

# 聚类融合算法研究

秦 锋, 陈奇明, 程泽凯

(安徽工业大学 计算机学院, 安徽 马鞍山 243002)

**摘 要:** 聚类是发现数据分布和隐含模式的一项重要技术, 但单一聚类算法却很难达到预期的效果。在缺乏样本集先验知识的前提下, 目前的分类融合技术很难应用到聚类技术中, 导致聚类融合技术起步很晚。近几年的研究发现, 聚类融合方法对提高聚类算法的稳定性和高效性发挥了重要的作用。文中对近年来聚类融合的方法和国内外研究现状进行了简单综述, 并且以基于投票的聚类融合算法为例, 实验证明了其比单一聚类算法的优越性, 展望了聚类融合算法的未来。

**关键词:** 聚类; 融合技术; 差异度; 投票

**中图分类号:** TP18

**文献标识码:** A

**文章编号:** 1673-629X(2010)07-0106-03

## An Overview of Clustering Ensemble Approaches

QIN Feng, CHEN Qi-ming, CHENG Ze-kai

(School of Computer Science, Anhui University of Technology, Maanshan 243002, China)

**Abstract:** Clustering is a technique for the discovery of data distribution and latent data pattern. Single clustering is hard to reach good result. However, in unsupervised learning, researches of ensemble approaches are concerned only in recent years. Because of the premise of the prior knowledge of sample sets is insensible, the ensemble approaches of classifier can't be utilized in the same way directly. Recent studies proof that clustering ensemble approaches can enhance robustness and stabilities greatly. Makes an overview and the research status at home and abroad of the clustering ensemble approaches. It makes an example of clustering ensemble algorithm based on voting, show superiority than single. Finally, prospect the future of clustering ensemble approaches.

**Key words:** clustering; ensemble technique; diversity; voting

## 0 引言

聚类<sup>[1,2]</sup>是一项重要的发现数据分布和隐含模式的数据挖掘技术。衡量高质量聚类算法的标准是具有类内相似性强, 而类间相似性弱, 同时还取决于算法能否发现隐藏在数据中隐含模式和有价值的信息。

融合聚类技术<sup>[3~5]</sup>是利用不同的算法或同一算法下不同参数得到的聚类结果进行融合, 得到比单一聚类算法更高效稳定的聚类结果。文中对近年来聚类融合的方法和国内外研究现状进行了综述, 并且以基于投票的聚类融合算法为例, 实验证明了其比单一聚类算法的优越性, 最后展望了聚类融合算法的未来。

再通过融合技术合并。详细说明如下所示:

设有  $N$  个样本的样本集  $X$ , 对数据集  $X$  进行  $H$  次不同聚类算法得到  $H$  次聚类结果。通过共识函数  $G$ , 把  $H$  次聚类得到的聚类成员融合得到  $C^*$ , 融合过程如图 1 所示。

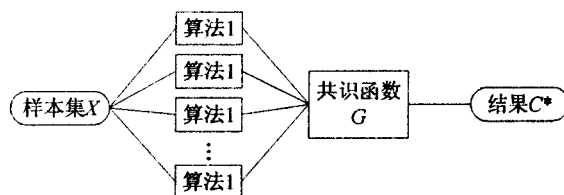


图 1 融合过程图

聚类融合算法比单一聚类算法具有的优势如下:

- ① 在各个领域和试验中表明了平均性能的优越性能, 也就是鲁棒性。
- ② 能够适应大部分领域和实验, 也就是适用性。
- ③ 杂音数据和异常数据对聚类融合的结果影响很小, 也就是稳定性。

聚类融合的过程是: 首先对样本集进行多次聚类产生聚类成员, 其次对聚类成员融合得出聚类结果。

## 1 聚类融合方法和国内外现状概述

聚类融合最早是由 A. L. Fred 和 A. Strehl<sup>[6]</sup>提出的, 被定义为把多个数据样本进行划分, 分为多个类。

收稿日期: 2009-11-13; 修回日期: 2010-02-04

基金项目: 安徽省自然科学基金资助项目(KJ2007A051)

