分类号： TP391　　单位代码： 10335

密　级： 公开　　学　　号： 21021238

浙江大学

硕士学位论文

中文论文题目： 符号属性数据的半监督聚类与属性选择

英文论文题目： **Semi-supervised Clustering and Feature Selection for Symbolic Data**

申请人姓名： 王文涛

指导教师： 代建华副教授

专业名称： 计算机应用技术

研究方向： 人工智能　　机器学习

所在学院： 计算机科学与技术学院

提交日期 2013-01-05

符号属性数据的半监督聚类与属性选择

论文作者签名： _____

指导教师签名： _____

论文评阅人1: 梁荣华 教授 浙江工业大学

评阅人2: 曹飞龙 教授 中国计量学院

评阅人3: 廖备水 副教授 浙江大学

评阅人4: _____

评阅人5: _____

答辩委员会主席: 张志华 教授 浙江大学

委员1: 廖备水 副教授 浙江大学

委员2: 郑能干 副教授 浙江大学

委员3: 代建华 副教授 浙江大学

委员4: _____

委员5: _____

答辩日期： 2013-03-07

Semi-supervised Clustering and Feature Selection for Symbolic Data

Author's signature:_____

Supervisor's signature:_____

External Reviewers: Ronghua Liang Professor

Feilong Cao Professor

Beishui Liao Associate Professor

Examining Committee Chairperson:

Zhihua Zhang Professor

Examining Committee Members:

Beishui Liao Associate Professor

Nenggan Zheng Associate Professor

Jianhua Dai Associate Professor

Date of oral defence: 2013-03-07

浙江大学研究生学位论文独创性声明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作及取得的研究成果。除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得浙江大学或其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示谢意。

学位论文作者签名：　　签字日期：　　年　　月　　日

学位论文版权使用授权书

本学位论文作者完全了解浙江大学有权保留并向国家有关部门或机构送交本论文的复印件和磁盘，允许论文被查阅和借阅。本人授权浙江大学可以将学位论文的全部或部分内容编入有关数据库进行检索和传播，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。

（保密的学位论文在解密后适用本授权书）

学位论文作者签名：　　导师签名：

签字日期：　　年　　月　　日　　签字日期：　　年　　月　　日

摘　要

在机器学习领域中，聚类和属性选择问题已经成为分析数据的有效手段。聚类是无监督学习的基本技术，其目的是在没有数据先验信息下分析数据的结构。一般而言，一个好的聚类算法要遵循类内(Intra-class)对象相似度最大而类间(Inter-class)对象相似度最小的原则。属性选择是在属性全集中选择重要的属性去掉冗余的属性，在提高学习效率和预测精度，降低算法复杂度都有明显的效果。

近年来，半监督学习成为一个研究热点，而其中的半监督聚类与半监督属性选择是重要研究内容。然而，绝大多数现有半监督聚类与半监督属性选择方法关注的是连续性属性数据，对符号属性数据相关研究还比较少。在现实应用中符号数据大量存在，因此符号属性数据的半监督聚类和属性选择是也是亟待研究的内容。本文对符号属性数据的半监督聚类与属性选择进行了研究，分别提出两种半监督聚类方法和半监督属性选择方法。

基于聚类集成思想，提出了一种符号属性数据半监督聚类的集成策略。为了有效进行集成，构造了四种基于权重的投票策略去获得最终的聚类结果。此外，提出了一种分裂再组合的聚类方法，利用无监督和有监督信息的形成等价关系，将样本划分成一个个小的簇，然后再将这些小簇通过基于不同集簇间距离度量策略的层次聚类方法进行组合得到最终的聚类结果。

受到监督学习中属性选择算法mRMR的启发，本文重新定义了半监督环境下的属性相关性和冗余性，构造一种最小冗余最大相关的符号数据半监督属性选择算法。此外，将粗糙集理论中传统的依赖度拓展到了半监督领域，提出了耦合依赖度的概念，它不仅可以度量条件属性对决策属性的依赖程度，还能度量条件属性间的冗余程度。基于耦合依赖度，构造了一种符号数据半监督属性选择算法。

实验结果表明，所提出的半监督聚类和属性选择方法能有效实现符号属性数据半监督聚类和特征选择。

关键词： 机器学习，半监督学习，聚类，属性选择，符号属性数据，聚类集成

Abstract

In the fields of the machine learning, clustering and feature selection have been providing an effective and efficient method for data analysis. Clustering is a fundamental technique of unsupervised learning, where the task is to find inherent structure form unlabeled data. A good cluster should divide the data into several clusters so that the intra-cluster similarity is maximized while the inter-cluster similarity is minimized. On the other hand, feature selection is applied to reduce the number of features in many applications where datasets have hundreds or thousands of features. It has proven in both theory and practice effective in enhancing learning efficiency, increasing predictive accuracy, and reducing complexity of learned results, especially for high-dimensional datasets.

Recently, Semi-supervised learning has become an attractive methodology for improving classification models and is often viewed as using unlabeled data to aid supervised learning. Similarly with supervised learning, semi-supervised clustering and semi-supervised feature selection have been active areas. However, most of those semi-supervised methods are applied to continual-value data, few methods are suitable for clustering and feature selection of symbolic value data in semi-supervised learning. This paper proposes two semi-supervised clustering algorithms and two semi-supervised feature selection approaches for symbolic data respectively.

The first method is based on clustering ensemble and the clusterers are generated by k-Modes. In addition, we provide four voting strategies to obtain the final clustering results. Base on split-merge model, we propose a semi-supervised symbolic clustering method. Though equivalent relations using unsupervised and supervised information, we obtain small partitions where objects are similar. The final cluster partition are merged by four different distances measurements between clusterings.

Inspired feature selection method mRMR in supervise learning, semi-supervised relevance and redundance measurement are redefined and a novel stop criterion is proposed to control the number of feature selection. In the last method, we extend the classical dependence degree in

rough set to the semi-supervised framework, called daul dependence degree. The dual degree measures not only the dependence with respect to the decision attribute, but also the redundancy between conditional attributes. For the proposed semi-supervised feature selection, we present two different search feature subset strategies.

Experiments show that our semi-supervised clustering and feature selection methods for symbolic data are effective and efficiency, which can provide an alternative solution for semi-supervised learning.

Keywords: Machine Learning, Semi-supervised Learning, Clustering, Feature Selection, Symbolic Data, Clustering Ensemble

目　　录

图 目 录

表 目 录

第1章 绪论

近年来，机器学习(Machine Learning)已经成为人工智能领域的一个研究热点。它被广泛地应用于机器人、生物信息学、图像识别、语音识别、传感器网络、信息安全、工业过程控制等多个领域。

1.1 机器学习简介

机器学习是研究计算机怎样模拟或实现人类的学习行为，以发现新的知识，重新组织已有的知识结构并不断改善自身的性能。Tom M.Mitchell认为，机器学习是“计算机利用经验改善系统自身性能的行为”[1]。美国航空航天局JPL实验室指出了“机器学习对科学研究的整个过程正起到越来越大的支持作用，.....，该领域在今后的若干年内将取得稳定而快速的发展”[2]。在机器学习领域中，按照传统可以划分为：有监督学习(Supervised Learning)和无监督学习(Unsupervised Learning)，近年来，半监督学习(Semi-supervised Learning)引起了许多研究机构的兴趣。半监督学习主要包括：半监督分类，半监督回归，半监督聚类以及半监督属性选择。

1.1.1 监督学习

监督学习是指通过对数据集进行学习和训练并构建模型，然后对新的样本进行预测的学习方法。实质上，监督学习提供一种函数映射，用来学习数据集中给定的对象到类别的映射关系并对新的样本给出映射结果。

1.1.2 无监督学习

无监督学习是指待训练的数据集没有人工标注的类标。无监督学习的目的是找出数据集中蕴含的各种结构信息，从而发现数据集的规律，找出有价值的信息。无监督学习大体包括：无监督聚类，无监督属性选择和降维等。

1.1.3 半监督学习

对于监督学习，为了训练一个分类函数或构建模型，需要数据集为带类标数据(Labeled Data)，而得到标记数据通常是很困难也很费时，往往需要人工进行标注。另一方面，不带类标数据(Unlabeled Data)容易收集，但无监督学习中，对解空间的搜索具有一定的盲目性，结果在有些情况下较差[3]。半监督学习是一种介于监督和无监督学习之间的方法，通过少量的带类标数据来“指示”或“引导”对未知样本的学习。这样，假设整个数据集$X=\{x_1,x_2,\ldots,x_n\}$，则$X=X^L\cup X^U$。其中X^L表示有类标数据，X^U表示无类标数据。这种事半监督学习中最为基本的描述，随着半监督学习的不断研究，出现了许多新的表现形式，如约束指导学习，直推学习等。

1.2 研究背景及主要内容

1.2.1 研究背景

在机器学习领域中，聚类和属性选择问题是重要的研究方向。聚类问题就是在没有任何数据的先验信息下对数据进行聚类分析，它是一种有效的分析数据结构的手段。在半监督学习中，国内外的研究重点主要针对半监督学习中的分类和回归问题，对于半监督聚类问题的研究则相对较少。于此同时，对于符号属性数据的聚类一直是研究的热点和难点。因此，符号属性数据的半监督聚类具有研究价值。

另一方面，属性选择是降低特征维度的一种重要方式，是解决“维度灾难”的有效手段。但是大多数属性选择方法是考虑与类别之间的相关性的，也就是在有监督情况下的，而对于半监督监督属性选择相关的方法较少，因此设计出既能考虑无监督信息又能考虑半监督信息的属性选择方法是一个值得重点研究的问题。

1.2.2 研究内容

针对上一节介绍的两方面问题，本文对符号数据半监督聚类和属性选择问题进行了详细的分析与研究，并分别提出了两种符号属性数据的半监督聚类方法和两种属性选择方法，为半监督学习提供了新的思路和解决方案。主要内容如下:

1) 结合聚类集成思想，提出了基于权值投票的半监督聚类集成方法，对于每个聚类成员，我们分别计算有监督的权重和无监督的权重，共同投票产生最终的聚类结果。并提出了四种不同的投票策略。最后分别对不同投票策略做了验证与比较实验。

2) 通过对整个数据集先分裂再组合，提出了基于分裂再组合的半监督聚类方法，并比较了不同的分裂策略和组合策略。实验表明，该方法能随着带类标的比重不断加大，效果不断提升。

3) 受到mRMR算法启发，提出了一种最小冗余最大相关的半监督属性选择方法，方法中重新定义了属性的相关性和属性间的冗余性，不仅考虑无监督数据的信息，而且还考虑了单个属性对整个属性子集的作用。

4) 拓展了经典粗糙集中的依赖度定义，使其不仅能度量属性与类别相关性，而且能度量两个属性之间的冗余性。并提出了属性一致性度量的概念和对应的属性选择算法。

1.3 论文结构

本文总共分为7章，每章的主要内容如下:

第 2章主要介绍了聚类和属性选择的基本概念，并详细阐述了聚类和属性选择的研究现状及进展。

第 3章介绍了聚类集成的相关概念，并提出了一种基于聚类集成思想的半监督聚类方法，该方法适用于符号属性数据聚类。并对该方法做了大量实验进行验证与比较。

第 4章对等价类以及划分进行了简单的介绍，提出了一种分裂再组合的半监督聚类方法，该方法先通过无监督信息和有监督信息构建等价类，然后再以不同的组合策略进行层次聚类，最后做了不同集簇距离策略的对比实验。

第 5章简单阐述了mRMR算法，提出了一种基于最小冗余最大相关的半监督属性选择方法，利用有监督信息和无监督信息，我们重新定义了属性的冗余性和相关性，并提出了一种启发式算法。实验验证了其有效性。

第 6章介绍了粗糙集的相关概念，并提出了一种耦合依赖度的度量方式和相应的属性选择算法。最后做了验证性实验，并统一比较了本文提出的四种方法。

第 7为本文的总结，并展望了符号属性数据的半监督聚类和属性选择仍需努力的研究方向。

第2章 聚类与属性选择概述

本章将主要介绍聚类和属性选择的基本概念，并将介绍聚类和属性选择的研究现状及大体分类。

2.1 聚类算法

2.1.1 聚类概念与相关定义

聚类分析是一种基本的无监督学习方法，同时也是多元统计分析的一个重要研究内容，已被广泛的运用到许多应用领域中，例如：模式识别，数据挖掘，图像音频处理，信息论，生物信息学等领域[4–9]。所谓聚类分析就是将无标签数据按照某种度量，根据数据自身的特征，组织成具有不同特点的集簇[10]。聚类分析的基本目标是发现样本集合的自然分组方法[10]。为此，必须先定义度量尺度，借以度量对象之间的联系，这个定量尺度就叫做对象相似性。为了从复杂的数据集中产生出比较简单的结构，多数做法都要求有一个“接近程度”或“相似性”的量度。在选择相似性量度时，通常带有相当大的主观性。

