聚类融合方法综述

阳琳斌，王文渊
(清华大学 自动化系，北京 100084)

摘　要：在分类算法和回归模型中，正广泛而且成功地使用着融合方法，该方法能克服分类、回归中的不稳定性，并给出较好的结果。在非监督机器学习领域，由于缺乏数据集的先验知识，所以分类和回归中的融合方法就不能直接用于聚类算法，这导致了该领域中对融合方法研究的起步较晚；近几年的研究和实验表明，聚类融合方法能很好地提高聚类算法的鲁棒性和稳定性。对近年来聚类融合的方法进行了综述，阐述了近年来对聚类融合方法进行研究的主要内容与特点，并讨论了聚类融合方法的研究方向。

关键词：聚类融合；数据重抽样；共识函数；差异度
中图法分类号：TP391　　**文献标识码**：A　　**文章编号**：1001-3695(2005)12-0008-03

Clustering Ensemble Approaches: An Overview

YANG Lin-yun, WANG Wen-yuan
(Dept. of Automation, Tsinghua University, Beijing 100084, China)

Abstract: Ensemble approaches are widely and successfully used in classification algorithms and regression models. It can offer better results for overcoming instabilities in classification algorithms and regression models. However, in unsupervised learning, the researches of ensemble approaches are concerned only in recent years. Because the prior information of data sets in unsupervised learning is unknown, the ensemble approaches of classification algorithms and regression models can't be utilized in the same way directly. Recent researches and experiments show that clustering ensemble approaches can enhance the robustness and stabilities of unsupervised learning greatly. This paper makes an overview of the clustering ensemble approaches in recent years. It illustrates the contents and characteristics of recent clustering ensemble approaches research and discusses the future directions of clustering ensemble study.

Key words: Clustering Ensemble; Data Resampling; Consensus Function; Diversity

如果不在事先对数据作出假设，那么从本质上说，就不会存在比随机猜测或其他任何算法更为优越的学习算法[1]。

聚类算法是一种非监督机器学习算法，与监督学习算法不同，我们事先对数据集的分布没有任何的了解。聚类算法的目的就是将集中的数据人为地划分成若干类，以揭示这些数据分布的真实情况。事实上，任何聚类算法都对数据集本身有一定的预先假设，根据文献[1]中的"No Free Lunch"理论，如果数据集本身的分布并不符合预先的假设，则算法的结果将毫无意义，甚至可以说该结果只是给出了一个错误的分布，或是给数据集强加了一个虚构的分布。因此，面对特定的应用问题，如何选择合适的聚类算法是聚类分析研究中的一个重要课题。

融合方法将不同算法或者同一算法下使用不同参数得到的结果进行合并，从而得到比单一算法更为优越的结果。在分类算法和回归模型中，融合方法的使用已经比较成熟。但在聚类分析领域，聚类融合方法的研究在近几年才开始出现。本文对近年来聚类融合的方法进行了综述，介绍了聚类融合的概念和一些术语，阐述了近年来聚类融合方法研究的主要内容与特点，并讨论了聚类融合方法的研究方向。

1　聚类融合方法概述

聚类融合(Clustering Ensemble/Clustering Combination)这个词是 A. Strehl 和 J. Ghosh 于 2002 年在文献[2]中提出的，但是在 2001 年 A. L. Fred 在文献[3]中已经进行了类似的研究。在文献[2]中，作者对聚类融合的定义是：将多个对一组对象进行划分的不同结果进行合并，而不使用对象原有的特征。它的具体表达如下：

假设有 n 个数据点 $X=\{x_1, x_2, x_3, \cdots, x_n\}$，对数据集 X 用 H 次聚类算法得到 H 个聚类结果，$H=\{\pi_1, \pi_2, \pi_3, \cdots, \pi_H\}$(以下称之为聚类成员)，其中 $\pi_k(k=1, 2, 3, \cdots, H)$ 为对第 k 次算法得到的聚类结果。设计一种共识函数 Γ，对这 H 个聚类成员的聚类结果进行合并，得到一个最终的聚类结果 π'，如图 1 所示。

