聚类集成方法研究

杨草原 刘大有 杨 博 池淑珍 金 弟
(吉林大学计算机科学与技术学院 长春 130012)
(吉林大学符号计算与知识工程教育部重点实验室 长春 130012)

摘 要 聚类集成通过对原始数据集的多个聚类结果进行学习和集成,得到一个能较好地反映数据集内在结构的数据划分。聚类集成能够较好地检测和处理孤立点,提高聚类结果质量。综述了聚类集成的相关知识,介绍了聚类集成的相关概念和优点;根据使用的聚类算法介绍了3种产生聚类成员方法,分析了各自的优缺点及适用条件;介绍了目前已有的一致性函数,阐述了其基本原理,并指出了其局限;最后讨论了未来的研究方向。

关键词 聚类集成,聚类成员,一致性函数,聚类算法
中图法分类号 TP391 **文献标识码** A

Research on Cluster Aggregation Approaches

YANG Cao-yuan LIU Da-you YANG Bo CHI Shu-zhen JIN Di
(College of Computer Science and Technology, Jilin University, Changchun 130012, China)
(Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun 130012, China)

Abstract Clustering aggregation can offer a partition that could better reflect the inherent structure of the data set by studying and integrating many clustering results of the original data set. Clustering aggregation could detect and deal with the isolated points preferably, which improves the quality of clustering. This paper made an overview of the relevant knowledge of the clustering aggregation, presented the concepts and advantages of clustering aggregation. It presented three approaches to get clustering members according to the uesd cluster algorithms; analysed their respective advantages, disadvantages and application conditions; presented the existing consensus functions; explained the basic principles and pointed out their limitations. Finally, it discussed the future research directions.

Keywords Clustering aggregation, Clustering member, Consensus function, Clustering algorithms

1 引言

聚类分析通过计算数据对象间的相似度把数据集划分为若干簇,使同一个簇的对象具有较高的相似度,而不同簇的对象差异较大[1]。作为一种不同于分类的非监督学习方法,聚类分析在数据挖掘[2]、信息检索[3]、图像分割[4]、模式识别和机器学习等很多领域都具有广泛的应用。

传统的聚类方法主要包括划分方法(partitional method)、层次方法(hierarchical method)、基于密度的方法(density-based method)、基于网格的方法(grid-based method)和基于模型的方法(model-based method)[1],其代表算法如图1所示。

虽然传统的聚类算法[5]有很多,但存在以下局限性[6]:1)聚类结果很大程度上取决于参数及初始化;2)大多数聚类算法难于判断数据集的真实类簇个数;3)对同一数据集不同的聚类算法可能产生不同的结果。没有一种聚类算法能准确揭示各种数据集所呈现出来的多种多样的簇结构[7]。现实世界的多维数据集可能具有各种形状或结构,因此使用单一聚类算法、识别其类簇结构是一件困难的事情[8]。

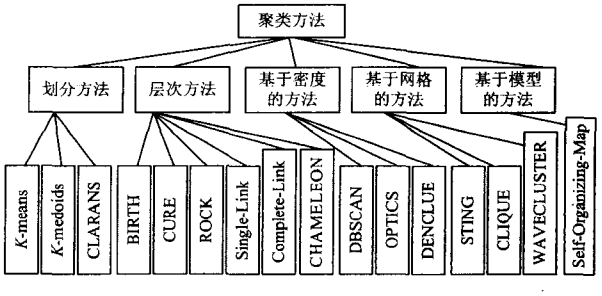

图1 聚类方法的代表算法

集成学习通过集成多个不同的学习器来解决同一个问题,提高系统的学习能力,被广泛用于机器学习、神经网络、统计学等领域[9]。聚类集成利用集成学习技术,通过学习合并数据集的多个聚类结果,得到一个新的聚类结果。

到稿日期:2010-03-31 返修日期:2010-06-29 本文受国家自然科学基金项目(60773099,60873149,60973088),国家863高技术研究发展计划项目(2006AA10Z245,2006AA10A309),中央高校基本科研业务费专项资金资助。