2.1.2 聚类算法的分类

聚类方法主要包括基于划分的方法、层次方法、基于密度的方法、基于网格的方法以及基于模型的方法等。

1) 基于划分方法：通过给定待对象为n的数据集以及期望生成集簇的数目$k(k \leqslant n)$，根据划分策略将样本分成k个集合，算法过程通过重复迭代将对象分配到最近的类中，直到聚类结果达到相应的收敛准则。典型的算法有：k-Means算法、CLARANS算法以及PAM算法。

2) 基于层次方法：采用分裂或凝聚的方式在不同层次对数据进行划分，聚类结果以树形聚类图表示。凝聚聚类初始将每个数据作为一个单独的类，随后逐层合并对象直

到所有对象被合并为一个类。分裂聚类则相反，层次聚类算法的典型算法有Birch算法和Chameleno算法等。

3) 基于密度方法：主要依据数据集的密度分布来完成聚类过程。密度聚类的优点是不受数据集的形状影响，并可以过滤噪声对象。算法将密度大的区域的对象划分到一起，密度小的对象则被分离开来，DBSCAN算法和OPTICS算法是基于密度的聚类代表性算法。

4) 基于网格方法：将数据首先转化为网格结构，然后基于网格结构进行聚类，该算法拥有处理高效以及算法性能与数据对象的数量无关等优点。典型的算法有：Sting算法、Clique算法以及WaveCluster算法。

5) 基于模型方法：基于数学模型，在给定的样本和模型间建立关联的一种方法。EM算法、COBWEB算法、SOM算法和ART神经网络算法是基于模型聚类方法的代表性算法。

2.1.3 半监督聚类

半监督学习是介于无监督和监督学习之间的一种学习方法，待聚类的数据集中如果包含已标记的数据，则利用半监督学习方法可提高学习的性能。半监督学习按照功能不同可分为半监督分类和半监督聚类[11]两种。半监督分类不仅利用有标记数据进行分类器训练，同时会加入无标记数据辅助训练分类器[12]。半监督聚类利用少量先验知识(有标签的数据或数据间的约束信息)来指导聚类的过程。

先验知识即领域知识或背景知识，是基于数据集本身的约束信息，在半监督聚类方法中的先验知识形式可分为类标签和成对约束。

1) 类标签

对带有部分类标的数据集X可分为两部分X^l和X^u，$X = X^l \cup X^u$，并且$X^l \cup X^u = \varnothing$，其中$X^l = \{x_1^l, x_2^l, \ldots, x_{n_l}^l\}$表示$n_l$个含有类标的对象集合，$X^u = \{x_1^u, x_2^u, \ldots, x_{n_u}^u\}$表示$n_u$个没有类标的对象集合，一般$n_l$远小于$n_u$。

2) 成对约束关系

成对约束是半监督学习中更为普遍的一种先验知识，不同于类标签，成对约束不关心单个样本的标签，而是定义数据对象间的关联关系，即属于同一类或不属于同一类，通常Must-Link 约束和Cannot-Link约束来表示正关联约束和负关联约束。

Must-Link约束表示数据集中的两个数据对象之间的相似度很大，在聚类的结果中应该被聚集在同一个簇中，Cannot-Link 约束反之。

对于数据对象x_i和x_j，如果其所属的类分别是C_i 和C_j，则基于C_i 和C_j:

$$(x_i, x_j) \in \begin{cases} \text{Must-Link}, & if\ i = j \\ \text{Cannot-Link}, & if\ i \neq j \end{cases} \tag{2.1}$$

一般地，成对约束具有对称性以及传递性的特征:

(a) 对称性

$$\begin{aligned} (x_i, x_j) \in \text{Must-Link} &\Rightarrow (x_j, x_i) \in \text{Must-Link} \\ (x_i, x_j) \in \text{Cannot-Link} &\Rightarrow (x_j, x_i) \in \text{Cannot-Link} \end{aligned} \tag{2.2}$$

(b) 传递性

$$\begin{aligned} (x_i, x_j) \in \text{Must-Link} \wedge (x_j, x_k) \in \text{Must-Link} &\Rightarrow (x_i, x_j) \in \text{Must-Link} \\ (x_i, x_j) \in \text{Cannot-Link} \wedge (x_j, x_k) \in \text{Cannot-Link} &\Rightarrow (x_i, x_j) \in \text{Cannot-Link} \end{aligned} \tag{2.3}$$

半监督聚类是半监督学习中的一个重要组成部分，半监督聚类则是在无监督聚类的基础上，通过有标签数据(或约束关系)指导聚类过程，以提高聚类效果和质量[13]。目前，半监督聚类算法在很多实际领域中已获得广泛应用，例如：图像处理、生物信息工程、文本挖掘等[13–17]。通过对现有的半监督聚类算法进行比较和分析，可以根据使用先验信息方法的不同，将半监督聚类算法大致归为一下三类[3]:

1) 基于限制的方法：利用有监督信息来指导聚类过程来找到一个较好的数据划分。这种方法主要是通过有监督信息对聚类算法的收敛过程进行限制和约束，从而驱使它得到更好的结果。典型的算法有Seeded k-means算法和Constrained k-Means算法[18]。

2) 基于相似性度量的方法：首先训练对象间相似性度量用以满足限制信息，然后通过基于该相似性度量的聚类算法进行聚类。典型的算法有：Klein等人在Must-Link和Cannot-Link对点限制的基础上，融合了二值传递关系方法的半监督聚类算法[19]。

3) 潜藏信息共同指导的方法：探索利用存在于数据集合本身可获得的聚类先验信息，如有监督信息和无监督信息，并将这种信息与已知信息相结合。典型的算法是：王玲等人提出的基于密度敏感的半监督谱聚类[20]。

2.1.4 符号属性数据聚类算法

针对数值属性数据聚类算法的研究已取得了丰硕的成果，然而，随着符号属性数据的不断增多，针对符号属性数据聚类算法的研究得到了越来越多的关注，并取得了一定的研究成果。符号属性聚类算法的模型可分为基于类型转换的聚类模型、基于相异测度的聚类模型、基于概率统计的聚类模型以及其它类型的模型等。

1) 基于类型转换的聚类模型：将符号属性转化成连续型数值属性，然后通过现有的数值属性数据聚类算法对其进行聚类。如Ralambondrainy提出的Conceptual k-Means 聚类算法[21]，它将每个符号属性值变换到一个二值属性，然后利用k-Means聚类算法对其聚类。

2) 基于概率统计的聚类模型：针对符号属性的取值有限的特点，用概率统计来对其进行建模，将类原型定义为概率分布的形式，且对象与类间的相似性也用概率来表示，相应的聚类算法根据对象与类间的隶属概率来实现不同类的划分。通常这类聚类算法要用到概率统计中的Bayesian定理和极大似然估计法。COBWEB[22]，ECOBWEB[23]和COP-COBWEB算法[24] 等。

3) 基于相异测度的聚类模型：参照数值属性数据聚类算法的设计，重新定义适合于符号属性数据的相异测度，用它来代替距离测度，设计出类似于数值属性数据聚类算法。其中最具代表性的算法是由Huang在1997年提出的k-Modes聚类算法[25]。它采用简单匹配差异法来计算符号属性数据之间的差异程度，并用Modes 代替k-Means算法中的均值，该算法将在以后的章节中详细介绍。另一种典型的算法是ROCK聚类算法[26]，它根据相似度闭值和共享近邻来构建一个凝聚的层次聚类算法，得到样本集的聚类后在对整个数据集进行聚类。

4) 除了以上的聚类模型外，有些学者还将熵的概念引入进了符号属性数据聚类算法中。如Daniel从熵的角度出发，认为类包含相似对象越多熵越小，并将这一观点应用于符号属性数据聚类，提出了COOLCAT算法[27]。

2.2 属性选择算法

属性选择问题是模式分类、数据挖掘、图像处理等许多不同领域的重点问题[28]。近几年来，机器学习方法在实际应用中不断增长的重要性，使得属性选择问题成为十分热门的

研究课题，广泛应用于模式识别、统计学、机器学习等领域，已经有很多国内外研究人员提出了独特的思想和解决方案[29–33]。

属性选择是指在初始的M个属性中选择出一个有$m(m¡M)$个属性的属性子集，这m个属性可以像原来的M个属性一样用来正确区分数据集中的每个数据对象。随着域维度的增大，属性的数量也在不断增大。发现最优属性子集通常是难以实现的[30]，许多与属性选择相关的问题都已被证明是NP-Hard问题[34]。

现有的属性选择算法大致可以分为三类：过滤(Filter)模型、封装(Wrapper)模型和混合(Hybrid) 模型。

1) 过滤模型：定义属性间的度量来滤除无关属性和冗余属性，属性选择过程是独立于学习算法。这种模型的优点是耗时少，灵活性好。

2) 封装模型：使用分类正确率作为评价函数[30]。这种方法能够得到较高的分类预测精度，但由于在属性选择过程中必须使用学习算法对每一个搜索到的属性子集进行学习，需要耗费大量的时间，属性选择的效果依赖于分类器的好坏。

3) 混合模型：结合过滤模型和封装模型，首先采用过滤模型选择出多个候选属性子集，之后用封装模型在候选属性子集中选择最优属性子集。这种方法的时间复杂度、分类预测精度和适用性都介于过滤模型与封装模型之间。

从属性选择的实现方面来讲，属性选择是由属性评价方法(Evaluation)和属性子集的搜索策略(Search Strategy)两部分构成的。

2.2.1 属性评价方法

属性评价方法主要包括：距离度量、信息度量、依赖性度量等。

1) 距离度量:又称为可分离性、分歧法、区别度量法。对于两个类别的问题而言，选择属性X而不是选择属性Y，当使用X产生的两个类别的距离要比使用Y产生的距离大。如果差别是0，那么成X和Y是不可区分的。

2) 依赖性度量:依赖性度量或相关性度量限定了依据一个变量的值来预测另一个值的能力。例如粗糙集里面的依赖度。

3) 信息度量:主要是值通过信息论的方法计算一个属性的信息增益。

2.2.2 属性子集的搜索策略

属性子集的搜索策略主要可以分为：启发式搜索，穷尽式搜索和随机式搜索等。

1) 启发式搜索：贪心的在每次循环中，所有剩余的属性都会被考虑用来选择(或丢弃)。这个简单的过程有很多种变化的方法，但产生的属性子集基本上是增加的(增加属性或减少属性)。搜索空间是$O(2^N)$或更少。

2) 穷尽式搜索：通过计算评估函数对属性子集空间进行穷尽式搜索，搜索最优的属性子集。穷尽式的搜索就是完全的搜索。为了尽可能在不减少找到最优子集的概率的情况下，使用不同的启发函数被用来减少搜索所需要的时间。这样，尽管搜索属性空间仍旧是$O(2^N)$，但是评估的子集个数减少了。依赖于评估函数的最优属性子集能够被保证是因为在这个过程使用了回溯策略。回溯策略有很多种实现的方式，如:分店模式(Branch and Bound)，最优搜索(Best First Search)，宽度搜索(Beam Search)。

3) 随机式搜索：通过设置一个可能大的搜索次数。搜索到的属性子集的最优程度依赖于可用的资源。在属性子空间中随机的选择，并通过一定的评价函数判断是否停止。如基于遗传算法的属性选择等。

2.2.3 半监督属性选择算法

半监督属性选择方面的研究目前还比较少，主要从两个方面考虑:

1) 有标签数据和无标签数据结合

Wu和Li[35]利用直推式支持向量机[36](TSVM，Transductive Support Vector Machine)方法进行属性选择，TSVM 利用了无标签数据的信息，通过逐步剔除权重小的特征达到属性选择的目的Zhao和Liu[37] 提出了一种基于谱分析的半监督特征选择方法，通过评价聚类可分性和一致性对所有属性进行评分。Yubo等人[38] 在目标函数中添加未标记样本信息以实现半监督属性选择。Yaslan等人[39] 利用选取两个相关性属性子集对Co-Training算法[40]进行改进。

2) 成对约束与无标签数据结合Zhang等人[41]提出了一种用于属性选择的新半监督评价方法，CS方法，该方法通过利用成对约束集来引导属性选择，并给出了一个约束分数(CS,Constraint Score)。Sun和Zhang提出了BCS(Bagging Constrained Score)方法[42]，将Bagging方法引入其中，利用多个约束集来代替原有单一约束集的想法，并通过投票表决的方式学习属性选择空间。

2.3 小结

本章主要介绍了聚类和属性选择的基本概念，介绍了聚类算法划分方法和半监督聚类的研究现状，然后又讲了符号属性数据的相关聚类算法。接着对属性选择进行了介绍，并对属性选择两方面问题：属性评价和搜索策略进行了阐述。最后对半监督属性选择算法的研究成果做了总结。可以看出，半监督聚类和属性选择在符号属性数据的方法相对较少，有些算法并不适合符号属性数据，对此，本文提出了几种解决方案，并进行了实验验证。