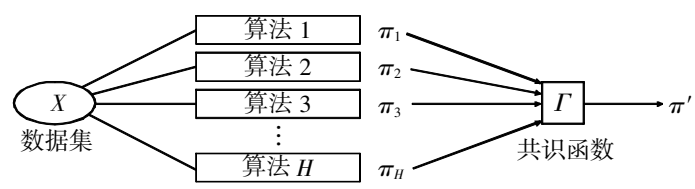

图 1　聚类融合示意图

聚类融合可以比单一聚类算法得到更好的结果，在文献[4]中，A. Topchy 总结出了以下几个方面：

(1) 鲁棒性。在各领域和数据集中的平均性能更为优越。

(2) 适用性。能找到单一聚类方法难以得到的聚类结果。

(3) 稳定性。噪声、孤立点和抽样方法等对聚类结果的影响较小。

(4) 并行性和可扩展性。能对数据子集进行并行聚类并

收稿日期：2005-03-31；修返日期：2005-04-30

进行合并; 能对分布式数据源或数据属性的聚类结果进行合并。

2　聚类融合方法的主要进展

在聚类融合中, 先要产生对数据集 X 的 H 个聚类成员, 然后对这 H 个聚类成员的聚类结果进行合并。因此, 近年来的研究主要集中在两个方面: ①如何产生有效的聚类成员, 不同的聚类成员之间需要有什么样的差异度; ②如何设计共识函数以便对聚类成员进行合并, 包括对聚类成员结果标志的匹配(在不同聚类结果中, 对实质相同的类给予同样的标志) 等。在文献[5] 中, B. Minaei-Bidgoli 等图示了当前聚类融合的研究方向(图 2)。

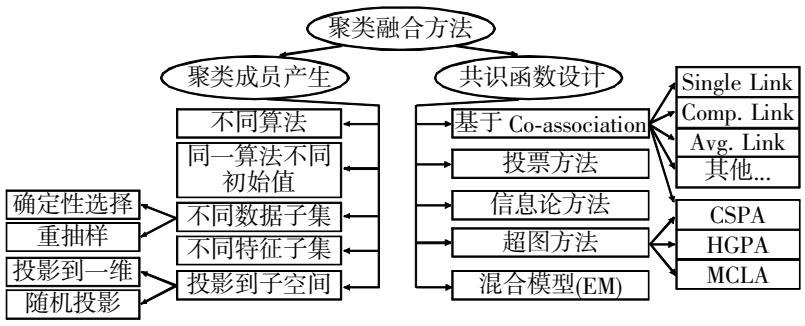

图 2　聚类融合研究方法图

从图 2 可知, 我们可以通过选择不同的算法、对一个算法选择不同的初始值、选择不同的对象子集(确定性选择或数据重抽样)、选择不同的特征子集投影到数据子空间(投影到一维空间或随机投影) 等来产生聚类成员。而在设计共识函数时, 有基于 Co-association 的分级聚合算法(SL, CL, AL 等)、投票算法、信息论算法、超图方法(CSPA, HGPA, MCLA) 和混合模型的 EM 算法等。

2.1　聚类成员产生方法

产生聚类成员最常用的方法是采用经典的 K-means 方法, 随机选择不同的初始点和聚类个数, 运行 H 次, 从而产生所需要的聚类成员。该方法最大的优点是算法的复杂度低, 运行方便等; 但是对于边界难以区分的数据、分布非球形的数据以及在处理高维数据时, 该方法的效果并不理想。

B. Minaei-Bidgoli 等人在文献[6] 中使用随机抽样的方法产生数据子集, 他们认为随机抽样产生的数据子集能从不同的层面反映整个数据集的结构, 从而能较好地揭示数据集的真实分布。他们对每个数据子集使用 K-means 算法生成聚类, 从而得到聚类成员。文献[7, 8] 中则采用 Bootstrap 抽样方法来产生聚类成员。B. Minaei-Bidgoli 等在文献[5] 中比较了随机抽样与 Bootstrap 方法产生聚类成员对聚类融合的影响。

