据挖掘方法中,聚类已经被用来发现复杂数据库之间的模式和关系。最近,聚类集成方法作为聚类算法的一种改进开始被研究人员提出,以提高无监督分类方法的鲁棒性和稳定性。很多研究已经证明聚类集成对任意形状和规模的数据聚类时,其性能优于单一的聚类算法[1]。更重要的是,聚类集成使用户避免选用不恰当的聚类算法的风险。应当指出,聚类集成受益于多个聚类结果[2]。聚类集成的准确性和差异性有着特定的关系,差异性被引入以提高聚类集成的性能。另外,为了评估聚类集成的准确性,研究人员提出了ARI(Adjusted Rand Index)[3],该指数具有以下的性质:(1)如果两个划分独立,则其值是0;(2)相对于其他指数,它更易于挑选出正确的划分。

在本文中,研究了聚类集成算法中三种层次聚类方法作为融合函数的差异性,并用ARI对其准确性做了量化评价,然后在一些数据集上进行了实验并对实验结果进行了比较,最后对论文做了总结并指出未来的研究方向。

2 聚类集成

假设$P_1,P_2,\cdots,P_L$为数据集Z的L个划分,其中所使用的聚类算法可以是相同的但每次运行具有不同的参数,也可以使用不同的聚类算法。聚类集成的目的是找到一个最能体现Z的结构的划分P^*,P^*是通过对L个划分聚类集成得到的结

基金项目: 国家自然科学基金(the National Natural Science Foundation of China under Grant No.60773062);河北省自然科学基金(the Natural Science Foundation of Hebei Province of China under Grant No.F2009000236)。

作者简介: 李凯(1963-),男,教授,博士,主研方向:机器学习、计算智能、模式识别、数据挖掘、神经网络等;王兰(1983-),女,硕士研究生,主研方向:机器学习。

收稿日期: 2009-03-03 **修回日期:** 2009-05-11

果。研究人员已经证明，在分类器集成中，当构建的一组分类器具有一定的差异性时，分类器的集成效果较好[2]。按照类似的方法，可以将差异性引入到聚类方法之中，以此提高聚类集成的性能，也就是说，在聚类集成中，如果集成成员在数据集上的划分一致，那么聚类结果也不会有所改进。

聚类集成主要由两部分构成，即如何产生具有较大差异性的个体成员以及融合方法的选取。产生个体成员的方法主要包括：

(1)基于特征的聚类[1]：使用同一数据集的不同的特征子集。

(2)基于不同算法的聚类[4]：生成个体成员时使用不同的聚类方法。

(3)基于随机参数的聚类：例如使用k-means时，随机初始化参数K，或者产生L个随机的映射，将这些高维数据映射到低维空间中，然后对每个映射运行k-means[5]。

(4)基于随机抽样技术的聚类：例如，放回或无放回抽样技术[6]。

融合方法主要有：(1)直接方法。找两个划分之间簇标记的一致性，然后融合相同标记的簇。对同一个数据所属的簇标记出现的次数，采用多数投票法，出现次数最多的为此数据的簇标记[1]。基于特征的方法。将每个聚类的结果输出视为一个明确的特征。L个特征集可以被视为一个“中间特征空间”，然后在其之上运行某种聚类算法[7]。超曲线图方法。将L个划分组织成一个超曲线图，然后对超曲线图划分某种方法进行融合[1]。(2)成对方法。利用两个对象间的关系得到一个划分的协矩阵，采用某种聚类算法融合所有划分的协矩阵得到最后的划分[7]。

3　层次聚类方法

给出N个数据的数据集和一个$N\times N$的距离矩阵，层次聚类的基本算法如下：

步骤1　将每个数据点作为一个簇，簇间的距离(相似)等于相应数据点的距离。

步骤2　找到最相近(最相似)的两个簇，然后把它们合并。

步骤3　计算新产生的簇和原来每个簇之间的距离。

步骤4　重复步骤2和步骤3，直到所有的数据都聚到一个簇中。

可以看到，步骤3可以用不同的方法来实现，这些方法可以是单连接法、全连接法和平均连接法。

(1)单连接法：单连接法也就是最短距离法。两个类之间的距离用从两个类中抽取的每对样本的最小距离表示。作为距离度量，一旦最近的两个类的距离超过某个任意给定的阈值，算法就自动结束。