第3章 基于权值投票的半监督聚类集成

在过去的十几年中，集成学习方法已经逐渐成为机器学习中最热门的研究领域之一。该方法是训练多个分类器然后将它们组合起来对一个新的对象进行预测[43]。文献[44]阐述了集成学习的优点，由于训练集的规模有限，不能精确得到假设模型或模型，导致模型产生结果的准确率不一致，集成学习即是通过把多种学习方法融合来规避单一算法的不足。因为一个集成模型的分类效果和归纳能力大多数情况要比单个分类器的好[45]，所以集成学习方法被广泛应用于多个领域，其中具有代表性的算法有Bagging[46]，AdaBoost[47]，随机森林[48]等。

2002年, Strehl等人提出“聚类集成”(Cluster Ensembles)的概念，其定义是指关于一个数据集的多个划分(Partitions)组合成为一个统一聚类结果的方法[49]。这多个划分也称作聚类集合，每个划分称为聚类成员。2007年, Gionis 等人从另一个角度给出了一种描述: 给定一个聚类集合,聚类集成(Clustering Aggregation)的目标就是要寻找一个聚类使其所有的聚类成员尽可能一致[50]。因此我们可以总结为, 聚类集成是利用多个聚类结果组合起来成为一个统一的聚类结果, 这个结果在最大程度上与聚类成员的结果相符合或一致[51]。聚类集成法相对于单个聚类算法有如下优势[52]:

1) 鲁棒性：在不同的数据集上，聚类集成的均衡性能更好。

2) 适用性：可以获得单个聚类算法无法获得的结果。

3) 稳定性：由于聚类集成的平衡规则，噪声数据对最终的聚类结果影响不大。

4) 并行和可扩展性：多个聚类方法可并行进行，然后对聚类结果进行集成。

这里需要指出的是，由于训练数据集无类别标签，因此聚类集成比传统的有监督集成要困难的多[53]。下一节，我们对聚类集成做一个简单的概述。

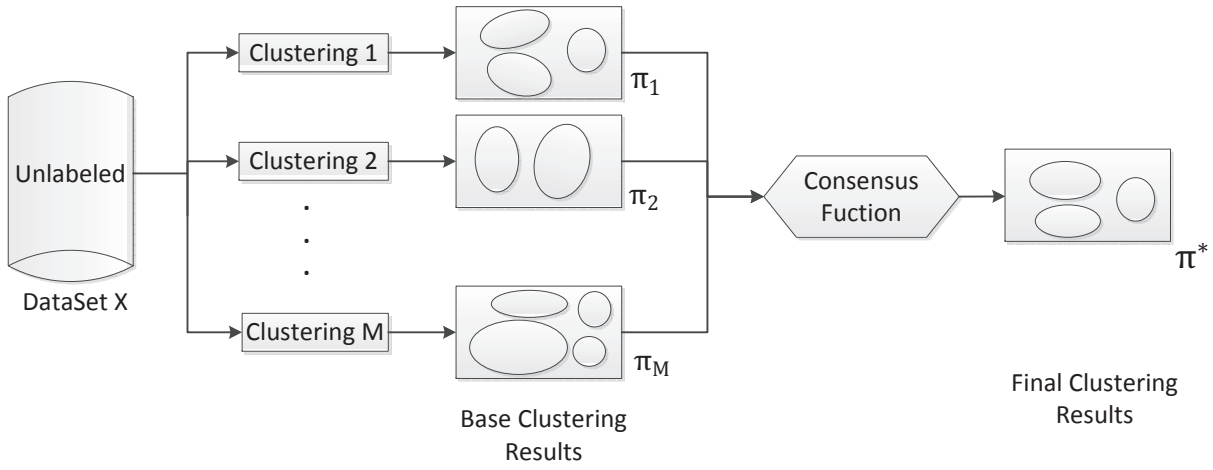


图 3.1 无监督聚类集成过程示意图

3.1 聚类集成

在Strehl[49]等提出的最初的聚类集成的定义中，假定数据集的个数为N，表示为$\mathbf{X} = \{x_1, x_2, \ldots, x_N\}$，在数据集$\mathbf{X}$上执行$M$次有差异的聚类算法，并得到$M$个聚类结果（聚类成员）$\Pi = \{\pi_1, \pi_2, \ldots, \pi_M\}$，每个聚类成员$\pi_j, (j = 1, 2, \ldots, M)$表示第$j$个聚类的划分结果的标签集。$x_i \rightarrow \{\pi_1(x_i), \pi_2(x_i), \ldots, \pi_M(x_i)\}$表示第$i$个对象对应所有聚类成员中的划分结果，$\pi_j(x_i)$是第$i$个对象在$\pi_j$中的类标签。最后通过设计一个一致性函数对$M$个聚类结果进行集成，得到一个最终的聚类结果$\pi^*$，整个聚类集成过程如图 3.1。

与监督集成学习类似，目前聚类集成算法要解决的主要问题有两个[54]:

1) 集簇的多样性(Diversity): 如何产生差异性的聚类成员从而形成一个聚类集合。

2) 一致性函数(Consensus Function): 如何组合聚类集合成为一个统一聚类结果。

在监督集成学习中保证多样性是集成学习方法好坏的关键因素，因此，聚类成员的差异性与多样性也是决定最终集成效果的重要特征；一致性函数的作用是最大程度地利用各个聚类结果的信息，并融合所有聚类结果结合成为一个统一的聚类结果。现阶段国内外研究的热点还放在第二个问题上，也就是如何从聚类集合中得到一个统一的聚类结果，所谓的一致性函数的研究上。

3.1.1 聚类成员的产生方法

在文献[43]中，Thomas G.等人对聚类成员差异性和多样性的影响问题进行了研究，并指出较大差异的聚类成员可以组合得到聚类效果。而在文献[55]中，Kuncheva, L.I 通过实

验发现只有适量的差异才能获得较好的集成结果，即差异性和聚类结果质量并非单调递增的关系，差异性越大不一定能得到越好的集成结果，差异性过大反而会降低聚类集成的质量。现有产生基聚类成员的方法如下:

1) 不同算法：使用不同的聚类算法对数据集进行聚类划分，从而得到多个不同的聚类成员，如文献[56]。

2) 相同算法不同参数：对同一个聚类算法，通过调整算法的初始化参数，即对参数赋予不同的初始值，得到有差异性的基聚类成员。例如k-Means算法初始点的选取对结果影响很大，通过选取初始点的不同就可以产生不同的聚类结果[53;57]。

3) 不同的对象子集：针对原始数据集，采用确定性、重抽样等采样方法从对象空间得到一组对象子集，然后基于对象子集再聚类从而获得有差异的聚类成员，如文献[58]。

4) 不同的特征子集：通过对特征空间选取待聚类样本的不同特征，然后再特征子集上进行聚类得到不同的聚类成员，如文献[54;59]。

5) 映射子空间：通过对数据变换空间，对投影到新的子空间的数据集进行聚类从而得到有差异的聚类成员，常用的特征映射方法如随机投影、PCA 等[54]。

6) 添加噪音：在待聚类样本中人为地添加部分噪声数据从而达到生成有差异的聚类成员[60]。

3.1.2 一致性函数

一致性函数是一个函数或方法，它将聚类成员进行集成，并最终得到一个统一的聚类结果[51;60]。目前存在许多一致性函数，如投票法，共联矩阵，超图法，互信息法，混合模型法和其他方法等。

1) 投票法

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，如文献[49;53;61]。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。

2) 共联矩阵(Co-association Matrix)

该方法将得到的聚类成员构造成$N \times N$的共联矩阵，该共联矩阵的意义在于统计对象之间在所有聚类划分中处于同一类的频率，共联矩阵可以被作为度量数据间相似性的一个矩阵，每个元素代表对象间的相似度，如文献[62;63]

基于互联合矩阵的一致性函数的缺点是它的计算和存储复杂性是二次的，所以不太适用规模较大的数据。

3) 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[49]。

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；但这几种方法都有一定的局限性，因为它们都基于图形划分算法，所以结果也受所采用的图形划分算法的影响。比如METIS和HMETIS算法要求用户输入类的个数，然后它们趋向于将图形划分成相似大小的部分，也就是说它们趋向于发现具有相似大小的类[60]。

4) 互信息法

信息论中的互信息是度量两个事件关联性的一种方式。在聚类集成中，互信息是一种可测量不同数据分布统计信息的参考指标。假设π^a和π^b是两个聚类成员，k^a 和k^b 分别表示π^a和π^b 的簇的数量，n_i表示π^a 中归属于类C_i^a 的对象数量，n_j表示π^b中归属于类C_j^b的对象数量，n_ij 表示同时属于π^a中类C_i^a和π^b 中类C_j^b 的对象数量，n 表示总共的对象个数。则范围是$[0,1]$ 的归一化互信息定义为

$$\mathbf{\Phi}^{NMI}(\pi^a,\pi^b)=\frac{\sum_{i=1}^{k^a}\sum_{j=1}^{k^b}\log\left(\frac{n\cdot n_{ij}}{n_i n_j}\right)}{\sqrt{\left(\sum_{i=1}^{k^a}n_i\log\frac{n_i}{n}\right)\left(\sum_{j=1}^{k^b}n_j\log\frac{n_j}{n}\right)}} \tag{3.1}$$

通过互信息法设计的共识函数的目标是寻找一个与所有聚类成员之间互信息最大的聚类划分，目标划分可定义为[49]:

$$\pi^{opt}=\underset{\pi}{\operatorname{argmax}}\sum_{m=1}^{M}\mathbf{\Phi}^{NMI}(\pi,\pi^m) \tag{3.2}$$

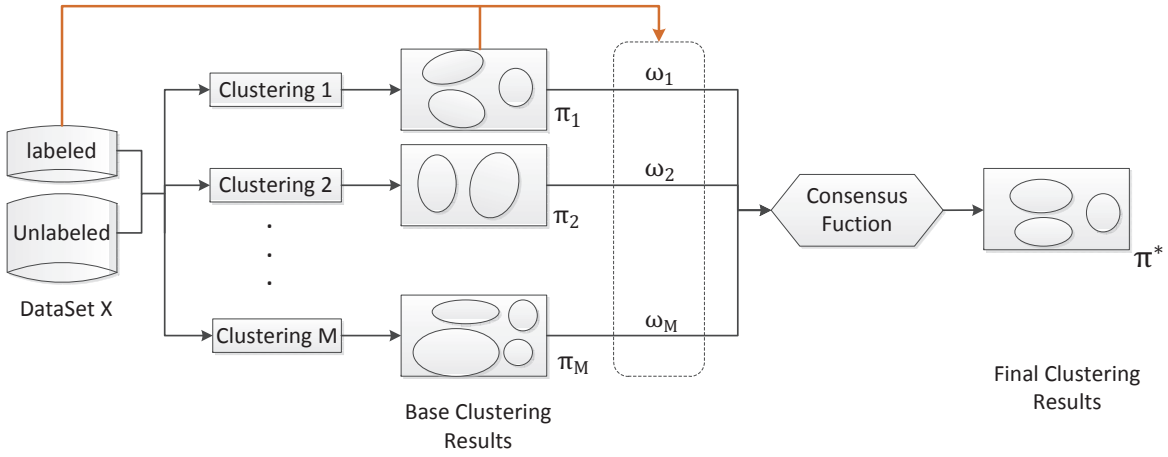


图 3.2 半监督集成聚类流程图

当目标簇的数量确定时，搜索最佳划分的问题就转换为求解经典类内的最小方差问题。

5) 混合模型法

混合模型法事先假设聚类集成的结果是一个多元多项式分布的混合模型[64]，然后使用EM算法作为后续聚类得到该混合模型的极大似然估计，从而得到集成后聚类划分。该方法需要估计的参数较多，但它不需要进行诸如投票法所面临的标签匹配问题，具有处理丢失数据的能力，并且算法复杂性也比共联矩阵法低。

6) 其他

文献[60]基于信息理论，通过求解目标函数最优化来求解聚类集成问题，最后利用遗传算法求解最后的聚类划分。Yang等[65]使用ART神经网络作为共识函数，对由改进的蚁群算法产生的多个聚类结果进行最后的聚类。

3.2 基于权值投票的半监督聚类集成算法

本文提出了一种基于权值投票的符号属性数据半监督聚类集成算法，如图 3.2所示。下文将对该算法进行详细的阐述。首先，我们先对半监督聚类问题进行描述。

假设$\mathbf{X} = \mathbf{X}^L \bigcup \mathbf{X}^U = \{x_1, x_2, \ldots, x_n\}$表示一个属性个数为$m$的数据集。其中，$\mathbf{X}^L$代表有类标的数据集，对象$X_i^l \in \mathbf{X}^L$可以表示成为$X_i^l = [x_{i1}, x_{i2}, \ldots, x_{im}, d_i]$，其中的$d_i$表示$X_i^l$的类标；$\mathbf{X}^U$代表无类标的数据集对象$X_i^u \in \mathbf{X}^U$可以表示成为$X_i^u = [x_{i1}, x_{i2}, \ldots, x_{im}]$。

一个聚类器将数据集$\mathbf{X}$划分成k个集簇，并且可以表示成一个标签向量(label vector) $\pi \in \mathbf{N}^n$，指示对象x_1被分配到第π_i个集簇中，也就是集簇C_{π_i}中，其中$\pi_i \in \{1, 2, \ldots, k\}$。