A. Topchy 等人在文献[9] 中主要研究了由 "弱" 聚类组成的聚类成员对聚类融合的影响。"弱" 聚类是指仅比随机划分好一些的聚类, 文中用两种方法得到 "弱" 聚类: ①将多维数据随机投影到一维空间; ②用一组随机的超平面将数据空间切割, 然后用 K-means 算法分别对得到的子空间数据进行聚类, 组成聚类成员。文中选用了三种共识函数对聚类成员进行聚类融合。实验结果表明, 由 "弱" 聚类组成聚类成员实现的聚类融合算法, 其结果普遍好于单一的经典聚类算法。文中同时指出, 选用不同的共识函数对融合结果的影响很大, 对于某些数据集, 一些共识函数产生的结果并不理想, 甚至有不收敛的情况。

X. Z. Fern 等人在文献[10] 中使用随机投影法将高维数据投影到低维空间, 通过多次投影得到若干个数据子集, 然后用 EM 聚类算法对每次的投影子集聚类, 从而得到聚类成员。他们还在文献[11] 中使用了随机投影、PCA 与随机抽样、PCA 与随机投影等方法综合产生聚类成员。

A. Topchy 等人在文献[12] 中提出自适应的数据成员产生方法, 并对数据重抽样提出了新观点。他们认为数据集中的数据点对聚类结果的贡献是不相等的, 新的子样本集应集中在数据空间中有问题的区域, 即类与类之间的决策边界区域, 该区域的数据点应被偏重考虑。在文献[12] 中, 聚类成员是顺序产生的, 每次对抽样的数据子集用 K-means 算法聚类。作者用一个参数来衡量数据点的连贯性, 以决定下一次抽样中该点被抽取的概率, 第一次抽样时所有点的抽取概率是相等的。该参数被定义为数据点最多地被分到某一类中的次数除以当前聚类算法运行过的总次数, 并以此来调整该数据点下一次被抽取的概率。该参数取值在 0 到 1 之间, 值越小表明该点越处于问题区域, 下次被抽取的概率也越大。这里存在标志匹配的问题, 文中使用第一次聚类算法的聚类结果作为参考标志, 并使用 Hungarian 算法[13] 进行标志匹配。实验结果表明使用自适应的数据抽样方法能较好地提高聚类融合的结果。

L. I. Kuncheva 等人在文献[14] 中就聚类成员的差异度对聚类融合的影响进行了详尽的讨论, 通过实验作者发现聚类成员间较大的差异度将能提高聚类融合的质量。聚类成员之间的差异度可以通过 Rand Index, Jaccard Index, Adjusted Rand Index, Mutual Information 等来衡量。S. T. Hadjitodorov 等[15] 进一步研究发现, 聚类成员间的差异度与聚类融合质量之间的关系是非单调的, 当差异度过大时, 聚类融合的质量反而下降。文中指出, 适中的差异度将能得到较好的聚类融合结果, 该文还给出了寻找适中差异度的方法。

2.2　共识函数设计

在 A. Strehl 和 J. Ghosh 提出聚类融合这个概念之前, 已经有一些文献对不同聚类结果的合并进行了研究。A. L. Fred 在文献[3] 中提出了 Co-association 矩阵, 用于衡量数据点之间的相似度, 其中第 i 个数据点与第 j 个数据点之间的相似度为

$$A_{ij} = \frac{i \text{ 与 } j \text{ 属于同一聚类的次数}}{\text{聚类算法总次数 } H}$$

文中提出了 Voting-K-means 的方法, 随机产生初始点进行 K-means 算法, 运行 H 次得到聚类成员, 然后计算出聚类成员的 Co-association 矩阵。该方法选取 0.5 为阈值, Co-association 矩阵中大于 0.5 的点对即认为属于最终聚类结果中的同一类。