(2)全连接法：全连接法又称为最长距离法。全连接与单连接聚类方式相同，只是类与类之间的距离定义不是选取最小距离，而是找两类数据对象之间距离最大者。

(3)平均连接法：平均连接法的距离度量是选取两类数据对象之间平均距离。

4　聚类集成的评价

为了评价聚类集成的结果，使用了ARI(Adjusted Rand Index)指标评价聚类结果的正确性，ARI定义如下[3,8]：

$$t_1=\sum_{i=1}^{K_A}C_{N_i}^2,\ t_2=\sum_{j=1}^{K_B}C_{N_j}^2,\ t_3=\frac{2t_1t_2}{N(N-1)},$$

$$ARI(A,B)=\frac{\sum_{i=1}^{K_A}\sum_{j=1}^{K_B}C_{N_{ij}}^2-t_3}{(t_1+t_2)/2-t_3} \tag{1}$$

其中A和B是数据集Z的两个划分，分别有K_A和K_B个簇；N_{ij}表示在划分A的第i个簇中的数据同时也在划分B的第j个簇中数据的数量；N_i，N_j分别表示划分A中第i个簇与划分B的第j个簇中数据的数量。

由ARI定义知道，如果K_A，K_B以及每个簇中数据的数量一定，那么随机取两个划分A和B，则$ARI(A,B)$的值是0，这就表明，当$ARI(A,B)$的值接近0时，则划分A和B是独立的，反之亦然。

在实验研究中，使用$ARI(P_i,P^T)$来测量聚类结果的准确性，P_i是第i个聚类个体得到的划分，P^T是给定数据的正确划分。聚类集成的准确性可以用$ARI(P^*,P^T)$来计算，P^*是聚类集成的结果。

5　聚类集成算法的框架

给定一个含有N个数据的集合Z，设L和K分别表示集成的规模和簇的个数。为了使聚类集成中的个体之间存在差异性，可以通过改变聚类算法中的随机参数k，来生成L个个体成员，在实验研究中，k的取值范围是[2，K+10]。因此，在大多数情况下，对于每一个个体成员，k值是不同的。为了融合L个划分，对每个划分生成一个协矩阵，通过计算L个协矩阵的平均值，获得最后的矩阵$\boldsymbol{M}$，可以将其视为点和点之间的相似矩阵[2]。最后，在矩阵$\boldsymbol{M}$上应用层次聚类算法得到最终的划分。本文使用单连接法、全连接法以及平均连接法进行聚类集成，并通过$ARI(P^*,P^T)$分别计算三种融合函数集成结果的准确性。聚类集成算法的框架如下：

(1)初始化：对集成规模L和簇个数K赋初始值。

(2)生成个体成员：使用聚类算法通过随机选取k值生成L个划分。

(3)生成相应的划分矩阵：为每一个划分形成一个$N\times N$协矩阵，$\boldsymbol{M}^{(S)}=\{m_{ij}^{(s)}\}$，$s$=1，2，…，$L$，在划分$s$中，如果$Z_i$和$Z_j$在一个簇中，$m_{ij}^{(s)}=1$，否则$m_{ij}^{(s)}=0$。

(4)聚类结果的融合：利用不同的划分$M^{(s)}$，s=1，2，…，L生成协矩阵，$M=(M^{(1)}+M^{(2)}+\cdots+M^{(L)})/L$；应用层次聚类算法进行聚类。

实验中，研究了聚类集成算法的三种融合函数，包括单链接法，全连接法和平均连接法。为了评估聚类集成的性能，实验中使用了ARI指标，最后，对使用三种融合函数的实验结果进行了比较研究。

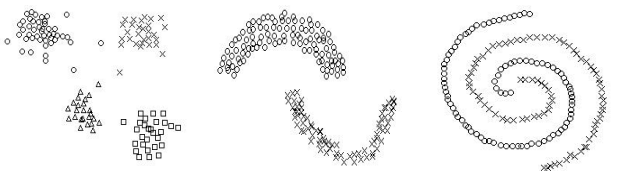
6　实验结果和分析

6.1　实验数据

为了评价聚类集成的性能，实验中选择了10个数据集，其中3个是人工数据，7个是UCI的数据，分别如图1与表1所示。图1中显示人工数据的簇的个数。

6.2　实验结果

当层次聚类达到K个簇时，就得到了聚类集成的结果，然后使用$ARI(P^*,P^T)$来计算结果的准确性。为了比较三种融合



(a)Four-gauss(4) (b)Half-rings(2) (c)Two spirals(2)