一个大小为t的聚类集成框架包含t个聚类器，这t个聚类器得到的聚类结果表示为$\Pi = \{\pi_1, \pi_2, \ldots, \pi_t\}$。通过一个一致性函数$\mathbf{F}$得到一个最终的聚类结果$\pi^* = \mathbf{F}(\Pi)$。

3.2.1 生成不同的聚类结果

为了产生多个不同的聚类结果，本文采用k-Modes算法作为基本的聚类算法。1998年，Huang在文献[25]中提出符号属性数据聚类算法k-Modes，它将k-Means改变成概念型数据聚类算法。k-Modes聚类算法是通过对k-Means聚类算法的扩展，使其应用于符号属性数据聚类。它采用简单匹配方法度量同一符号属性下两个属性值之间的距离,用模式(Modes) 代替k-Means聚类算法中的均值(Means)，使用基于频率的方法来更新模式点以收敛到聚类准则函数的极值点。

与k-Means算法相同，k-Modes算法也对初始点的选择敏感，因此对k-Modes算法，我们选取不同的初始中心点来得到不同的聚类结果。

k-Modes聚类算法其字符型数据描述为：设$\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$为$n$个对象构成的非空有限集合，$\mathbf{A} = \{A_1, A_2, \ldots, A_m\}$是由$m$个符号属性构成的非空有限集合，$X_i$被符号属性集$\mathbf{A}$描述为$X_i = [x_{i,1}, x_{i,2}, \ldots, x_{i,m}]$，每个属性$A_j$的值域为$DOM(A_j) = \{a_j^{(1)}, a_j^{(2)}, \ldots, a_j^{(p_j)}\}$，其中$p_j$为属性$A_j$符号数据类别的个数。传统的k-Modes聚类算法采用简单的0-1方法作为对象与中心点之间的相异度度量。

对于符号属性的数据，定义的差异测度为

$$d(X_i, Z_l) = \sum_{j=1}^{m} \delta(x_{ij}, z_{lj}) \tag{3.3}$$

其中，

$$\delta(x_{ij}, z_{lj}) = \begin{cases} 1, & x_{ij} = z_{lj} \\ 0 & x_{ij} \neq z_{lj} \end{cases}$$

x_{ij}和z_{lj}是在第j个符号属性上的取值，m是属性的个数。

在相似性度度量函数确定的基础上，对于如何判断当前初始聚类结果的优劣，我们还需要一定的判断标准，即聚类分析中的聚类准则函数。聚类准则函数是用来将所有对象有效的并且尽可能准确的划分到相应的子类当中的函数。其重要作用体现在聚类准则函数越好，聚类结果准确度越高。k-Modes聚类算法采用的是误差平方和方法，具体是使用以下

目标函数最小作为聚类准则函数:

$$\begin{aligned} P(U,Z) &= \sum_{l=1}^{k}\sum_{i=1}^{n} u_{i,j} d(X_i, Z_l) \\ \text{Subject to} &\sum_{l=1}^{k} u_{i,l} = 1, \quad 1 \leqslant i \leqslant n \\ & u_{i,l} \in \{0,1\}, \quad 1 \leqslant i \leqslant n, 1 \leqslant l \leqslant k \end{aligned} \tag{3.4}$$

在上式中，U是$n \times k$的隶属度矩阵；n表示对象集$\mathbf{X}$中所包含元素的个数，对应于隶属度矩阵中的行向量；k 表示所划分子类的个数，对应于隶属度矩阵中的列向量；$u_{i,l} = 1$ 表示对象X_i被分配到划分子类C_l中。$Z = \{Z_1, Z_2, \ldots, Z_k\}$ 是k个对象构成的非空有限集合，表示所划分子类的类中心,其中的元素Z_l可以表示为$Z_l = [z_{l,1}, z_{l,2}, \ldots, z_{l,m}](1 \leqslant l \leqslant k)$。$d(X_i, Z_l)$表示对象$X_i$, 与类中心$Z_l$ 的新的相异度量函数。

k-Modes的最优化问题可以分下面两个最小化问题迭代完成:

1) 问题P_1：固定$Z = \widetilde{Z}$，$P(U, \widetilde{Z})$为最小值，当且仅当

$$u_{i,l} = \begin{cases} 1 & if\ d(X_i, Z_l) \leqslant d(X_i, Z_h), \forall h, 1 \leqslant h \leqslant k \\ 0 & otherwise \end{cases}$$

2) 问题P_2：固定$U = \widetilde{U}$，$P(\widetilde{U}, Z)$为最小值，当且仅当

$$z_{l,j} = a_j^{(r)},\ \text{satisfies}\ \underset{r, 1 \leqslant i \leqslant n}{\operatorname{argmax}} \left| \{u_{i,l} | x_{i,j} = a_j^{(r)}, u_{i,l} = 1\} \right|$$

其中，$a_j^{(r)}$是子类C_l在属性A_j下的mode。

k-Modes聚类算法的基本流程如算法 1所示。

算法 1 k-Modes算法流程

输入: 数据集$X = \{X_1, X_2, \ldots, X_n\}$和聚类个数$k$

输出: 聚类划分结果$C = \{C_1, C_2, \ldots, C_k\}$

1: 随机选择或指定k个对象作为类中心的初始值$Z^{(1)} = \{Z_1^{(1)}, Z_2^{(1)}, \ldots, Z_k^{(1)}\}$

2: 计算每一个数据点到这个类中心之间的相异度$d(X_i, Z_l)$，并通过所得的相异度值来计算隶属度矩阵$U^{(1)}$，令$t = 1$。

3: 使用更新聚类中心的方法确定$Z^{(t+1)}$ 使得$P(U^{(t)}, Z^{(t+1)})$最小，如果$P(U^{(t), Z^{(t+1)}}) = P(U^{(t)}, Z^{(t)})$，算法结束。

4: 使用更新隶属度矩阵的方法确定$U^{(t+1)}$使得$P(U^{(t+1)}, Z^{(t+1)})$最小，如果$P(U^{(t+1)}, Z^{(t+1)}) = P(U^{(t)}, Z^{(t)})$，算法结束。否则，令$t = t + 1$，然后转向Step3。

3.2.2 组合不同的聚类结果

3.2.2.1 聚类成员权重

我们认为，得到的不同聚类成员的重要程度是不同的，它们在最终的聚类结果所起的作用也是不同的，因此我们要给每个聚类成员加一个权重，通过权重再来投票得出最终的聚类结果。我们有监督和无监督两个方面考虑:

1) 有监督数据: 我们定义了一个一致率评价度量，也就是单单有监督这部分数据的聚类结果和这部分的类标一致的个数占总数的比例。

定义 3.1. 假设给定数据集$X = X^L \bigcup X^U$，X^L为有类标数据，X^U为无类标数据。某聚类成员π_h 的有监督部分的权重可以表示为

$$\omega_h^L = \frac{\sum_{x_i \in X^L} \sum_{x_j \in X^L} Cons(x_i, x_j)}{|X^L| \cdot |X^L| \cdot Z^L} \tag{3.5}$$

其中$|\cdot|$表示集合的个数，Z^L 是归一化因子，使得$\omega_h^L > 0$并且$\sum \omega_h^L = 1$，$Cons$为两个对象的聚类结果与类标的一致度量，表示为：

$$Cons(x_i, x_j) \begin{cases} 0 & \pi_h(i) = \pi_h(j) \oplus d(i) = d(j) \\ 1 & otherwise \end{cases} \tag{3.6}$$

$\pi_h(i) = \pi_h(j)$代表x_i和x_j 在同一个聚类结果里，$d(i) = d(j)$代表类标相同，$\oplus$ 为异或符号，该式子表示聚类结果和类标不一致结果为0，一致结果为1。

对于有类标数据，我们认为聚类结果越和类标一致，就说明这个聚类成员的聚类效果越好，就应该给它赋更大的权值。

2) 无监督数据（全体数据）: 这部分我们用公式 3.12NMI来度量某一个聚类成员与其他的聚类成员的相似性。因为X^L去掉类标也是无监督信息，因此计算时我们考虑的是全体数据集X。

定义 3.2. 假设聚类成员的个数为t，某聚类成员π_h的无监督部分的权重可以表示为：

$$\omega_h^U = \frac{\sum_{l=1, l \neq h}^{t} \Phi^{NMI}(\pi_h, \pi_l)}{(1-t) \cdot Z^U} \tag{3.7}$$

$\Phi^{NMI}(\pi_h, \pi_l)$为聚类成员π_h和π_l归一化互信息，Z^U是归一化因子，使得$\omega_h^U > 0$并且$\sum \omega_h^U = 1$

通过上面有监督和无监督两个方面得到的权重综合起来，对于某个聚类成员，我们得到了其最终的权重。

定义 3.3. 假设数据集$X = X^L \bigcup X^U$，通过生成不同的聚类结果，得到t 个聚类成员为$Pi = \{\pi_1, \pi_2, \ldots, \pi_t\}$，其中任意一个聚类成员$\pi_h$的权重定义为：

$$\omega_h = \frac{\alpha\omega_h^L + (1-\alpha)\omega_h^U}{2} \tag{3.8}$$

这里的α控制有监督信息与无监督信息在最终聚类结果里起作用的比例。

3.2.2.2 对齐步骤

基于投票的聚类集成算法在结合之前都要有对齐这一步骤[53]。这是因为两个不同的聚类成员可能分配给同一集簇不同的标号，例如，有两个聚类成员，它们对应的标签向量分别为$[1,2,2,1,1,3,3]^T$和$[2,3,3,2,2,1,1]^T$，尽管每一维度上对应的标签不同，但实际上这两个聚类成员的结果是一样的，因此要首先进行对齐操作。

算法 2 聚类成员π^b对齐成员π^a的算法

输入： 两个聚类成员$\pi^a = \{C_1^a, C_2^a, \ldots, C_k^a\}$和$\pi^b = \{C_1^b, C_2^b, \ldots, C_k^b\}$。

输出： 新的聚类标签π^{b*}

1: **for** i=1 to k, j=1 to k **do**

2: OVERLAP$_{ij}$=Count(C_i^a,C_j^a)

//OVERLAP是一个$k \times k$的矩阵，Count(A,B) 计算集合A和B相交部分元素的个数

3: **end for**

4: $\pi^{b*} = \varnothing$

5: **while** $\pi^{b*} \neq \{C_1^b, C_2^b, \ldots, C_k^b\}$ **do**

6: (u,v)=argmax(OVERLAP$_{uv}$))

7: Match(C_u^a,C_v^b) //将C_u^a和C_v^b匹配，即标签改为相同

8: $\pi^{b*} = \pi^{b*} \cup \{C_v^b\}$ 并删除OVERLAP$_{u*}$ 和OVERLAP$_{*v}$

9: **end while**

聚类成员的对齐是基于两个相似的集簇含有相似的对象。例如有一个聚类集体中含有两个聚类成员π_a和π_b将数据集划分成k类，分别表示为$\{C_1^a, C_2^a, \ldots, C_k^a\}$和$\{C_1^b, C_2^b, \ldots, C_k^b\}$。对聚类成员中的每一对不同的集簇$C_i^a$和$C_j^b$中重叠的对象计数，换句话说对同时出现

在C_i^a和C_j^b中的对象计数，然后选择个数最多的集簇对设置为匹配并将它们的标签改为相同，重复直到所有集簇都匹配完毕，如算法 2。

当聚类集成个数t，也就是聚类集体中聚类成员的个数大于2时，必须要选出一个聚类成员作为基准，其他聚类成员需要向它对齐。需要指出的是在文献[53]中，基准聚类成员是随机选择的，但是我们通过实验发现，选择权重公式 3.8最大值的聚类成员作为基准，具有更好的效果，因此，本文基准聚类成员的选择可以表示为:

$$\pi^{base} = \underset{h}{\operatorname{argmax}}(\omega_h) \tag{3.9}$$

3.2.2.3 投票策略

受到文献[53]启发，我们也提出四种投票策略，假设数据集为X聚类集成的个数为t:

1) W_Voting:将数据集X分别送到t个聚类器中，得到聚类成员π^i和对应的权重ω_i，所有聚类器对每个对象作带权重的投票得到最终的聚类结果。

2) S_W_Voting:将数据集X分别送到t 个聚类器中，得到聚类成员π^i和对应的权重ω_i，权重$\omega_i \geqslant \lambda$的聚类器对每个对象作带权重的投票得到最终的聚类结果。这里的λ是一个阈值，仿照文献[53]，我们也将其设置为$1/t$。

3) Random_W_Voting:对于每个聚类器在数据集X随机选择属性，组成新的数据集X_i'，然后进行聚类，得到聚类成员π^i和对应的权重ω_i，所有聚类器对每个对象作带权重的投票得到最终的聚类结果。

4) Random_S_W_Voting:对于每个聚类器在数据集X随机选择属性，组成新的数据集X_i'，然后进行聚类，得到聚类成员π^i 和对应的权重ω_i，权重$\omega_i \geqslant \lambda$的聚类器对每个对象作带权重的投票得到最终的聚类结果。