作者简介: 秦 锋(1962-), 男, 安徽马鞍山人, 教授, 硕士生导师, 研究方向为人工智能、计算机网络方向。

所以,目前研究的热点应该是:a.如何产生高效的聚类成员;b.共识函数如何构建才能产生高效的聚类融合算法。B. Minaei - Bidgoli<sup>[7]</sup>对聚类融合今后的研究方向作了图2说明。

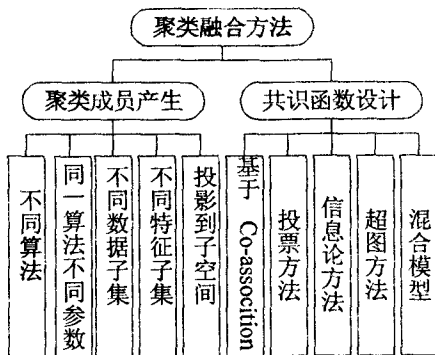


图2 聚类融合前景展望图

### 1.1 聚类成员的产生

通过图2所示可以知道聚类成员有很多的方法来产生:X. Z. Fern<sup>[8]</sup>提出了随机投影法,方法是将通过随机投影产生多个样本子集,然后用EM聚类算法对子集进行聚类产生聚类成员。A. Topchy 主要是通过研究弱聚类产生聚类成员来试验聚类融合的效果和影响,事实证明,弱聚类产生聚类成员的聚类融合效果比单一的聚类效果要优越得多。B. Minaei - Bidgoli<sup>[9]</sup>则采用了数理统计中随机抽样技术来产生聚类成员,并利用经典k-means算法对其进行聚类产生聚类成员。后来,A. Topchy 又提出聚类成员的自适应产生的方法<sup>[10]</sup>,实验证明了可行性和高效性。

### 1.2 共识函数的设计

据图2可知,共识函数的设计方法主要有共联矩阵法、投票法、信息论方法、超图法和混合模型法,具体介绍如下:

A. L. Fred<sup>[11]</sup>提出了共联矩阵的方法,用于记录样本之间的相似度,然后运行 single-linkage 算法进行融合产生聚类结果。A. L. Fred 提出基于投票策略的聚类融合算法,方法是利用经典K-means聚类算法随机对样本进行H次聚类,并且通过共联矩阵进行聚类融合产生最后的聚类结果。A. Topchy 主要利用信息论中的方差理论构建了共识函数完成聚类的融合,实验证明了有效和高效性。基于超图的方法主要是由A. Strehl 和 J. Ghosh<sup>[6]</sup>等人提出,CSPA(Cluster-based Similarity Partitioning Algorithm)方法主要是采用了共联矩阵法和图论方法论中的METIS算法来完成聚类融合;HGPA(Hyper Graph Partitioning Algorithm)则是采用超图的方法,并通过超图算法HMETIS进行融合得到聚类最终结果;最后,MCLA(Meta-Clustering Algorithm)是一种对聚类成员再聚类的超图算法,实验

证明了比单一算法更加的高效和稳定。A. Topchy 利用分布式的混合模型来构建共识函数并且采用EM聚类算法完成最终的结果,具有算法复杂度低、算法执行速度快和避免标签匹配的问题的优点。

## 2 基于投票的聚类融合算法

从聚类融合的方法中得知,基于投票策略的融合聚类算法首先要解决三个问题:

- ①如何产生N个具有差异性的聚类结果;
- ②在N个差异性聚类结果中,如何解决样本的类标签不一致性问题;
- ③如何采用投票的方式对样本进行归类,得到最终的聚类结果。

### 2.1 生成聚类成员

文中采用的是经典的k-means聚类算法,由于每次随机选取K个不同的聚类中心进行N次聚类,则必然会产生N个具有差异性的聚类结果。

设定N个样本集 $X = \{X_1, X_2, \dots, X_N\}$ ,初始给定聚类个数K和随机产生K个初始聚类中心,最后聚类结果 $C = \{C_1, C_2, \dots, C_K\}$ ,算法的伪代码如下:

- ①从样本集中随机选取K个数据作为聚类的初始中心;
- ②利用欧氏距离公式计算每一个数据到K个聚类中心的距离,并将自己分配到距离最小的聚类当中;
- ③重新计算新聚类的中心;
- ④重复②、③,直到新的聚类中心不变为止。