图1 人工生成的数据集

表1 数据集的特性

数据集名	类的个数	数据的个数	属性的个数
Glass	6	214	9
Hayes-roth	3	132	4
Iris	3	150	4
Lung-cancer	3	32	56
Soybean-small	4	47	35
Tae	3	151	5
Wine	3	178	13

函数的性能,分别选择了集成规模L为10、30、50、70、90与100,对每个数据点进行了10次实验,然后取这些值的平均值作为集成的最终结果。当集成规模是100时,三种方法比较的结果如表2所示。

表2 实验比较结果

数据集名	单连接法	全连接法	平均连接法
Four-gauss	0.933 9	0.958 5	0.970 2
Glass	0.359 1	0.510 1	0.512 2
Hayes-roth	0.477 2	0.491 3	0.531 6
Half-rings	0.193 9	0.396 8	1.000 0
Iris	0.680 8	0.724 2	0.768 0
Lung-cancer	0.443 3	0.603 7	0.682 7
Soybean-small	0.714 1	0.770 1	0.839 1
Tae	0.175 0	0.695 1	0.647 5
Two spirals	1.000 0	0.131 7	1.000 0
Wine	0.666 0	0.630 6	0.730 3

图1是修改后的三张图片。

实验结果表明在数据集Iris,Wine,Glass,Hayes-roth,Lung-cancer,Soybean-small,Half-rings,Two spirals和Four-gauss上,平均连接法的准确性优于另外两种方法,也就是说平均连接法聚类集成的性能优于单连接法和全连接法。

另外,与单个聚类算法 k-means(对每个数据集也进行了10次实验)相比,平均连接法也具有较好的性能。在人工数据集Half-rings,Two spirals和Four-gauss上,两种方法的实验结果如图2所示。

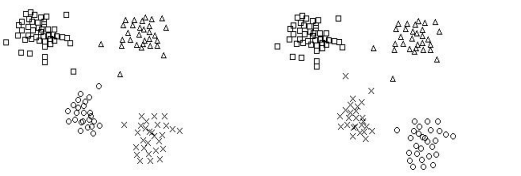
平均连接法在数据集Two spirals和Half-rings上的集成结果都是1,所以在这里没有将图显示出来。

图2是修改后的四张图片。

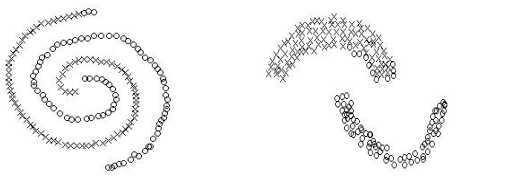
同样,在这些数据集上,又选择了集成规模为3 000进行了实验,结果如表3所示,它说明了平均连接法集成效果较好,具有较稳定的聚类集成性能。

为了更好地研究三种方法的正确性和集成规模之间的关系,实验中选择了集成规模分别为500、1 000、1 500、2 000和2 500,并针对这些数据集各自的集成规模,在每一个数值上进行了10次实验。实验结果如图3所示。

由图3可以发现,随着集成规模的增长,在数据集Iris,Wine,Glass,Lung-cancer,Soybean-small,Tae和Half-rings上,平均连