整个算法流程如算法 3所示。先通过多个k-Modes聚类器得到多个聚类成员(聚类结果)，组成一个聚类集体，然后分别计算有标签部分权值和无标签部分权值，最后通过投票策略得到最终的聚类结果。

算法 3 基于权值投票的半监督聚类集成

输入: 数据集$X = \{X_1, X_2, \ldots, X_n\}$，权值比例系数$\alpha$，聚类集成个数$t$和聚类个数$k$

输出: 最终的聚类结果π^*

1: **for** i=1 to t **do**

2: π^i=k-Modes(X)，ω_i=Weight(X)

//k-Modes可以换成其他聚类算法，这里我们用生成初始点不同得到t个不同的聚类成员；Weight函数是利用公式 3.8计算得到权重

3: **end for**

4: 利用公式 3.9选出基准聚类成员；

5: 利用算法 2将其余聚类成员与基准聚类成员标签对齐;

6: 根据某种投票策略，通过聚类成员和对应的权重，得到最终的聚类结果π^*。

3.3 实验与分析

3.3.1 数据集

本章的测试用的数据集有18个，来自UCI(University of California Irvine)[66]上的符号属性的数据集。数据集的详细情况见表 3.1。$\#Instances$，$\#Attribute$和$\#Classes$分别表示对象个数，属性个数和类别个数。这里的数据集对象全部都是有类标的，便于后面的实验及评价。

3.3.2 评价方法

本文主要采用两种评价方法，准确率Accuracy(ACC)和归一化互信息Normalized Mutual Information(NMI)来评级聚类效果[67]。

定义 3.4. 对一个对象x_i，假设r_i和s_i分别是聚类结果标签和真实的类标，准确率ACC则可以定义为

$$ACC = \frac{\sum_i^n \delta(s_i, map(r_i))}{n} \tag{3.10}$$

其中，n是数据集对象的个数，$\delta(x, y)$的值为1，如果$x = y$，反之亦然。与公式 3.3相同。$map(r_i)$ 是一个映射函数，最佳映射方法采用的是Kuhn-Munkres算法[68]。

表 3.1 数据集信息

DataSet	*#Instances*	*#Attribute*	*#Classes*	*Abstract*
Audiology	226	69	24	Standardized version of the original audiology database
Balloons	20	4	2	Data previously used in cognitive psychology experiment
BreastCancer	286	9	2	Breast Cancer Data (Restricted Access)
CarEvaluation	1728	6	4	Derived from simple hierarchical decision model
Chess	3196	36	2	King-Rook versus King-Pawn on a7
Hayes	132	4	3	Hayes-Roth Data Set from topic: human subjects study
Lenses	24	4	3	Database for fitting contact lenses
Lympho	148	18	4	Lymphography domain data set
Monks	124	6	2	A set of artificial domains over the same attribute space
Mushroom	8124	22	2	Mushrooms described in terms of physical characteristics
Nursery	12960	8	5	Derived from a hierarchical decision model
Promoters	106	57	2	E. Coli promoter gene sequences (DNA)
Shuttle	15	6	2	Shuttle Landing Control Data Set
Soybean	302	35	3	Large version of Michalski's famous soybean disease database
SPECT	267	22	2	Single Proton Emission Computed Tomography (SPECT) images.
Trains	10	32	2	2 data formats (structured, one-instance-per-line)
Tumor	339	17	22	Primary Tumor Data Set from Ljubljana Oncology Institute
Voting	435	16	2	1984 United Stated Congressional Voting Records

定义 3.5. 假设C表示数据集真实的划分，C'是通过聚类算法得到的聚类结果，则它们的mutual information(MI)定义为：

$$MI(C,C') = \sum_{c_i \in C, c'_j \in C'} p(c_i, c'_j) \log_2 \frac{p(c_i, c'_j)}{p(c_i) \cdot p(c'_j)} \tag{3.11}$$

其中，$p(c_i)$和$p(c'_j)$分别表示一个对象分别被c_i和c'_j选中的概率，$p(c_i, c'_j)$表示一个对象同时被c_i和c'_j选中的概率，C和C'的归一化互信息可以表示为：

$$NMI(C,C') = \frac{MI(C,C')}{max(H(C), H(C'))} \tag{3.12}$$

式子中，$H(C)$和$H(C')$分别表示C和C'的熵。不难发现$NMI \in [0,1]$。如果$NMI = 1$，说明两个划分完全相同；相反，如果$NMI = 0$，说明两个划分完全独立。

3.3.3 结果及分析

因为符号属性数据的半监督聚类相关的研究很少，所以为了验证本文提出的算法，我们将四种投票策略和单独k-Modes算法进行比较，分别计算它们的准确率和归一化互信息。

我们首先从三个方面分别对算法进行分析:

1) 权值比例系数α(Alpha): 我们假设聚类集成个数Ensemble Size(Ensize)为15，有类标数据对象个数占全部数据对象个数比例Labeled Percent(Lpercent)为15%，Alpha分别取0 到1之间，步进为0.5的11个数，分别聚类计算准确率和归一化互信息。为了降低随机性，我们对每个Alpha重复10次，计算得到的准确率和归一化互信息取平均值。

   结果如图 3.3和图 3.4所示。由图中我们可以观察到，对于准确率ACC来说，不管alpha取何值，数据集Audiology 和Lenses 四种投票策略都要优于单独的k-Modes算法，而在数据集Balance和Shuttle 上，半监督聚类集成表现的要差。但是整体而言，四种投票策略要好于单独的k-Modes算法。对于四种投票策略来说，W_Voting和S_W_Voting表现大体相同，Random_W_Voting 和Random_S_W_Voting表现趋势也是一样的，并且这四种投票策略在不同的数据集效果也有不同，有好有坏但效果大致相当。在NMI上的结果，可以得出与ACC相同的结论。

2) 聚类集成个数Ensemble Size(EnSize): 我们赋值权值比例系数Alpha为0.6，有类标数据比例Lpercent为15%，聚类集成个数EnSize分别为5,13,15,18,20,25和30。计算10次得到平均值如图 3.5和 3.6。

   通过实验结果可以看到，尽管Ensize的大小变化，聚类效果的评价指标ACC和NMI在绝大多数的数据集上得到了较好的效果，比单独k-Modes有一定的优势，特别是在Lenses，Promoters和SPECT上，优势更为明显。但也要看到，在Shuttle数据集上，效果不如k-Modes。我们还可以得出，聚类效果ACC和NMI不会随着聚类集成个数而提高，换句话说，聚类集成个数增加不会对聚类集成效果产生明显提高。相反，多数情况下会降低聚类效果。例如数据集Lenses所展示。

3) 有类标数据比例Labeled percent(Lpercent): 我们将权值比例系数Alpha赋值为0.6，聚类集成个数EnSize等于15，Lpercent分别取取0到0.5之间，步进为0.05的11个数，计算10次得到平均值如图 3.7和 3.8所示。

   实验结果表明，随着有类标数据比例Lpercent增大，聚类效果评价指标ACC和NMI大体平稳，但是会有波动。对于四种投票策略来说，在大多数的数据集上，聚类效果要优于单独k-Modes，我们还发现，有10组数据集W_Voting和S_W_Voting策略要好，而在另外8组数据集上Random_W_Voting和Random_S_W_Voting策略表现的更加出色。因此我们可以这么说不同策略适用于不同数据集。至于什么特质的数据集适用于怎样的投票策略有待于将来进一步研究。