### 2.2 类标签的转换

k-means聚类算法是随机初始化K个类中心点进行多次聚类,则可能会出现某个样本在多次聚类中隶属于不同的类标签,这就导致了标签的不一致性问题,因此,在采用基于投票的聚类融合算法之前必须要解决类的标签不一致性问题,使得同一样本在多次聚类算法中隶属的类标签号一样,具体解决算法如下:

假设运行N次k-means算法,每次随机选取K个类中心点,得到N次不同的聚类结果 $C_1, C_2, \dots, C_N$ ,即 $C_1 = \{P_{11}, P_{12}, \dots, P_{1K}\}$ , $C_2 = \{P_{21}, P_{22}, \dots, P_{2K}\}, \dots, C_N = \{P_{N1}, P_{N2}, \dots, P_{NK}\}$ 。选择第一次k-means聚类算法作为基点,依次与其他的进行比较,假设选定来自第一次k-means算法聚类结果 $C_1$ 中的类成员 $C_{1i}$ 和第二次k-means算法 $C_2$ 中的类成员 $C_{2j}$ ,定义一个矩阵N,用 $N_{ij}$ 来记录类成员 $C_{1i}$ 和 $C_{2j}$ 中相同的数据数,依次计算出所有类似类对之间重叠的数据个数,找出重叠数最多的两个聚类,用共同的标签表示。具体算法描述如下:

for(int K - means = 1; K - means < N; K - means

```
++){\n  for(int P1 = 0;P1 < K;P1 ++){\n    for(int Pi = 0; Pi < K;Pi ++){\n      计算分别属于 C1 中的类成员 P1i 与属于其他 k -\n      means 算法聚类结果 Ci 中的类成员 Pij 的重叠样本数,\n      存放到矩阵 Nij 当中;\n    }\n  }\n\n  扫描矩阵 N 中的最大值 Nij, 即聚类结果 C1 中的\n  第 i 个类 P1i 与聚类结果 Ci 中的类成员的 Pij 样本的重\n  叠数, 返回类 i 和类 j, 并且用相同的类标签进行标识,\n  文中用第一次 k - means 算法为基准, 即 j = i\n}
```

2.3 投票法则

经过标签转换之后,所有的聚类结果的类标签就一致了,下面将详细介绍投票的过程:设定一个矩阵 Matrix[ N][K],N:样本数据的个数,K:类的个数,用来存放每一个样本属于某个类的次数,最后扫描矩阵 Matrix[N][K],记录每一个样本属于某个类的最大值,把样本归于次数最大的列所标识的类,得到最终的聚类结果。而矩阵 Matrix[N][K] 产生的算法详细介绍如下:

设定一个矩阵 MergeMtr[X][ Y],X:属于某个类的总样本数,Y:样本的维数,用来存放隶属于某个类的所有样本,Pattern[N][ Y],N :样本的总数,Y:样本的维数,用于存放原始的样本,扫描并计算样本矩阵 Pattern[N][ Y] 中的每一个样本 Pattern[ i][ Y] 分别在每一个类样本矩阵 MergeMtr[X][ Y] 中出现的次数,填充 Matrix[N][K]。

3 实验设计与结果分析

在本实验中,采用 UCI 中的 iris 数据集作为挖掘对象,其真实数据特征描述如表 1 所示,实验环境是 Celeron(R) CPU 2.8GHz,512M 内存,Windows XP 操作系统,编程语言采用的是 Win32 console 控制台程序。

表 1 UCI 数据的属性描述

数据集	属性数	类数	数据个数	数据分布
iris	4	3	150	50-50-50

首先设定聚类的个数 K,文中选取与 UCI 真实数据中相同的类数,通常用平方误差准则作为衡量聚类结果好坏的原则,其表达式为:

$$E = \sum_{i=1}^K \sum P \in ci | P - mi |^2$$

其中 E 是所有样本的平方误差的总和,K 是类的个

数,P 是空间中的点即样本,mi 是隶属于某个类 ci 的平均值。文中用 iris 数据集作为实验对象,用平方误差准则来衡量单一 k - means 算法与基于投票的聚类融合算法性能的优劣,其结果如表 2、图 3 所示。