A. Strehl 和 J. Ghosh 在文献[2] 中提出了三个基于超图的方法: CSPA, HGPA 和 MCLA。①CSPA(Cluster-based Similarity Partitioning Algorithm) 首先得到聚类成员的 Co-association 矩阵, 由 Co-association 矩阵生成图, 其中顶点为数据点, 边的权重为数据点之间的 Co-association 值; 然后用基于图论的聚类算法 METIS 算法进行聚类, 得到最终的聚类结果。② HGPA (Hypergraph Partitioning Algorithm) 生成一个超图, 其中顶点为数据点, 每一个聚类成员中的每个类均为一条封闭的边, 包含了属于该类的所有顶点, 所有的顶点和边的权重都设置为相同的值; 然后用基于超图的聚类算法 HMETIS 算法进行聚类, 得到最终的聚类结果。③MCLA(Meta-Clustering Algorithm) 是一

种对聚类成员再聚类的超图算法,它对超图边进行断裂操作,再由断裂操作来决定每个数据点最终属于的类。

A. Fred 等在文献[16]中提出了 EA(Evidence Accumulation)方法。文中首先得到聚类成员的 Co-association 矩阵,然后用基于 MST(Minimum Spanning Tree)的分级聚类算法(Single-Link, Complete-Link, Average-Link)得到最终的聚类结果。A. Fred 等在文献[17]中改进了 EA 方法:①在使用 K-means 算法生成聚类成员时,除了随机选择初始点外,还在一个预先设定的 k 值范围内随机选择 k 值;②在用基于 MST 的分级聚类算法时不是使用预先设定的阈值 t 进行合并,而是以 0.05 为间隔在 $[0.05, 0.95]$ 的范围内计算出所有 t 下的聚类结果,选取最终聚类个数保持不变最长范围的 t 作为阈值。

H. Ayad 等在文献[18]中借用了近邻法的思想,扩展了 CSPA 的图方法,提出了 WSnnG(Weighed Shared nearest neighbors Graph)方法。文中首先得到聚类成员的 Co-association 矩阵,与 CSPA 不同,它并不直接根据 Co-association 矩阵生成图,而是将 Co-association 矩阵稀疏化,只保留矩阵中值大于某个阈值(文中选为 0.5)的项,其他的项设为 0。这一步骤即是只保留了每个数据点与其最近邻的那些点的联系,然后以稀疏后的矩阵生成图,同时根据各点最近邻点的数量计算顶点和边的权重,再用基于图论的聚类算法得到最终的聚类结果。H. Ayad 等在文献[19]中改进了 WSnnG 方法,文中只使用每个数据点的 k 个最近邻的邻点来生成边,从而得到一个精简的图而减少了计算复杂度,实验表明聚类融合结果并不受太大的影响。文中还调整了计算边的权重的方法,直接使用 Co-association 矩阵中相应的值,更能体现点与点之间联系的程度。

X. Z. Fern 等在文献[20]中使用双元图进行聚类融合,提出了 HBGF(Hybrid Bipartite Graph Formulation)方法。文中生成一个双元图,顶点不仅由数据集中的数据点构成,还包括了聚类成员的每一个聚类。图中数据点与数据点之间及聚类点与聚类点之间不相连,若某一个聚类包含了某个数据点,则该聚类点与数据点之间生成一条边,权重设为 1。对生成的图用 METIS 算法进行基于图论的聚类,得到最终的聚类结果。

A. Topchy 等在文献[4]中用一个多项式分布的混合模型构建共识函数,然后使用 EM 算法求解最终的聚类。这种方法完全避免了标志匹配的问题,而且能够处理丢失的数据,计算复杂度比基于 Co-association 矩阵的算法低。H. Ayad 等则在文献[21]中利用信息论的度量建立一个概率模型,提出了 JS-DCC(Jensen-Shannon Divergence based Clustering Combination)方法。文中从 Co-association 矩阵中提取出一组联合分布模型,通过计算数据点与模型的 Jensen-Shannon 值[22]而决定数据点最终所属的聚类。