杨草原(1986—),女,硕士生,主要研究方向为数据挖掘,E-mail:yangcaoyuan2009@163.com;刘大有(1942—),男,教授,博士生导师,主要研究方向为知识工程与专家系统、数据挖掘等;杨 博(1974—),男,博士,教授,CCF会员,主要研究方向为Agent系统、数据挖掘与复杂网络分析;池淑珍(1986—),女,硕士生,主要研究方向为模式识别、图像处理;金 弟(1982—),男,博士生,主要研究方向为数据挖掘。

本文主要对近年来的聚类集成技术进行综述。首先介绍聚类集成的概念和优点；其次介绍聚类集成要研究的问题：产生聚类成员和设计一致性函数；介绍了3种产生聚类成员的方法，分析了各种方法的优缺点以及选取方法；介绍了目前已有的一致性函数的代表算法，指出其局限；随后介绍其它一些具有扩展性和对混合数据集进行聚类的聚类集成方法；最后针对聚类集成当前的研究现状，讨论了未来的研究方向。

2 聚类集成的概念

2002年，Strehl等人提出“聚类集成”(cluster ensembles)的概念，并给出了定义。聚类集成概念系指关于一个对象集合的多个划分(partitioning)组合成为一个统一聚类结果的方法，该方法不使用确定这些划分的特征和聚类算法[10]。2007年，Gionis等人也给出该问题的一种描述：给定一个聚类结果集合，聚类集成(clustering aggregation)的目标就是要寻找一个聚类，使其相对于所有的输入聚类结果来说，尽可能多地符合(或一致)[11]。由此可见，聚类集成是利用(经过选择的)多个聚类结果找到一个新的数据(或对象)划分，这个划分在最大程度上共享了所有输入的聚类结果对数据(或对象)集的聚类信息。

聚类集成过程为：假设数据集 X 有 n 个数据对象，$X=\{x_1,x_2,\cdots,x_n\}$，首先对数据集 X 使用 N 次聚类算法，得到 N 个聚类，$P=\{P_1,P_2,\cdots,P_N\}$(以下称之为聚类成员)，其中 P_i $(i=1,2,3,\cdots,N)$ 为第 i 次聚类算法得到的聚类结果。然后一致性函数 T 对 P 中的聚类结果进行集成得到一个新的数据划分 P'，如图2所示。

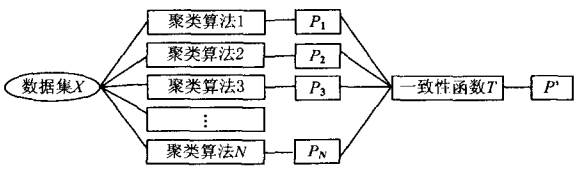

图2 聚类集成过程示意图

与单一聚类算法相比，聚类集成有以下优点[10-12]：

1)提高了聚类结果的质量和聚类的健壮性。不同的聚类结果(或划分)，从不同方面反映了数据集合的结构，聚类集成反映了多个聚类算法的综合特性。

2)聚类结果的集成重用和充分利用。

3)能划分具有分类属性的数据集。根据数据集的每个属性的不同值划分数据，每个属性得到一个聚类结果，对得到的划分进行集成，得到最终结果。

4)能检测和处理孤立点、噪音。若一个数据对象不属于任何一个簇或聚类成员，则无法对其得到一致划分，聚类集成将此数据对象划为单独的一个簇，不会影响其它数据对象及聚类结果。

5)能并行处理数据集，也能处理分布式数据源。对大规模的数据集，每个主机只对数据集的一部分并行地进行聚类，然后再集成，提高了时间效率；对分布式数据源，先在本地聚类，然后主机访问各个聚类结果进行集成，避免将整个数据集传输到主机。