表 2 iris 数据集平方误差实验结果

次数 N	平方误差 E	次数 N	平方误差 E
1	94.183	6	143.454
2	78.941	7	78.941
3	99.801	8	142.859
4	78.941	9	84.273
5	158.467	10	81.056
k - means 平均值		104.092	
基于投票算法		71.282	

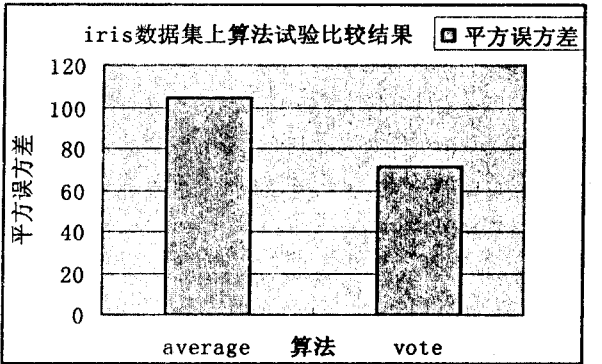


图 3 iris 数据集平方误差实验比较结果

图 3 中列出的是 k - means 算法运行 10 次,每一种算法运行 100 次后取平均值的平方误差,根据平方误差准则,E 越小则聚类的效果越好,通过图 3 的实验数据证明了基于投票策略的聚类融合算法所得到实验数据普遍优于单一的 k - means 算法的实验数据。由于文中算法的时间复杂度为 O(M \* N<sup>2</sup>),M 表示重复执行 k - means 或者融合算法的次数,N 表示样本的个数,关联矩阵空间复杂度为 O(N<sup>2</sup>),因此文中方法及同类方法不适用于大规模 N 海量数据的聚类,如何将融合技术应用于大规模数据集的聚类是进一步要研究的问题。

4 结束语

目前在聚类融合方法的研究上面还远远没有达到成熟的程度,将来的研究方向大致可以具体描述如下:

- ①关键参数的确定。参数的确定是目前聚类融合算法的一个研究热点。
- ②聚类成员的产生。在聚类融合中,利用什么样的算法来产生高效聚类成员将是一个新的研究方向。
- ③共识函数设计。要充分考虑到软聚类和硬聚类的融合给聚类结果带来的高效性。