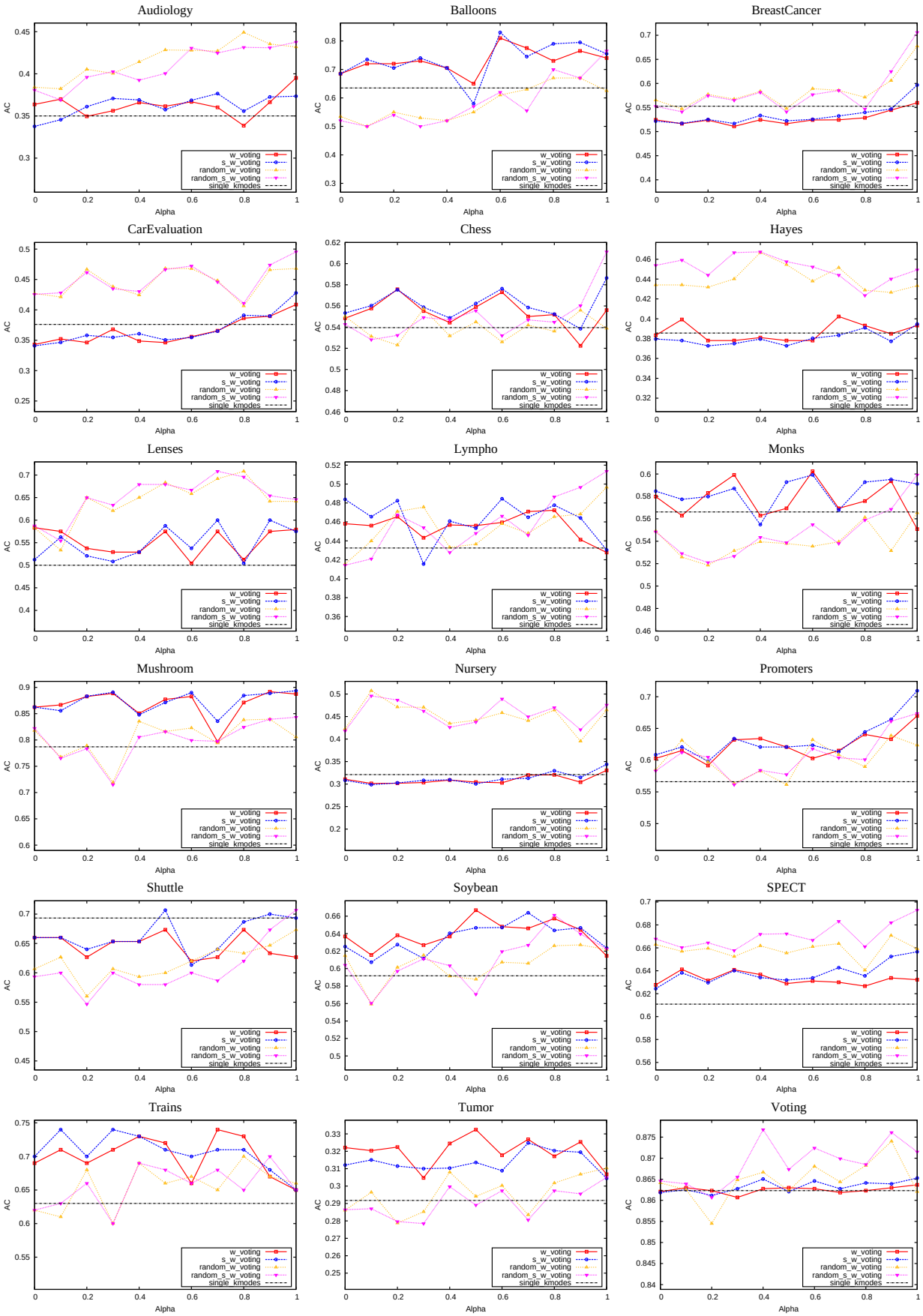


图 3.3 不同权值比例系数Alpha比较ACC

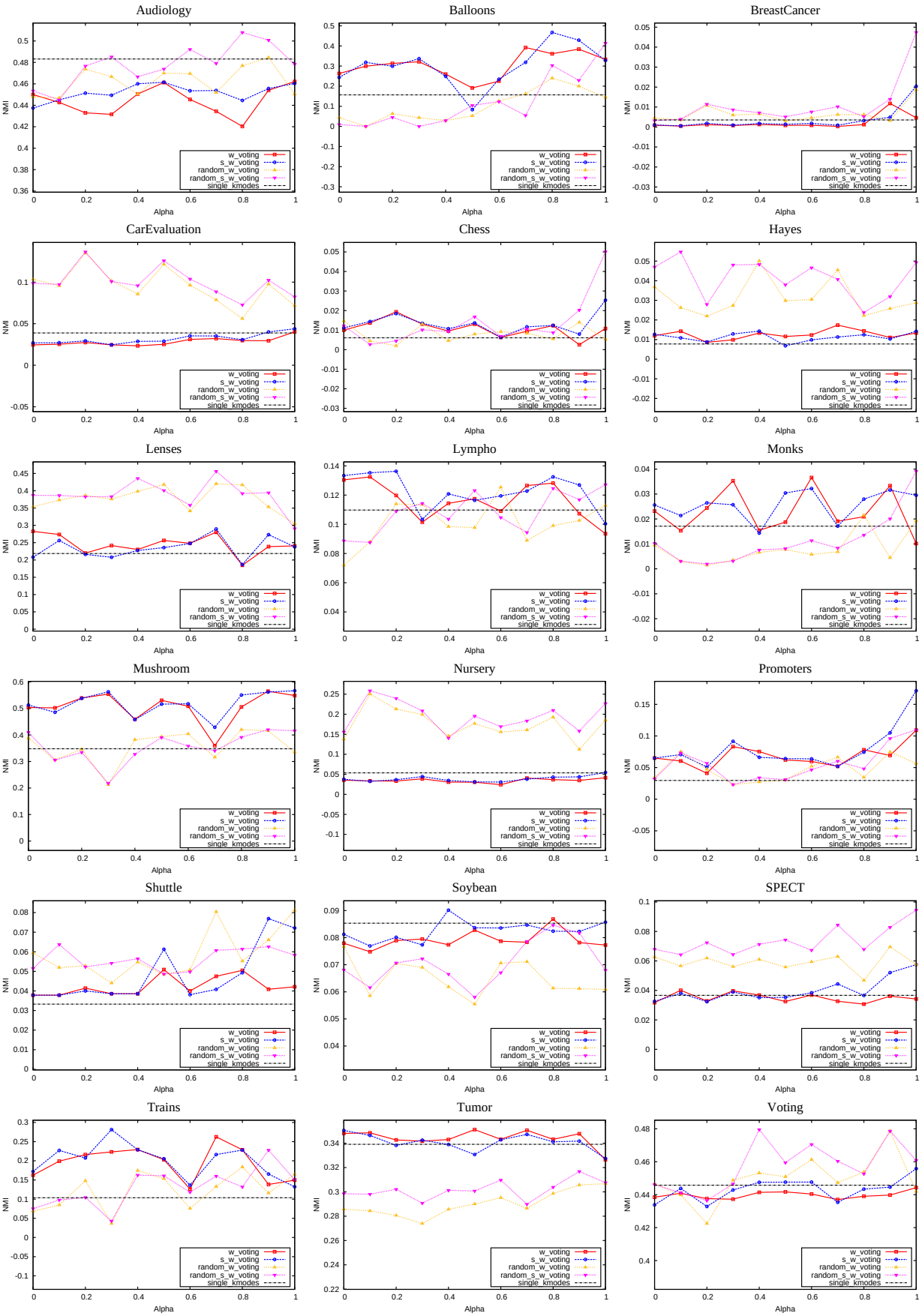


图 3.4 不同权值比例系数Alpha比较NMI

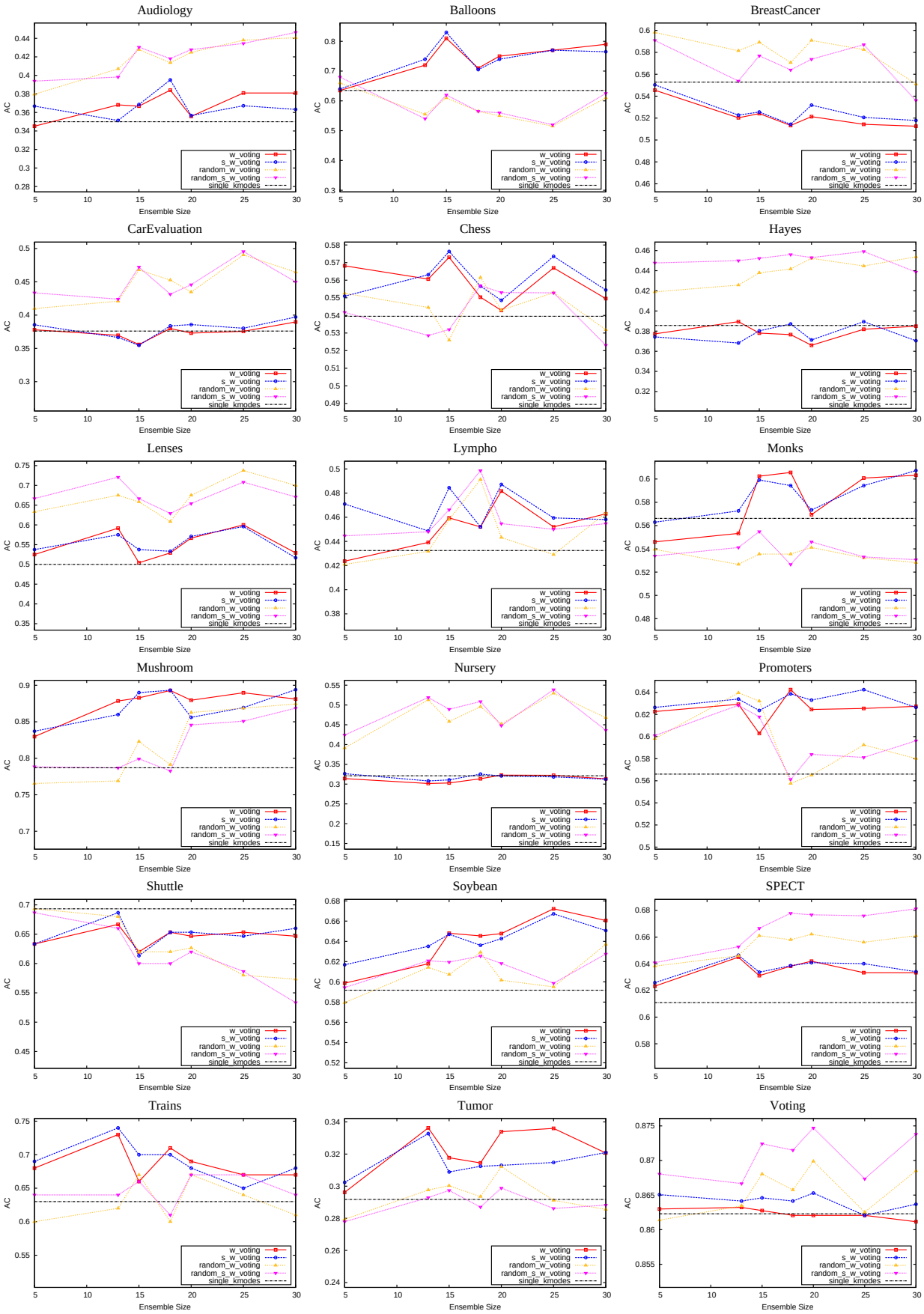


图 3.5 不同聚类集成个数EnSize比较ACC

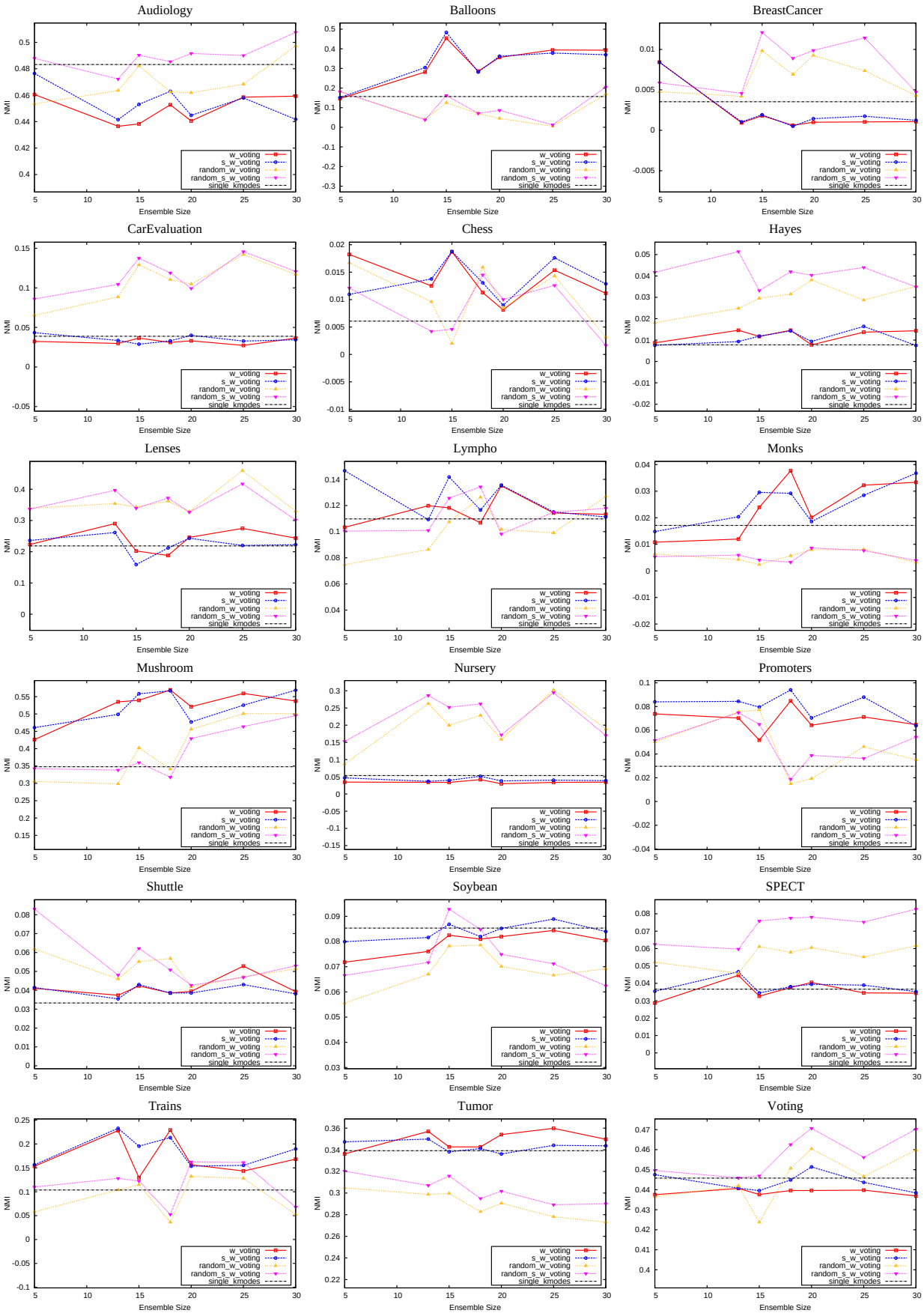


图 3.6 不同聚类集成个数EnSize比较NMI

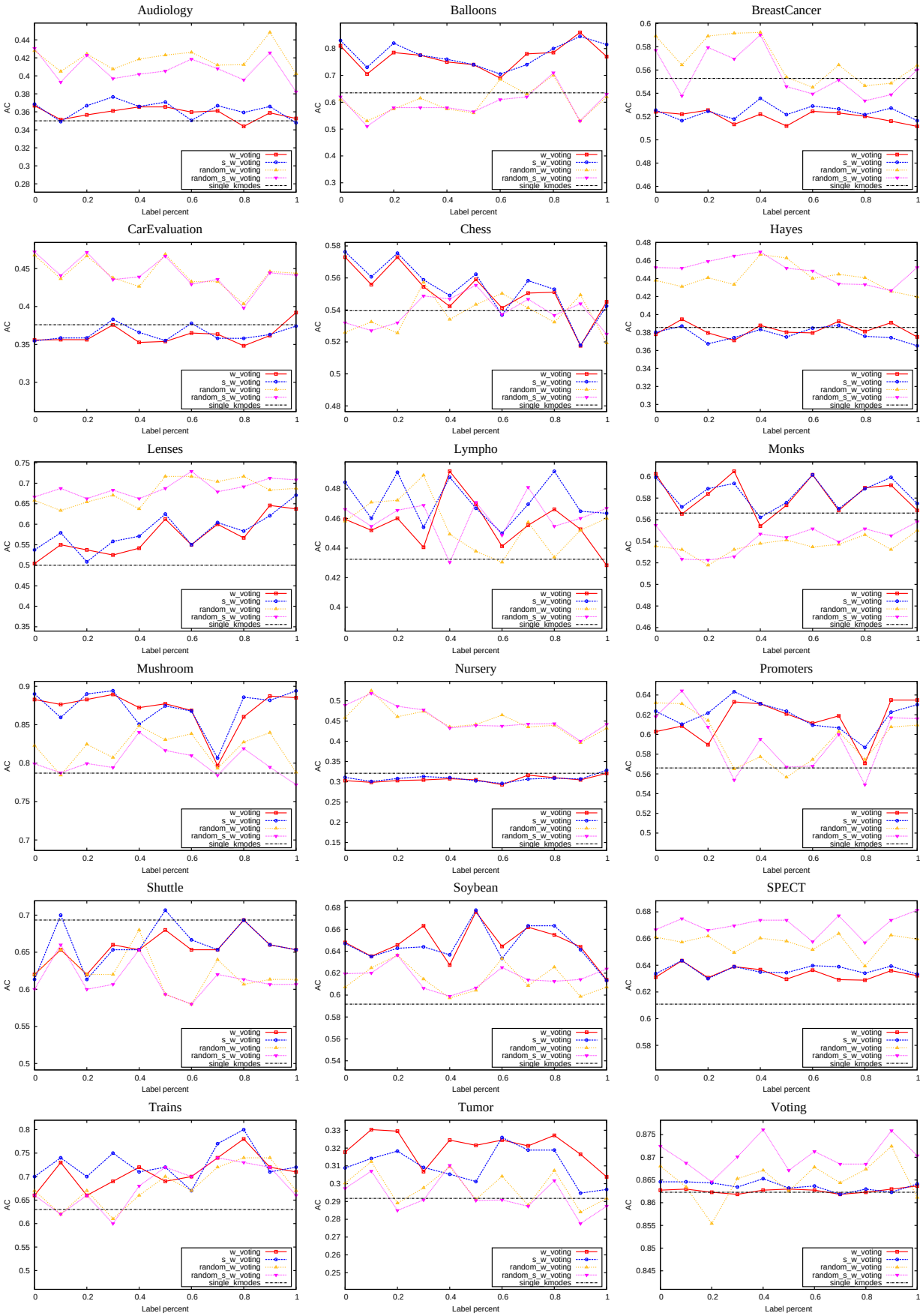


图 3.7 不同有类标数据比例Lpercent比较ACC

图 3.8 不同有类标数据比例Lpercent比较NMI

表 3.2 比较ACC:参数为Alpha=0.60 Ensize=15 Lpercent=0.15

Datasets	*w_voting*		*s_w_voting*		*random_w_voting*		*random_s_w_voting*		*single_kmodes*
	ACC±Std	ρ-Value	ACC±Std	ρ-Value	ACC±Std	ρ-Value	ACC±Std	ρ-Value	ACC±Std
Audiology	36.68±4.77	0.34-	36.86±5.29	0.32-	42.79±6.98	0.00✓	43.05±8.65	0.01✓	35.00±2.74
Balloons	81.00±6.58	0.00✓	83.00±4.83	0.00✓	61.00±15.78	0.69-	62.00±17.51	0.84-	63.50±10.29
BreastCancer	52.41±1.93	0.08-	52.55±1.54	0.06-	58.92±7.95	0.18-	57.69±8.81	0.41-	55.28±3.12
CarEvaluation	35.58±4.22	0.41-	35.46±4.28	0.33-	46.78±5.24	0.00✓	47.22±4.66	0.00✓	37.60±5.12
Chess	57.30±2.74	0.04✓	57.63±2.54	0.01✓	52.59±1.49	0.14-	53.20±2.84	0.59-	53.95±2.46
Hayes	37.80±2.07	0.24-	38.03±2.22	0.64-	43.79±5.66	0.02✓	45.23±3.75	0.00✓	38.56±2.19
Lenses	50.42±9.10	0.91-	53.75±7.47	0.31-	65.83±10.72	0.00✓	66.67±9.21	0.00✓	50.00±7.08
Lympho	45.95±3.70	0.27-	48.45±4.43	0.05-	45.81±6.08	0.18-	46.62±3.74	0.06-	43.24±5.14
Monks	60.24±4.15	0.09-	59.92±2.90	0.03✓	53.55±2.61	0.02×	55.48±2.87	0.42-	56.61±3.56
Mushroom	88.28±2.39	0.05✓	89.00±0.74	0.03✓	82.29±8.27	0.34-	79.92±8.81	0.73-	78.70±12.62
Nursery	30.29±3.56	0.06-	31.08±2.80	0.33-	45.82±14.87	0.01✓	48.92±12.85	0.00✓	32.11±3.48
Promoters	60.28±7.63	0.38-	62.36±9.83	0.25-	63.21±8.39	0.13-	61.79±8.43	0.23-	56.60±6.93
Shuttle	62.00±4.50	0.24-	61.33±4.22	0.20-	62.00±7.06	0.13-	60.00±7.70	0.09-	69.33±15.14
Soybean	64.80±4.71	0.07-	64.70±4.36	0.09-	60.73±6.72	0.67-	61.95±6.23	0.42-	59.17±6.75
SPECT	63.11±1.08	0.36-	63.37±0.91	0.29-	66.10±2.90	0.01✓	66.67±5.39	0.01✓	61.09±6.03
Trains	66.00±11.74	0.54-	70.00±11.55	0.30-	67.00±9.49	0.27-	66.00±8.43	0.56-	63.00±10.59
Tumor	31.77±2.01	0.00✓	30.88±1.81	0.01✓	30.03±2.08	0.25-	29.73±2.47	0.44-	29.17±2.07
Voting	86.28±0.19	0.85-	86.46±0.17	0.36-	86.80±1.07	0.11-	87.24±1.16	0.04✓	86.23±0.63
Average	56.12		56.94		57.50		57.74		53.84
Win/Tie/Lose	4/14/0		5/13/0		6/11/1		7/11/0		

我们分别取权值比例系数Alpha为0.6，聚类集成个数Ensize为15，有类标数据比例Lpercent为15%，计算10次得到一组准确率和一组归一化互结果，然后再和单独k-Modes算法的10次准确率和归一化互结果做置信率为0.05的T-test显著性检测，如表格 3.2和3.3 所示。其中，符号“✓”表示该策略结果有显著性差异并且优于单独k-Modes算法，符号“×”表示劣于单独k-Modes 算法，符号“–”表示两组结果没有显著性差异。

实验结果表明，本文提出的基于权值投票的半监督聚类集成的四种不同策略都要比单独的k-Modes算法有明显的优势，不管是从均值还是从T-test的结果上看。从全部的数据集来看，Random_W_Voting和Random_S_W_Voting策略比W_Voting和S_W_Voting策略稍占优势。而这两两策略之间差异不是很明显。

3.4 小结

本章提出了一种基于聚类集成思想的半监督聚类方法，该方法适用于符号属性数

据聚类。首先有用符号数据聚类算法k-Modes得到多个聚类成员组成一个聚类集体，因为k-Modes对初始点敏感，我们每次选取不同的初始点，这样就可以保证聚类成员的多样性。结合无监督信息和半监督信息，我们定义了无监督部分的权值和有监督部分的权值，以及组合这两部分权值的方式。最后又提出了四种投票策略，并采用权值投票得到最终的投票结果。在本章的最后，我们通过UCI数据集，对单个k-Modes算法进行了比较，并且也相互之间比较了四种投票策略。实验结果，在大多数的情况下，我们提出的方法明显优于单独k-Modes的效果，而四种投票策略则是针对不同的数据集有好有坏。

表 3.3 比较NMI:参数为Alpha=0.60 Ensize=15 Lpercent=0.15

Datasets	*w_voting*		*s_w_voting*		*random_w_voting*		*random_s_w_voting*		*single_kmodes*
	NMI±Std	ρ-Value	NMI±Std	ρ-Value	NMI±Std	ρ-Value	NMI±Std	ρ-Value	NMI±Std