3 聚类集成方法分析

由聚类集成过程可知，对一个数据集进行聚类集成，首先要产生多个聚类结果，然后对这些聚类进行集成。因此聚类集成主要研究两方面的问题：1)如何生成多个不同的聚类结果；2)如何设计有效的一致性函数对聚类结果进行集成，得到一个能反映数据集结构的数据划分。

3.1 生成聚类成员

通常在集成学习中，集体的差异度被认为是影响集成结果的关键因素之一[13]。聚类集成的第一步是产生多个具有差异度的聚类结果，差异从不同的方面反映数据集的结构，有利于集成。这一阶段是对数据集或其子集反复运行聚类算法，有以下几种方法：

(1)使用一个聚类算法，每次运行都设置不同的参数和随机初始化；

(2)使用不同的聚类算法，如K-means，SL(Single-Link)，CL(Complete-Link)，AL(Average-Link)等产生多个不同的聚类；

(3)对数据集的子集进行聚类，子集可通过采样如bagging，bootstrap，sub-sampling和随机采样等方法获得；

(4)将数据集的特征空间投影到数据子空间，有随机投影和一维投影等手段。

使用一个聚类算法，目前大多使用K-means算法。因为K-means实现简单，计算复杂度不高，执行速度快。但K-means适宜发现球形簇的数据集；对于结构复杂的数据集，尤其是边界不易区分、非球形分布以及高维数据，不能产生质量较好的聚类成员。使用K-means算法时有两种情况：1)固定 k 值，k 值在算法运行期间保持不变，每次运行随机选择 k 个簇心；2)变化 k 值，允许 k 在一定范围内变化，每次运行先产生一个合理的 k 值，然后随机选择 k 个簇心。

使用多个聚类算法有利于产生出一组具有一定差异度的聚类成员，因为不同的聚类算法对同一数据集有不同的处理能力。Fred等人[14]认为使用不同的方法来产生聚类成员，可以从不同的角度挖掘出模式间的关系。

对于超大规模的数据集，先使用采样技术获得子集，然后对子集进行聚类(体现了聚类集成的并行处理)，产生聚类成员；对于高维数据集，可先对其投影处理，再对投影得到的子集进行聚类，得到聚类成员。

2002年，Fred等人采用变化 k 值的K-means算法，首先产生一个合理的 k 值，$k\in[K\min,K\max]$，其中 $K\min\geqslant 2$，$K\max\leqslant n$(n 为数据集中数据对象数目)，然后随机选择 k 个数据对象作为初始的簇中心对数据集进行划分，生成聚类成员[15]。

2003年，Fern等人使用随机投影技术来解决高维数据的聚类集成。首先把高维数据随机投影到低维空间进行多次投影，得到数据集的多个子集，然后使用EM算法对得到的子集进行聚类，得到聚类成员[16]。

2004年，Minaei-Bidgoli等人比较了在生成聚类成员时使用的两种取样技术bootstrap和subsampling。实验结果表明在聚类集成中，小数目的取样足以代表整个数据集，也减少了计算时间和复杂性[17]。

2004年，Topchy等人使用subsampling设计了一种自适应的动态聚类集成，认为数据集中的数据对聚类结果的贡献是不等的。进行了两次取样，第一次所有数据的取样概率相等，第二次取样数据主要集中在第一次取样数据时聚类结果

不一致的区域，得到数据集的不同子集[18,19]。

2005年，唐伟和周志华在产生初始聚类成员后对其进行了选择，首先使用bootstrap对数据集进行采样，得到不同的子集，运行K-means算法得到多个聚类结果，然后使用规范化互信息(Normalized Mutual Information，NMI)计算每个聚类结果的权值，最后选择权值大于预设阈值的聚类结果作为聚类成员[9]。

2006年，Zhou等人使用固定k值的K-means算法，先随机初始化k个簇中心，然后根据簇中对象的平均值划分数据集，运行算法N次，得到N个聚类结果[20]。

2007年，Gionis等人对同一个数据集使用了SL，CL，AL，Ward's clusterings及K-means等5个聚类算法产生具有差异度的聚类成员[11]，其中SL，CL，AL和Ward's clusterings是层次聚类算法，K-means是划分聚类算法。