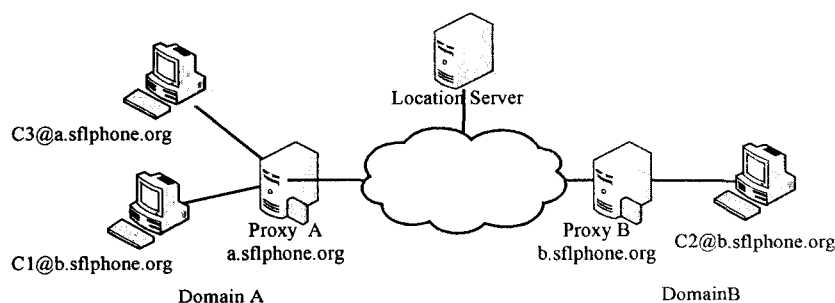


图 3 用户 C1 移动时的试验环境

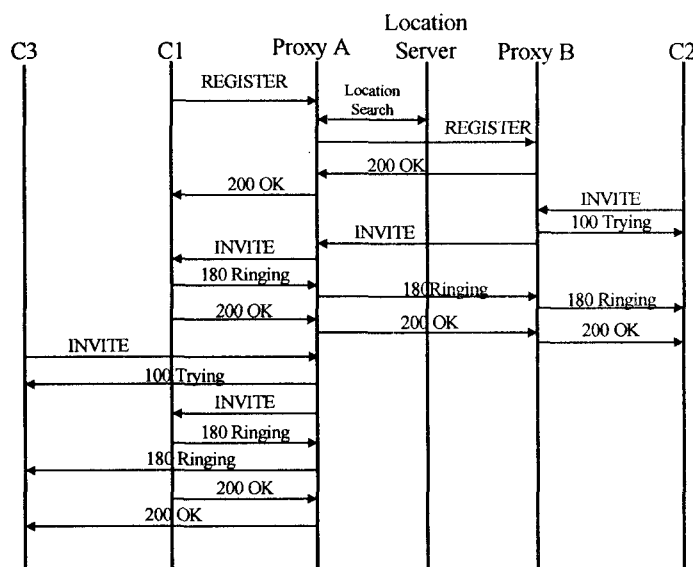


图 4 用户 C1 移动时的通信时序图

益的补充。随着人们对移动通信要求的提高,基于 SIP 在应用层解决移动性问题将具有更加广阔的应用前景。后继工作包括采用 SIP 机制来研究解决终端移

动性、会话移动性以及服务移动性所面临的技术问题。

## 参考文献:

- [1] Popescu L. Supporting Multimedia Session Mobility using SIP[C]// Communication Networks and Services Research Conference 2003. Moncton, New Brunswick, Canada; [s. n.], 2003.
- [2] Rosenberg J. SIP: Session Initiation Protocol[S]. IETF, RFC 3261. 2002.
- [3] 张永强, 张捍东, 赵金宝. SIP 协议栈研究[J]. 计算机技术与发展, 2007, 17(11): 55-57.
- [4] 张 荣, 武 波. SIP 协议的应用研究[J]. 计算机技术与发展, 2006, 16(6): 71-73.
- [5] Vakil F. Mobility Management in a SIP Environment[S]. IETF, Internet Draft, 2000.
- [6] Schulzrinne H, Wedlund E. Application - Layer Mobility Using SIP[J]. Mobile Computing and Communications Review, 2001, 4(3): 47-57.
- [7] 黄 斌, 李秉智. 基于 SIP 的 VoIP 的移动性研究[J]. 微计算机技术, 2008, 24(1): 145-147.
- [8] Wedlund E, Schulzrinne H. Mobility support using SIP [C]// ACM/IEEE International Conference on Wireless and Mobile Multimedia. Seattle, Washington: [s. n.], 1999.
- [9] 胡凌凌, 彭容修. SIP 协议在一个 IP 电话模型中的实现[J]. 微机发展(现更名: 计算机技术与发展), 2005, 15(2): 100-102.

(上接第 108 页)

④目前, 聚类融合方法科学研究和具体应用上前景广泛, 其研究有重要意义。

## 参考文献:

- [1] 韩家炜. 数据挖掘: 概念与技术[M]. 北京: 清华大学出版社, 2000.
- [2] 李雄飞, 李 军. 数据挖掘与知识发现[M]. 北京: 高等教育出版社, 2003.
- [3] 阳琳贤, 王文渊. 聚类融合方法综述[J]. 计算机应用研究, 2005(12): 8-10.
- [4] 蒋盛益. 基于投票机制的融合聚类算法[J]. 小型微型计算机系统, 2007(2): 306-309.
- [5] 邹远强, 李国徽, 赵梓屹. 基于遗传和蚁群算法融合的聚类新方法[J]. 科学技术与工程, 2006(23): 4700-4705.
- [6] Strehl A, Ghosh J. Cluster Ensembles: A Knowledge Reuse Framework for Combining Multiple Partitions[J]. Journal of Machine Learning Research, 2003, 3(3): 583-617.
- [7] Minaei - Bidgoli B, Topchy A, Punch W F. A Comparison of Resampling Methods for Clustering Ensembles[C]// Int. Conf. on Machine Learning, Models, Technologies and Applications (MLMTA 2004). [s. l.]: [s. n.], 2004: 939-945.
- [8] Fern X Z, Brodley C E. Random projection for High Dimensional Data Clustering: A Cluster Ensemble Approach[C]// Proceedings of the 20th International Conference on Machine Learning. [s. l.]: [s. n.], 2003: 186-193.
- [9] Minaei - Bidgoli B, Topchy A, Punch W F. Ensembles of Partitions via Data Resampling[C]// Proceedings International Conference on Information Technology, Coding and Computing (ITCC 2004). [s. l.]: [s. n.], 2004: 188-192.
- [10] Topchy A, Minaei - Bidgoli B, Jain A K, et al. Adaptive Clustering Ensembles[C]// Proceedings of the 17th International Conference on Pattern Recognition (ICPR 2004). [s. l.]: [s. n.], 2004: 272-275.
- [11] Fred A L. Finding Consistent Clusters in Data Partitions[C]// Proceeding of the 2nd International Workshop on Multiple Classifier Systems, Volume 2096 of Lecture Notes in Computer Science. [s. l.]: Springer, 2001: 309-318.