Audiology	0.44±0.05	0.02×	0.45±0.04	0.07-	0.48±0.04	0.93-	0.49±0.06	0.71-	0.48±0.01
Balloons	0.45±0.11	0.00✓	0.48±0.08	0.00✓	0.12±0.31	0.77-	0.16±0.32	0.97-	0.16±0.16
BreastCancer	0.00±0.00	0.52-	0.00±0.00	0.54-	0.01±0.02	0.38-	0.01±0.03	0.35-	0.00±0.01
CarEvaluation	0.04±0.02	0.87-	0.03±0.02	0.41-	0.13±0.04	0.00✓	0.14±0.05	0.00✓	0.04±0.03
Chess	0.02±0.01	0.03✓	0.02±0.01	0.01✓	0.00±0.00	0.11-	0.00±0.01	0.70-	0.01±0.01
Hayes	0.01±0.01	0.25-	0.01±0.01	0.21-	0.03±0.03	0.02✓	0.03±0.01	0.00✓	0.01±0.00
Lenses	0.25±0.13	0.56-	0.25±0.13	0.57-	0.34±0.13	0.02✓	0.36±0.17	0.01✓	0.22±0.11
Lympho	0.12±0.02	0.60-	0.14±0.04	0.12-	0.11±0.03	0.88-	0.13±0.03	0.39-	0.11±0.04
Monks	0.04±0.03	0.09-	0.03±0.02	0.06-	0.01±0.01	0.04×	0.01±0.01	0.32-	0.02±0.02
Mushroom	0.54±0.07	0.05-	0.56±0.02	0.02✓	0.40±0.16	0.46-	0.36±0.16	0.87-	0.35±0.25
Nursery	0.03±0.02	0.05×	0.04±0.02	0.10-	0.20±0.18	0.03✓	0.25±0.14	0.00✓	0.05±0.02
Promoters	0.05±0.07	0.52-	0.08±0.09	0.26-	0.08±0.06	0.18-	0.07±0.07	0.32-	0.03±0.06
Shuttle	0.04±0.01	0.26-	0.04±0.00	0.22-	0.06±0.03	0.07-	0.06±0.04	0.06-	0.03±0.02
Soybean	0.08±0.01	0.42-	0.08±0.01	0.82-	0.07±0.02	0.20-	0.07±0.03	0.19-	0.09±0.02
SPECT	0.03±0.01	0.74-	0.03±0.01	0.85-	0.06±0.03	0.02✓	0.08±0.04	0.01✓	0.04±0.03
Trains	0.13±0.14	0.66-	0.20±0.16	0.26-	0.11±0.10	0.79-	0.12±0.12	0.76-	0.10±0.12
Tumor	0.34±0.01	0.48-	0.34±0.01	0.49-	0.30±0.03	0.00×	0.31±0.03	0.00×	0.34±0.02
Voting	0.44±0.01	0.55-	0.45±0.02	0.88-	0.46±0.03	0.11-	0.47±0.03	0.03✓	0.45±0.02
Average	0.17		0.18		0.16		0.17		0.14
Win/Tie/Lose	2/14/2		3/15/0		5/11/2		6/11/1		

第4章 基于分裂重组合的半监督聚类算法

本章节提出了一种基于分裂合并思想的符号数据半监督聚类算法。首先通过有监督和无监督信息将数据集的对象分裂成多于k个很小的簇，然后通过合并小簇成k 个大簇，最终得到聚类结果。

4.1 分裂重组合的半监督聚类算法

4.1.1 分裂策略

通过利用有监督和无监督信息，将对象分裂成多个小簇，换句话说，也就是得到一个划分。这里，我们引出等价关系的概念。

定义 4.1. 设R是某个集合X上的一个二元关系。若R满足以下条件：

1) 自反性：$\forall a \in X$，aRa

2) 对称性：$\forall a, b \in X$，$aRb \Rightarrow bRa$

3) 传递性：$\forall a, b, c \in X$，$aRb \wedge bRa \Rightarrow aRc$

则称R是一个定义在X上的等价关系。

假设$\mathbf{X} = \mathbf{X}^L \bigcup \mathbf{X}^U = \{x_1, x_2, \ldots, x_n\}$表示一个属性个数为$m$的数据集。其中，$\mathbf{X}^L$代表有类标的数据集，对象$X_i^l \in \mathbf{X}^L$可以表示成为$X_i^l = [x_{i1}, x_{i2}, \ldots, x_{im}, d_i]$，其中的$d_i$表示$X_i^l$的类标；$\mathbf{X}^U$代表无类标的数据集对象$X_i^u \in \mathbf{X}^U$可以表示成为$X_i^u = [x_{i1}, x_{i2}, \ldots, x_{im}, -1]$。这里要指出的是，为了计算方便，无类标的数据我们将类标都赋值为-1。

我们可以得到

$$\begin{aligned} R^U &= \{(x_i, x_j) \in X \times X | \forall m, x_{im} = x_{jm}\} \\ R^L &= \{(x_i, x_j) \in X \times X | d_i = d_j\} \end{aligned} \tag{4.1}$$

不难证明，R^U和R^L都为等价关系，其中R^U利用无监督信息得到的等价关系，而R^L是利用有监督的信息得到的二元等价关系。我们有$R = R^U \cap R^L$也是一个二元等价关系。

定义 4.2. 假设在一个集合X上定义一个等价关系R，则X中的某个元素x_i 的等价类就是在X 中等价于x_i的所有元素所形成的子集：

$$[x_i]_R = \{x_j \in X | x_i R x_j\} \tag{4.2}$$

在X中的给定等价关系R的所有等价类的集合表示为X/R并叫做X 除以R的商集，此时，X/R就是集合X在关系R下的划分。

当给定一个数据X时，先通过利用无监督信息得到一个无监督的等价关系R^U，进而得到了一个划分X/R^U；再利用所给的有类标的数据得到了等价关系R^L，将第一步得到的划分进一步细化，得到一个既有有监督信息又有有监督信息的划分$X/(R^U \cap R^L)$。通过上述方法，我们将数据集X分成一个一个小簇，也就是一个个等价类，这些等价类有的只有无监督信息，有的既有无监督信息也有有监督信息。

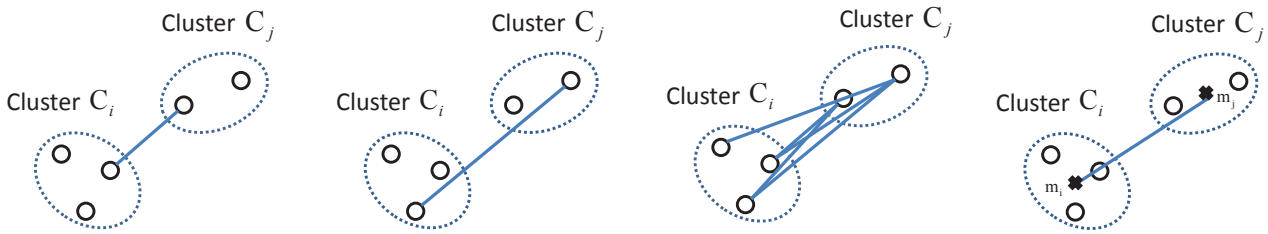
4.1.2 合并策略

在得到了一个个小簇(等价类)后，我们接下来将这些小簇合并，合并有多种方法，本文采用的是基于层次聚类的半监督合并方法。

层次聚类方法是根据给定的集簇距离度量策略，构造和维护一棵由簇和子簇形成的聚类树，直至满足某个终结条件为止。根据层次分解是自底向上还是自顶向下形成，层次聚类方法可以分为分裂方式(Divisive)和凝聚方式(Agglomerative)。