集成时使用哪种方法产生聚类成员，可从两方面考虑：集成者的目的和数据集的结构。如果集成者要求时间效率，而数据集的结构较单一，可以采用第一种方法；如果数据集的结构比较复杂，则可采用第二种方法；如果数据集规模较大或是高维数据集，可采用第三种方法，以减少计算时间和降低时间复杂度。

3.2 设计一致性函数

一致性函数(consenus function)是一个函数(或方法)，它将聚类成员进行合并(或称为集成)，得到一个统一的聚类结果[21]。目前存在许多一致性函数，如投票法、超图划分、基于共协矩阵的证据积累、概率积累等。

3.2.1 投票法

投票法在监督式学习的分类集成中很有效，但在聚类集成中存在一个问题：聚类的簇标签对应问题。例如，数据集有7个对象$\{a,b,c,d,e,f,g\}$，有两个划分C_1和C_2，其中$C_1=\{1,1,2,2,2,3,3\}$，$C_2=\{2,2,3,3,3,1,1\}$，如表1所列。

表1 C_1和C_2

	a	b	c	d	e	f	g
C_1	1	1	2	2	2	3	3
C_2	2	2	3	3	3	1	1

表面上看C_1和C_2不同，但逻辑上它们是等价的，都把数据集分为3个簇：$\{\{a,b\},\{c,d,e\},\{f,g\}\}$。在聚类集成中要利用簇标签，必须先解决标签的对应问题。Strehl等人提出一个簇表示方法，通过严格控制簇标签的生成顺序来避免多个数据划分逻辑上等价[10]。Zhou等人提出一个簇排列算法，对聚类成员中的簇，计算两个簇间重叠的数据点个数作为相似度，据此寻找聚类成员中逻辑等价的簇[20]。

投票法的基本思想是尽可能多地共享聚类成员对数据对象的分类信息，根据聚类成员对数据对象的划分进行投票，计算数据对象被分到每个簇的投票比例。依据多数投票超过一定阈值(一般大于等于0.5)来将其划分到这个簇中。

2001年，Fred认为可以在多个数据划分中找到一个一致的聚类结果。他利用聚类成员对数据点的划分信息，计算数据点对同时被分到同一个簇的次数，作为两个数据点是否属于同一个簇的投票。若投票过半，则这两个数据点划分到同一个簇中[22]，据此提出了共协矩阵(Co-association matrix)的概念，并作为相似度矩阵来划分数据。共协矩阵定义为：

$$co\_assoc(i,j)=votes_{i,j}/N$$

式中，N是聚类成员的数目，$votes_{i,j}$是在N个数据划分中模式(i,j)被分到同一个簇中的次数。

2006年，Zhou等人认为一个聚类器由数据集的多个数据划分组成，首先寻找出聚类成员中逻辑等价的簇，对齐所有聚类结果的簇，然后计算不同划分间重叠的数据点数目，判断两个簇是否相似。提出4个基于投票的聚类集成：voting，weigted voting，selective voting和selective weighted-voting，其中voting使用了多数投票策略，数据点被划分到出现次数最多的簇中；weigted voting在投票中使用互信息值作为权重；selective voting使用互信息值作为权重选择可用于集成的聚类成员，然后使用voting划分数据集；selective weighted-voting使用互信息值，既用于选择子集也用于投票[20]。

投票法的优点是简单，易于实现，充分利用了聚类成员对数据点的分类信息；缺点是需要处理簇标签对应问题，只依赖数据点和簇标签之间的关联划分数据。但这种关联较为脆弱，尤其是当聚类成员的质量普遍较差的时候，使用投票法可能得不到较好的数据划分。