P. E. Jouve 等在文献[23]中把聚类成员的聚类结果看成数据点的元特征,从而把聚类成员的标志映射成数据点的 H 维的属性,第 i 维的值标志该数据点在第 i 个聚类成员中的聚类结果;然后使用 Kerouac 方法[24]对数据集映射后的属性空间进行聚类,得到最终的聚类结果。该算法有较好的可扩展性,且不需要预先设定聚类个数。

3 未来研究方向

现有的聚类融合算法远未达到成熟的程度,未来研究方向将在以下方面展开:

(1)若干关键参数的确定。确定聚类的个数是聚类分析研究的热点之一,在聚类融合中,如何确定每个聚类成员的聚类个数、最终聚类的个数以及它们之间的关系将是一个值得探讨的问题。另外,在聚类成员质量差别很大时,聚类融合要给不同的聚类成员赋予不同的权重才可能得到较好的融合效果,权重的设计也是一个需要解决的问题。

(2)在聚类成员产生方面,需要进一步研究不同聚类算法之间的高效融合以及选择不同聚类算法的准则;而聚类成员相互不独立时的聚类融合,高维数据、海量数据的聚类融合,聚类之间存在混叠数据时的聚类融合等都对聚类融合提出了挑战。同时,差异度的量化分析仍需要更深入的研究。

(3)在共识函数设计方面,需要进一步研究“软”聚类(模糊聚类)的聚类融合及“硬”聚类与“软”聚类的聚类融合。而聚类融合的可扩展性算法、增量式算法的研究也是急需解决的问题。

(4)目前,聚类融合方法已经应用在医学诊断、基因表达分析、非数值型数据聚类等方面。研究聚类融合方法更广泛的实际应用以及现有机器学习方法在聚类融合中的应用都将有重要的意义。

参考文献:

[1] O Duda, P E Hart, D G Stork. Pattern Classification (2nd Edition) [M]. New York: Wiley, 2001. 454-458.

[2] A Strehl, J Ghosh. Cluster Ensembles: A Knowledge Reuse Framework for Combining Multiple Partitions [J]. Journal of Machine Learning Research, 2003, 3(3): 583-617.

[3] A L Fred. Finding Consistent Clusters in Data Partitions [C]. Proceedings of the 2nd International Workshop on Multiple Classifier Systems, Volume 2096 of Lecture Notes in Computer Science, Springer, 2001. 309-318.

[4] A Topchy, A K Jain, W Punch. A Mixture Model for Clustering Ensembles [C]. Proceedings of the 4th SIAM International Conference on Data Mining, 2004. 379-390.

[5] B Minaei-Bidgoli, A Topchy, W F Punch. A Comparison of Resampling Methods for Clustering Ensembles [C]. Intl. Conf. on Machine Learning, Models, Technologies and Applications (MLMTA 2004), 2004. 939-945.

[6] B Minaei-Bidgoli, A Topch, W F Punch. Ensembles of Partitions via Data Resampling [C]. Proceedings International Conference on Information Technology, Coding and Computing (ITCC 2004), Volume 2, 2004. 188-192.

[7] S Dudoit, J Fridlyand. Bagging to Improve the Accuracy of a Clustering Procedure [J]. Bioinformatics, 2003, 19(9): 1090-1099.

[8] B Fischer, J M Buhmann. Path-based Clustering for Grouping of Smooth Curves and Texture Segmentation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2003, 25(4): 513-518.

[9] A Topchy, A K Jain, W F Punch. Combining Multiple Weak Clusterings [C]. Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM 03), 2003. 331-338.

[10] X Z Fern, C E Brodley. Random Projection for High Dimensional Data Clustering: A Cluster Ensemble Approach [C]. Proceedings of the 20th International Conference on Machine Learning, 2003. 186-193.

[11] X Z Fern, C E Brodley. Cluster Ensembles for High Dimensional Clustering: An Empirical Study [EB/OL]. http://mow.ecn.purdue.edu/~xz/clustensem.pdf, 2004.