(a)平均连接法集成的结果 (b)k-means聚类的结果



(c)k-means聚类的结果 (d)k-means聚类的结果

图2 在数据集Four-gauss,Two spirals和Half-rings上平均连接法和k-means比较的结果

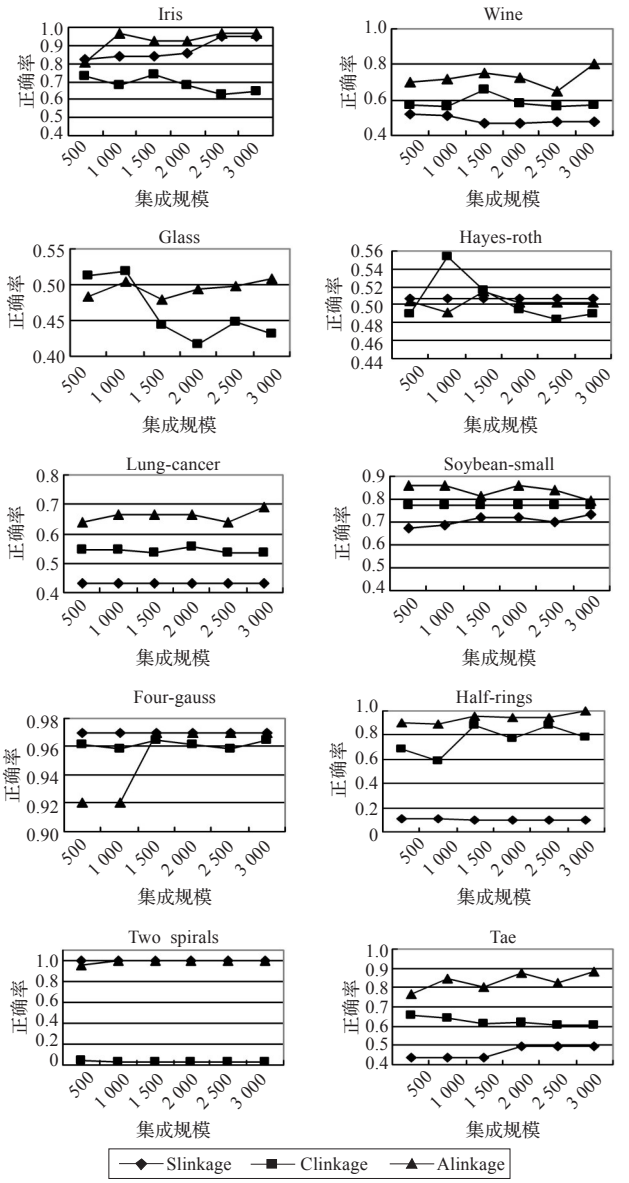


图3 三种融合方法的集成规模和正确率的关系

接法具有较高的准确性。三种方法的性能也趋于稳定。另外,三种方法在容易区分的数据集Four-gauss上都具有比较高的值,但单连接法的聚类集成的性能最稳定,且在数据集Two spirals也具有最佳的性能;相反,全连接法的性能是最差

的;对于这两种类型的数据集,平均连接法具有稳定且适中的性能。

7　结论

本文研究了聚类集成,将*k*-means作为基聚类算法生成个体成员,以及使用层次聚类方法作为融合方法对聚类结果进行融合;在实验中,分别在人工数据和UCI数据集上进行了实验研究并将三种融合函数的结果进行了比较;另外,为了评价聚类集成方法的性能,选用了ARI指标。最后,讨论并研究了三种方法的集成规模和正确率之间的关系。实验结果表明平均连接法相对于另外两种方法具有较好的聚类集成性能。在以后的研究中,需进一步研究聚类集成方法的性能和集成参数之间的关系,以及聚类集成的有效性。

参考文献:

[1] Strehl A,Ghosh J.Cluster ensembles—A knowledge reuse framework for combining multiple partitions[J].The Journal of Machine Learning Research,2003,3(3):583-617.

[2] Hadjitodorov S T,Kuncheva L I,Todorova L P.Moderate diversity for better cluster ensembles[J].Information Fusion,2005,7(3):264-275.

[3] Hubert L,Arabie P.Comparing partitions[J].Journal of Classification,1985,2(1):193-218.

[4] Hu X,Yoo I.Cluster ensemble and its applications in gene expression analysis[C]//Chen Y P P.Proc 2nd Asia-Pacific Bioinformatics Conference,Dunedin,New Zealand,2004:297-302.

[5] Topchy A,Jain A,Punch W.Combining multiple weak clusterings[C]//Proc Third IEEE International Conference on Data Mining,Melbourne Florida,2003:331-338.

[6] Minaei B,Topchy A,Punch W.Ensembles of partitions via data Resampling[C]//Proceedings of the International Conference on Information Technology on Coding and Computing,Las Vegas,NV,2004,2:188-192.

[7] Kuncheva L I.Evaluation of stability of *K*-means cluster ensembles with respect to random initialization[J].IEEE Transactions on Pattern Analysis and Machine Intelligence,2006,28(11):1798-1808.

[8] Rand W M.Objective criteria for the evaluation of clustering methods[J].Journal of the American Statistical Association,1971,66(336):846-850.