3.2.2 超图划分

一般图的边只有两个顶点，超图的一条超边可以有任意多个顶点。聚类成员可以用超图表示：超边表示簇，超边的顶点表示属于该簇的数据点。将聚类集成转化为超图的最小切割问题，使用基于图论的聚类算法进行聚类集成。

2002年，Strehl等人提出了3种基于超图划分的集成方法[10]：1) CSPA (Cluster-based Similarity Partioning Algorithm)是基于实例的超图划分，首先计算共协矩阵作为相似度矩阵，然后以数据点为顶点，以相似度值为边的权重构建超图，再使用图分割算法METIS划分数据集，得到聚类结果；2) HGPA(HyperGragh Partioning Algorithm)是基于聚类的超图划分，它把簇作为超边，超边包含了属于该簇的所有数据点，数据点作为顶点，并且超边和顶点的权重一样，然后使用HMETS算法分割超图；3) MCLA(Meta-CLustering Algorithm)也是基于聚类的超图划分，是对聚类成员再进行聚类。它以簇为顶点，簇之间共同的数据点占所有数据的比例作为边权重，使用METIS算法分裂超边把簇分成几个类，计算数据对象在每个类中的次数，把它分在次数最多的类中。

2004年，Fern等人对Strehl等人提出的超图划分算法进行了改进。他们认为CSPA只利用了点之间的相似性信息，MCLA只利用了簇之间的相似性信息，在构造超图过程中信息会有丢失[25]，因此他们提出一种基于实例和聚类的超图划分方法HBGF(Hybrid Bipartite Graph Formulation)，同时把数据点和簇作为顶点，把簇和在该簇中的所有数据点之间的连接作为超边，所有边的权重相同，构造一个双元图，然后利用图划分算法分割超图，得到最终的聚类结果[23]。

基于超图划分的聚类集成，优点是利用聚类成员来表示数据集的结构，考虑了同簇中数据点之间的关联和不同数据划分之间的关联；但采用的图分割算法有一定的局限性，趋向于将图形划分成大小相似的簇[21]。此外，由于使用互信息优化作为目标函数，使得计算复杂度较高。

3.2.3 证据积累

2002年，Fred等人提出了证据积累(Evidence Accumulation，EA)的聚类集成[24]，其基本思想是：处于同一个自然簇中的数据点在不同的数据划分中可能也属于同一个簇。把每

个聚类成员看作是一个独立的证据，计算两个数据点对被分到同一个簇中的次数，得到共协矩阵，然后使用基于最小生成树(Mininum Spaning Tree，MST)的层次聚类算法(如 SL，AL，CL 等)得到最后的结果。

随后，Ferd 对 EA 进行了改进[15,14]：1)生成聚类成员时，采用了变化 k 值的 K-means 算法，提高了聚类成员的差异度；2)使用层次聚类算法划分数据时，采用了变化的阈值 t，对所有的 t 生成的聚类结果用簇集最高寿命(cluster highest lifetime)作为标准来识别最终聚类结果的簇数目，从而避免了 k 值较大时易把实际上在一个簇中的数据点分到不同的簇中。

2006 年，Luo 等人使用 EA 方法，对混合数据(名词属性和数字属性)进行了聚类分析，其相似度矩阵分为两部分：一是数字属性的相似度；二是名词属性的相似度。先分别计算出两种属性的共协矩阵，然后将两个相似度矩阵叠加，得到总的相似度矩阵，最后使用光谱聚类算法得到一致的数据划分[25]。

基于 EA 的聚类集成使用共协矩阵作为相似度矩阵，计算较为简单，考虑了数据点与数据点之间的关联，缺点是计算的时间复杂度和空间复杂度较高，没有考虑簇集的其它特征(如形状和大小等)。