(下转第 14 页)

6 结束语

入侵检测作为一种积极主动的安全防护技术,处于快速发展之中,其研究和开发越来越受到学界和业界的重视。本文从 IDS 体系结构、入侵检测方法和 IDS 的测评研究三个方面对入侵检测进行了评述,分析了入侵检测的发展现状,讨论了当前入侵检测中存在的问题并指出今后的发展趋势。本文的工作为今后深入 IDS 的研究提供了必要的技术依据和发展方向。

参考文献:

[1] Lee, S J Stolfo, P K Chan. Real-time Data Mining-based Intrusion Detection[C]. Proceedings of the DARPA Information Survivability Conference and Exposition II (DISCEXII), Anaheim, CA: IEEE Computer Society, 2001. 85-100.

[2] P K Harmer, P D Williams, G H Gunsch. An Artificial Immune System Architecture for Computer Security Applications[J]. IEEE Transactions on Evolutionary Computation, 2002, 6(3): 252-280.

[3] Eugene H Spafford, Diego Zamboni. Intrusion Detection Using Autonomous Agents[J]. Computer Networks, 2000, 34: 547-570.

[4] Richard Lippmann, Joshua W Haines, David J Fried. Analysis and Results of the 1999 DARPA Off-Line Intrusion Detection Evaluation [C]. The 3rd International Symposium on Recent Advances in Intrusion Detection, Berlin Heidelberg: Springer, 2000. 162-182.

[5] 连一峰,戴英侠,胡艳,等.分布式入侵检测方法研究[J].计算机研究与发展,2003,40(8):1195-1202.

[6] 苏璞睿,李德全,冯登国.基于基因规划的主机异常检测方法[J].软件学报,2003,14(6):1120-1126.

[7] 高艳,管晓宏,孙国基.基于实时击键序列的主机入侵检测[J].计算机学报,2004,27(3):396-401.

[8] 冯力,管晓宏,郭三刚,等.采用规划识别理论预测系统调用序列中的入侵企图[J].计算机学报,2004,27(8):1083-1091.

[9] 徐明,陈纯,应晶.基于系统调用分类的异常检测[J].软件学报,2004,15(3):391-403.

[10] S Staniford Chen, B Tung. The Common Intrusion Detection Framework Architecture [EB/OL]. http://www.isi.edu/gost/cidf/drafts/2001/architecture.txt, 2003.

[11] J S Sherif, T G Dearmond. Intrusion Detection: Systems and Models [C]. Proceedings of the 11th IEEE International Workshops, Pennsylvania: IEEE Computer Society, 2002. 115-133.

[12] 卿斯汉,蒋建春,马恒太,等. 入侵检测技术研究综述[J].通信学报,2004,25(7):19-29.

[13] Carla Marceau. Characterizing the Behavior of a Program Using Multiple-length N-grams[C]. Proceedings of the New Security Paradigms Workshop, Cork, Ireland: USENIX, 2000. 68-74.

[14] Juan M Estevez-Tapiador, Pedro Garcia-Teodoro, *et al*. Anomaly Detection Methods in Wired Networks: A Survey and Taxonomy [J]. Computer Communications, 2004, 27(16): 1569-1584.

[15] Ke Wang, Salvatore J Stolfo. Anomalous Payload-based Network Intrusion Detection[C]. The 7th International Symposium on Recent Advances in Intrusion Detection, Berlin Heidelberg: Springer, 2004. 203-222.

[16] Mizuki Oka, Yoshihiro Oyama, Hirotake Abe. Anomaly Detection Using Layered Networks Based on Eigen Co-occurrence Matrix[C]. The 7th International Symposium on Recent Advances in Intrusion Detection, Berlin Heidelberg: Springer, 2004. 223-237.

[17] P A Porras, A Valdes. Live Traffic Analysis of TCP/IP Gateways [C]. Proceeding of the 1998 ISOC Symposium on Networks and Distributed Systems Security, San Diego, CA: Internet Society, 1998. 132-140.

[18] N Athanasiades, R Abler, J Levine, *et al*. Intrusion Detection Testing and Benchmarking Methodologies[C]. Proceedings of the 1st IEEE International Workshop on Information Assurance, Darmstadt, Germany: IEEE Computer Society, 2003. 63-72.