[9] Topchy A,Jain A K,Punch W.A mixture model for clustering ensembles[C]//Proceedings of SIAM Conference on Data Mining,2004:379-390.

(上接76页)

本CA研究的结果,即混沌型基本CA在45、30、90和150等规则下表现出混沌的非周期行为是产生随机数的原因[8]。

3.2　混合CA不同规则的对比实验

在混合CA模型中,进行了三次对比实验,采用元胞个数仍为20,迭代次数为1 280 000。在混合CA模型中,每运行2 048(M=8×256)步,对每个元胞特定的规则进行更新。

在混合CA对比实验1中,每个元胞按照随机获得规则进行更新和迭代演化,获得序列的熵值为3.728,大于基本CA平均熵值2.853,但远小于混沌型基本CA熵值最优值6.999。该实验反映出:混合CA输出随机序列的相关性,好于基本CA的平均表现,但远小于混沌型基本CA最优表现。

在混合CA对比实验2中,每个元胞基于混合CA与PSO融合的伪随机数产生算法,通过跟踪所有元胞中最优规则和自身最优规则,更新自身特定的规则,获得序列的熵值为4.025,大于混合CA对比实验1的熵值3.728。该实验反映出:混合CA和PSO融合的思想,有效地实现了每个元胞最佳CA规则的搜索,与对比实验1相比,一定程度上提高了生成随机序列的质量。

在混合CA对比实验3中,每个元胞在基本CA表现较优的45、30、90和150规则随机选取更新和迭代演化,获得序列的熵值为7.000。该实验反映出:即使混合CA的每个元胞都处于混沌型,但相比混沌型基本CA熵值最大值6.999,仍无法有效地提高生成随机序列的质量。

4　结束语

针对基本CA、混合CA的伪随机数发生器进行了深入的研究,通过对比实验发现,混合CA输出随机序列的相关性,尽管优于基本CA的平均表现,但远小于混沌型基本CA的最优表现。本文还运用PSO思想,针对混合CA的伪随机数发生器,提出了一种基于混合CA与PSO融合的伪随机数产生算法,其中元胞对应于PSO的粒子,其对应粒子在迭代规则空间中飞行。经过一定迭代次数后,通过计算每个元胞产生随机序列的熵值作为粒子的适应度函数值,有效地实现每个元胞最佳规则的搜索,一定程度上提高混合CA生成随机序列的质量。

基于对比实验结果与分析,观察到混沌型基本CA表现稳定并且较优。今后将引入小生境技术改进基于混合CA与PSO融合的伪随机数产生算法,将初始粒子群体分为若干个子粒子群体,且每个子粒子群体对应的元胞为混沌型基本CA;然后借助PSO基本思想实现每个子粒子群体之间的最佳规则搜索,构造出小生境技术下的最优CA-PSO耦合伪随机数发生器。

参考文献:

[1] Russel E C.Monte carlo and quasi-monte carlo methods[J].Acta Numberica,1998,7:1-49.

[2] 王千峰.元胞自动机与粒子群算法的融合在伪随机数发生器中的应用研究[D].上海:上海大学,2008.

[3] 杨自强,魏公毅.常见随机数发生器的缺陷及组合随机数发生器的理论与实践[J].数理统计与管理,2001,3(1):45-51.

[4] 杨自强,魏公毅.综述:产生伪随机数的若干新方法[J].数值计算与计算机应用,2001(3):201-216.

[5] 王许书,王新辉,夏宏.Montgomery方法及其在伪随机数发生器中的应用[J].计算机工程与应用,2001,37(11):52-53.

[6] Von Neumann.Theory of self-reproducing automata[M].Chicago:University of Illinois Press,1966.

[7] Codd E F.Cellular automata[M].New York:Academic Publishers,1968.

[8] Wolfram S.Random sequence generation by cellular automata[J].Adv Appl Mathematics,1986,7:123-169.

[9] Hortensius P D,McLeod R D.Parallel random number generation for VLSI systems using cellular automat[J].IEEE Trans Comput,1989,38(10):1466-1473.

[10] 朱保平,刘凤玉.基于耦合可控细胞自动机伪随机序列发生方法研究[J].计算机工程与应用,2006,42(29):69-70.

[11] Tomassini M.Generating high—quality Random Numbers in Parallel by Cellular Automata[J].Future Generation Computer Systems,1999,16(2):291-305.