3.2.4 概率积累

Wang 等人认为 EA 中的共协矩阵通过簇标签寻找两个数据对象间的关联，没有考虑簇的形状和大小等特征，不能很好地度量数据点对间的相似度[6]。2009 年，Wang 等人考虑了簇大小、样品维度和密度等因素，提出了概率积累(probability accumulation，PA)的聚类集成方法[6]。首先对得到的 N 个数据划分使用簇密度计算所有数据点对间的距离，生成每个数据划分的 component 矩阵，然后取所有 component 矩阵的平均值生成 p-association 矩阵，最后使用最高寿命标准，对 p-association 矩阵采用 MST 进行合并，得到最终的聚类结果。

PA 利用簇密度计算数据点对间的距离，考虑 3 种情况：(1)数据点对间的距离可直接使用时，直接计算数据集中所有数据点对间的距离；(2)数据点对间的距离不能直接用时，用簇集的平均数据对距离代表簇集中所有数据点对间的距离；(3)只有划分标签可用时，假设簇集均匀分布，即每个数据对象落在簇集中任何位置的概率是相同的，所有的概率密度函数变成一个常数，再使用概率密度计算平均距离。后两种情况使用簇条件概率密度计算平均距离。

由于 Wang 等人提出的 PA 假设数据集均匀分布，为使数据集满足条件，他们提出一个对原始数据集前/后处理过程[6]：1)前处理，把原始数据集的高维空间映射到大小相等的网格上，数据对象被映射为网格中的点，得到一个新的满足均匀分布的数据集；2)后处理，把新数据集中的划分还原回原始数据集中的划分。实验证明，经过前/后处理的 PA 聚类集成大大提高了聚类结果的质量。

PA 集成算法考虑了簇集的特征，其集成结果优于 EA 集成算法。带有前/后处理的 PA 集成算法对数据集处理之后，其执行效果得到了提高；但是不足之处是计算复杂性较高(同 EA 聚类集成相同)，且当数据集不满足均匀分布时需要对数据集进行前/后处理。

4 其它聚类集成

2005 年，He 等人分析了聚类集成(cluster ensemble，CE)和分类数据的聚类(categorical data clustering，CDC)的异同，指出两者在本质上是相同的，可以使用聚类集成的方法对具有分类属性的数据集进行聚类，把 CDC 作为一个优化问题，使用 Strehl 等人提出的 NMI 和 3 种超图划分方法对具有分类属性的数据集进行聚类分析[26]。

2008 年，He 等人以 NMI 为目标函数解决具有分类属性数据集的聚类分析，提出了 k-ANMI 算法：迭代遍历所有的数据点；改变数据点的簇标签并计算其相应的 ANMI 值，若是 ANMI 值增加，则把此簇标签赋给数据点，直到所有的簇标签都被计算过；然后遍历下一个数据点，直到某次迭代簇标签不再变化则算法终止[27]。

2007 年，Sevillano 等人基于 Borda 投票利用 Bordafuse 数据融合技术提出一个可用于信息检索领域的软聚类集成的一致性函数[28]：首先按照文档属于各个簇的概率大小进行排序；然后依据排序结果依次计算每个文档的成员概率矩阵；最后在此矩阵上进行投票，得到聚类结果。Sevillano 等人认为传统的硬聚类集成的一致性函数只利用簇标签，忽略了聚类成员信息，而软一致性函数考虑了簇成员的属性信息，从而可用来解决多类别属性的聚类问题。

2008 年，Tumer 等人把聚类集成当作一个动态优化问题，以最大化 ANMI 值为目标函数，提出了一种自适应的 VACs(Voting active clusters)投票方法。首先聚类成员中的所有簇对每个数据点进行投票，并使用一个加固学习算法计算每次投票的奖励值，然后把奖励值转化成结果聚类与聚类成员之间的 ANMI，数据点被分给投票最多的簇[29]。

2009 年，Hore 等人从时间和空间复杂性研究了聚类集成的规模扩展性问题，提出两个算法 Bipartite Merger 和 Metis Merger，不计算所有的数据点，而是以簇质心代表整个簇集，在超大规模数据集上进行聚类集成，不仅节省了空间，而且降低了时间复杂性[30]。