[19] Peter Mell, Vincent Hu, Richard Lippmann, *et al*. An Overview of Issues in Testing Intrusion Detection Systems[R]. NIST Interagency Report (NISTIR), 2003.

[20] Herve Debar, Benjamin Morin. Evaluation of the Diagnostic Capabilities of Commercial Intrusion Detection Systems[C]. The 5th International Symposium on Recent Advances in Intrusion Detection, Berlin Heidelberg: Springer, 2002. 177-198.

作者简介:

吴庆涛(1975-),男,江西会昌人,博士研究生,主要研究领域为网络异常行为检测;邵志清(1966-),男,江苏常熟人,教授,博士生导师,主要研究领域为软件开发与验证方法。

(上接第 10 页)

[12] Topchy, B Minaei-Bidgol, A K Jain, *et al*. Adaptive Clustering Ensembles [C]. Proceedings of the 17th International Conference on Pattern Recognition(ICPR 2004), Volume 1, 2004. 272-275.

[13] H Christos, Papadimitriou, Kenneth Steiglitz. Combinatorial Optimization: Algorithms and Complexity [M]. Prentice-Hall, 1982.

[14] L I Kuncheva, S T Hadjitodorov. Using Diversity in Cluster Ensembles [C]. Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, 2004. 1214-1219.

[15] S T Hadjitodorov, L I Kuncheva, L P Todorova. Moderate Diversity for Better Cluster Ensembles [J]. Information Fusion, 2005, (1).

[16] A Fred, A K Jain. Data Clustering Using Evidence Accumulation [C]. Proceedings of the 16th International Conference on Pattern Recognition(ICPR 2002), Volume 4, 2002. 276-280.

[17] A Fred, A K Jain. Evidence Accumulation Clustering Based on the K-means Algorithm [C]. Proceedings of the International Workshops on Structural and Syntactic Pattern Recognition(SSPR 2002), 2002. 442-451.

[18] H Ayad, M Kamel. Finding Natural Clusters Using Multi-Clusterer Combiner Based on Shared Nearest Neighbors [C]. Proceedings of the 4th International Workshop on Multiple Classifier Systems (MCS 03), Volume 2709 of Lecture Notes in Computer Science, Springer, 2003. 166-175.

[19] H Ayad, M Kamel. Refined Shared Nearest Neighbors Graph for Combining Multiple Data Clusterings, Advances in Intelligent Data Analysis [C]. Proceedings of the 5th International Symposium on Intelligent Data Analysis(IDA 2003), Volume 2810 of Lecture Notes in Computer Science, Springer, 2003. 307-318.

[20] X Z Fern, C E Brodley. Solving Cluster Ensemble Problems by Bipartite Graph Partitioning [C]. The 21st International Conference on Machine Learning, 2004.

[21] H Ayad, O A Basir, M Kamel. A Probabilistic Model Using Information Theoretic Measures for Cluster Ensembles [C]. Proceedings of the 5th International Workshop on Multiple Classifier Systems, Volume 3077 of Lecture Notes in Computer Science, Springer, 2004. 144-153.

[22] J Lin. Divergence Measures Based on the Shannon Entropy [J]. IEEE Transactionson Information Theory, 1995, 37(1): 145-151.

[23] P E Jouve, N Nicoloyannis. A New Method for Combining Partitions, Applications for Distributed Clustering [C]. International Workshop on Paralell and Distributed Machine Learning and Data Mining(ECML/PKDD03), 2003.

[24] P E Jouve, N Nicoloyannis. KEROUAC: An Algorithm for Clustering Categorical Data Sets with Practical Advantages [C]. Proceedings of DMAK, International Workshop on Data Mining for Actionable Knowledge, 2003.

作者简介:

阳琳斌,男,博士研究生,研究方向为数据挖掘算法与应用;王文渊,博士生导师,研究方向为信息处理、图像处理、模式识别和网络通信等领域的理论与应用研究。