2009 年，王红军等人提出一个基于隐含变量的聚类集成(latent variable cluster ensemble，LVCE)模型，把聚类成员看作是原始数据集的特征属性，从而把聚类集成问题转换成聚类问题，对聚类成员进行聚类，不仅简化了集成，而且当数据集增加时，只需对新增的数据集进行处理，使得算法具有扩展性[31]。

2009 年，阳琳赟等人利用粗糙集中的属性重要性理论，考虑了聚类成员集合中每个成员的质量，提出了一种基于属性重要性的加权聚类集成，进行两次集成：首先对聚类成员采用基于共协矩阵的集成算法进行第一次集成；然后利用决策表属性理论给聚类成员赋予不同的权重，生成新的加权共协矩阵，再次进行集成，得到最终的聚类结果[32]。

结束语 集成学习利用多个基学习器解决一个问题，提高了学习系统的泛化能力[10]。近年来，集成技术用于聚类分析，提高了聚类效果。本文介绍了聚类集成的基本概念及优点、聚类成员的产生及用于集成的一致性函数。虽然聚类集成能够获得比单一聚类算法更好的聚类结果，但其计算时间也高于单一聚类算法。单一聚类算法只进行了一次聚类，而聚类集成先产生多个聚类结果(相当于执行多次单一聚类)，

然后进行集成。

经过近几年的发展，还聚类集成已经有了很大的发展，尤其是在一致性函数方面，不断出现新的方法。但聚类集成仍需继续发展，有许多问题尚需进一步研究：

(1) 聚类成员的差异度研究。目前大多数聚类集成采用多次运行同一个基聚类算法产生聚类成员，关于聚类成员的差异度对集成结果的影响研究得较少，因此差异度对聚类结果的影响需要深入地研究。

(2) 设计一致性函数方面，有两个问题：一是目前集成效果较好的是基于共协矩阵的一致性函数，但共协矩阵的时间复杂度和空间复杂性较高，应当研究新的一致性函数；二是考虑结合不同的聚类集成，聚类分析的任务是找出数据集潜在的结构和数据对象之间的关系，而不仅仅是指出数据对象属于哪一个簇。若把软聚类和硬聚类结合起来，能更清晰地说明数据对象之间的关系。

(3) 高维数据集的研究较少。随着信息化的发展，现实数据库中的高维数据、大规模海量数据越来越多，且每天都有新数据不断加入。对于高维数据的聚类集成、可扩展性算法和增量式算法需要更深入地研究。

(4) 对混合型和分类属性数据集的聚类集成需进一步研究。目前，大都采用分别对数字属性和名词属性的数据进行聚类，然后集成，而对两者之间的关系没有深入地研究，可考虑对不同的属性赋予不同的权值。

参 考 文 献

[1] Han Jiawei, Kamber M. Data Mining concepts and techniques [M]. 范明，孟小峰等译. 北京：机械工业出版社，2001：223-259

[2] Judd D, Mckinley P, Jain A K. Large-scale parallel data clustering[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20(8): 871-876

[3] Bhatia S K, Deogun J S. Conceptual clustering information retrieval[J]. IEEE Transactions on Systems, Man, and Cybernetics, 1998, 28(3): 427- 436

[4] Frigui H, Krishnapuram R. A robust competitiveclustering algorithm with applications in computer vision[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1999, 21 (5): 450-465

[5] Jain A K, Murty M N, Flynn P J. Data clustering: A review[J]. ACM Computing Surveys, 1999, 31(3): 264 -323

[6] Wang Xi, Yang Chunyu, Zhou Jie. Clustering aggregation by probability accumulation[J]. Pattern Recognition, 2009, 42(5): 668-675

[7] 孙吉贵，刘杰，赵连宇. 聚类算法研究[J]. 软件学报，2008，19 (1)：48-61

[8] Fraley C, Raftery A E. How many clusters ? Which clustering method? Answers via model based cluster analysis [J]. The Computer Journal, 1998, 41(8): 578-588

[9] 唐伟，周志华. 基于 Bagging 的选择性聚类集成[J]. 软件学报，2005, 16(4): 496-502

[10] Strehl A, Ghosh J, Cardie C. Cluster ensembles: A knowledge reuse framework for combining multiple partitions[J]. Journal of Machine Learning Research, 2002(3): 583-617

[11] Gionis A, Mannila H, Tsaparas P. Clustering aggregation[J]. ACM Transactions on Knowledge Discovery from Data, 2007, 1 (1): 1-30

[12] Topchy A, Jain A K, Punch W. Clustering Ensembles: Models of Consensus and Weak Partitions[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(12): 1866-1881

[13] 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

[14] Fred A, Jain A. Combining multiple clusterings using evidence accumulation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(6): 835-850

[15] Fred A, Jain A K. Evidence accumulation clustering based on the k-means algorithm[A]// Proceedings of the Joint IAPR International Workshop on Structural, Syntactic, and Statistical Pattern Recognition[C]. 2002: 442-451

[16] Fern X Z, Brodley C E. Random projection for high dimensional data clustering: A cluster ensemble appRoach[C]// Proceedings of the Twentieth International Conference on Machine Learning (ICML'2003). 2003: 186-193

[17] Minaei-Bidgoli B, Topchy A, Punch W F. A comparison of resampling methods for clustering ensembles [C] // International Conference on Machine Learning, Models, Technologies and Applications. 2004: 939-945

[18] Topchy A, Minaei-Bidgoli B, Jain A K, et al. Adaptive clustering ensembles[C] // Proceedings of the 17th International Conference on Pattern Recog nition(ICRP 2004). 2004: 272-275

[19] 阳琳赟，王文渊. 聚类融合方法综述[J]. 计算机应用研究，2005，22(12)：8-10，14

[20] Zhou Zhihua, Tang Wei. Clusterer ensemble[J]. Knowledge-Based Systems, 2006, 19(1): 77-83

[21] 罗会兰. 聚类集成关键技术研究[D]. 杭州：浙江大学，2007

[22] Fern A. Finding consistent clusters in data partitionns[C]// Proceedings of the 2nd International Workshop on Multiple Classifier Systems. 2001: 309-318

[23] Fern X Z, Brodley C E. Solving cluster ensemble problems by bipartite graph partitioning[C]// Proceedings of the 21st International Conference on Machine Learning. 2004: 36-43

[24] Fred A, Jain A K. Data clustering using evidence Accumulation [C]// Proceedings of the 16th International Conference on Pattern Recognition(ICPR'02). 2002: 276-280

[25] Luo Huilan, Kong Fansheng, Li Yixiao. Clustering mixed data based on evidence accumulation: Advanced Data Mining and Applications[C] // LNCS 4093, 2006. Heidelberg-Berlin: Springer, 2006: 348-355

[26] He Zengyou, Xu Xiaofei, Deng Shengchun. A cluster ensemble method for clustering categorical data[J]. Information Fusion, 2005, 6(2): 143-151

[27] He Zengyou, Xu Xiaofei, Deng Shengchu. k-ANMI: A mutual information based clustering algorithm for categorical data [J]. Information Fusion, 2008, 9(2): 223-233

[28] Sevillano X, Alias F, Socoro J C. BordaConsensus: a new consensus function for soft cluster ensembles[C]// Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 2007: 743-744

[29] Tumer K, Agogino A K. Ensemble clustering with voting active clusters[J]. Pattern Recognition Letters, 2008, 29 (4): 1947-1953

[30] Hore P, Hall L O, Goldgof D B. A scalable framework for cluster ensembles[J]. Pattern Recognition, 2009, 42(5): 676-688

[31] 王红军，李志蜀，成颐，等. 基于隐含变量的聚类集成模型[J]. 软件学报，2009，20(4)：825-833

[32] 阳琳赟，周海京，卓晴，等. 基于属性重要性的加权聚类融合[J]. 计算机科学，2009，36(4)：243-245