1) 分裂方式:这种自顶向下的策略首先将所有对象设置在一个簇中，然后逐步细分为越来越小的簇，直到每个对象自成一簇，或者达到了某个终结条件，如达到了某个设定的集簇数目，或两个最近的集簇之间的距离超过了某个阈值。

2) 凝聚方式:这种自底向上的策略首先将每个对象作为一个簇，然后合并这些原子簇为越来越大的簇，直到所有的对象都在一个簇中，或者某个终结条件被满足。绝大多数层次聚类方法属于这一类，它们只是在集簇距离度量策略不同。



(a) Single策略　(b) Complete策略　(c) Average策略　(d) Mean策略

图 4.1　任意两个集簇之间的距离度量策略

对于任意两个集簇之间的距离度量[69;70]，有四种主要的策略:

1) 最小距离(Single-Link):如图 4.1a 所示，是指用两个集簇中所有对象的最近距离代表两个集簇间的距离。公式为 4.3

$$d_{single}(C_i, C_j) = \min_{\forall x \in C_i, y \in C_j} ||x - y|| \tag{4.3}$$

2) 最大距离(Complete-Link):如图 4.1b 所示，是指用两个集簇中所有对象的最远距离代表两个集簇间的距离。公式为 4.4

$$d_{complete}(C_i, C_j) = \max_{\forall x \in C_i, y \in C_j} ||x - y|| \tag{4.4}$$

3) 平均距离(Average-Link):如图 4.1c 所示，是指用两个集簇中所有对象间的距离的平均距离代表两个集簇间的距离。假设n_i是集簇C_i中对象的个数，n_j是集簇C_j中对象的个数，公式为 4.5

$$d_{average}(C_i, C_j) = \frac{1}{n_i n_j} \sum_{x \in C_i} \sum_{y \in C_j} ||x - y|| \tag{4.5}$$

4) 平均值距离(Mean-Link):如图 4.1d 所示，是指用两个集簇中各自中心点之间的距离代表两个集簇间的距离。假设m_i是簇C_i的平均值，m_j是集簇C_j的平均值，公式为 4.6

$$d_{mean}(C_i, C_j) = ||m_i - m_j|| \tag{4.6}$$

可以看到，上述公式都用到了对象x与对象y的距离度量，经典的层次聚类里的聚类度量采用的欧式距离，但是欧式距离不适合符号属性数据的度量。本文研究是的是符号属性数据的聚类，因此我们要重新的两个对象之间的距离度量，受到k-Modes算法的启发，这里我们也采用简单的差异测度作为符号属性对象度量，并且，考虑到半监督的因素，我们加大了有类标数据的相异或者相似程度。

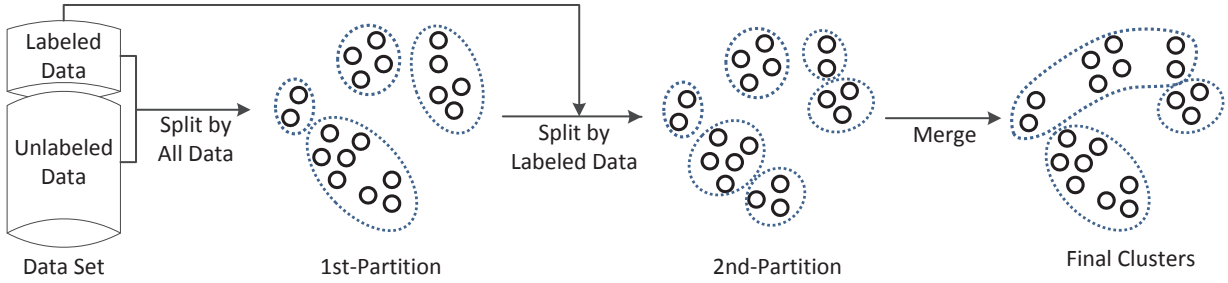


图 4.2 基于分裂组合的半监督聚类示意图

定义 4.3. 假设数据集$X = X^L \bigcup X^U \in R^m$，其中$X^L$为有类标数据，$X^U$为无类标数据，$m$为符号属性的个数，则$x_i, x_j \in X$的半监督差异测度定义为

$$d(x_i, x_j) \begin{cases} -m & x_i \in X^L \wedge x_j \in X^L \wedge d_i = d_j \\ m & x_i \in X^L \wedge x_j \in X^L \wedge d_i \neq d_j \\ \sum_{j=1}^{m} \delta(x_{ij}, z_{lj}) & otherwise \end{cases} \tag{4.7}$$

通过重新定义了两个对象的距离，也就是我们这里的差异测度，不仅可以度量两个符号属性对象，还可以将有监督的信息考虑进去。如果两个对象都是有类标的，如果类标相等则等于$-m$，这样就将两个簇之间距离“拉近”；反之，如果两个对象有类标并且类标不同，则它们的距离为最大值，这样就将所属这两个对象的两个簇“推远”。

算法 4 基于分裂重组合的半监督聚类算法

输入： 数据集$X = X^U \cup X^L$和聚类个数k

输出： 聚类划分结果$C = \{C_1, C_2, \ldots, C_k\}$

1: 计算R^U得到划分X/R^U

2: 计算R^L得到划分$X/(R^U \cap R^L)$，此时有t个集簇，记作$C' = \{C_1, C_2, \ldots, C_t\}$

3: 按照簇之间的度量策略，找到最符合策略的两个簇C_i和C_i组合，形成一个新簇并且$t = t - 1$。判断如果$t = k$，算法停止，返回剩下的簇；否则返回step3。

基于分裂重组合的符号属性数据半监督聚类算法示意图如图 4.2所示，给定数据集X，其中包括无监督数据X^U和有监督数据X^L。通过公式 4.1首先得到了等价关系R^U，进而得到了第一阶段划分X/R^U，记做1st-Partition，然后再根据有监督信息计算等价关系R^L，得到第二阶段划分2nd-Partition。最后在用层次聚类的方法合并各个划分，直到最后剩下k个簇算法停止。为了能让层次聚类适用于符号属性数据并且能利用有监督信息更好的“指

表 4.1 数据集信息

DataSet	*#Instances*	*#Attribute*	*#Classes*	*Abstract*
Audiology	226	69	24	Standardized version of the original audiology database
Balance	625	4	3	Balance scale weight & distance database
Balloons	20	4	2	Data previously used in cognitive psychology experiment
BreastCancer	286	9	2	Breast Cancer Data (Restricted Access)
Hayes	132	4	3	Hayes-Roth Data Set from topic: human subjects study
Lenses	24	4	3	Database for fitting contact lenses
Lympho	148	18	4	Lymphography domain data set
Monks	124	6	2	A set of artificial domains over the same attribute space
Shuttle	15	6	2	Shuttle Landing Control Data Set
Soybean	302	35	3	Large version of Michalski's famous soybean disease database
SPECT	267	22	2	Single Proton Emission Computed Tomography (SPECT) images.
TicTac	958	9	2	Binary classification task on tic-tac-toe game
Trains	10	32	2	2 data formats (structured, one-instance-per-line)
Tumor	339	17	22	Primary Tumor Data Set from Ljubljana Oncology Institute
Voting	435	16	2	1984 United Stated Congressional Voting Records

导”聚类，优化聚类效果，我们重新定义了距离公式，代替了传统的欧式距离，加大了有类标数据的相似或者相异程度，以影响聚类结果，具体算法如算法 4。

4.2 实验与分析

由于层次聚类的运行时间较长，所以我们去掉了上一章的数据集规模较大的几个数据，数据集的详细情况见表 4.1。为了验证本文提出的算法，我们将集簇度量策略和单独层次算法进行比较，分别计算它们的准确率ACC和归一化互信息NMI。

我们从有类标数据比例Lpercent对算法进行分析。我们将Lpercent分别取0到0.5之间，步进为0.05的11个数，计算10次得到平均值如图 4.3和 4.4 所示。

从实验结果我们可以明显的看出，本文提出的半监督聚类方法随着有类标数据比例Lpercent不断增大，相比于单独的层次聚类，聚类效果越来越好。再从四种不同的集簇距离策略来看，Average策略和Mean策略随着Lpercent增高聚类效果有明显的提高，而Single策略和Complete策略，聚类效果很不稳定，特别是数据集Shuttle和Soybean，明显的劣于其他集簇距离策略和单独的层次聚类方法。这可能是因为，在Single策略和Complete策略中，有时有类标数据所占比例很小，没有发挥优势，而在Average策略和Mean策略中，有类标数据对集簇间距离产生作用，从另一个方面讲，我们定义的对象

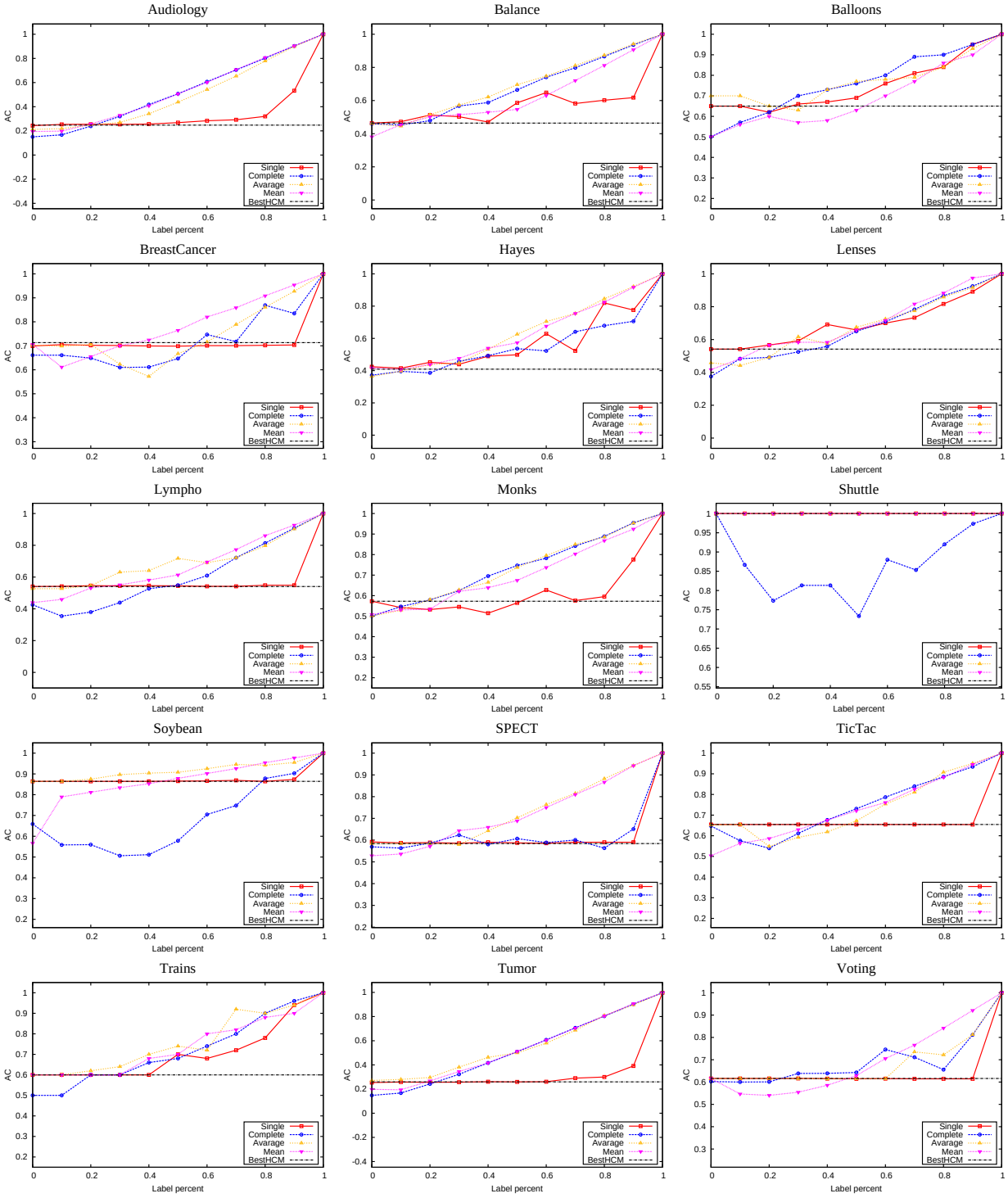


图 4.3 不同有类标数据比例Lpercent比较ACC

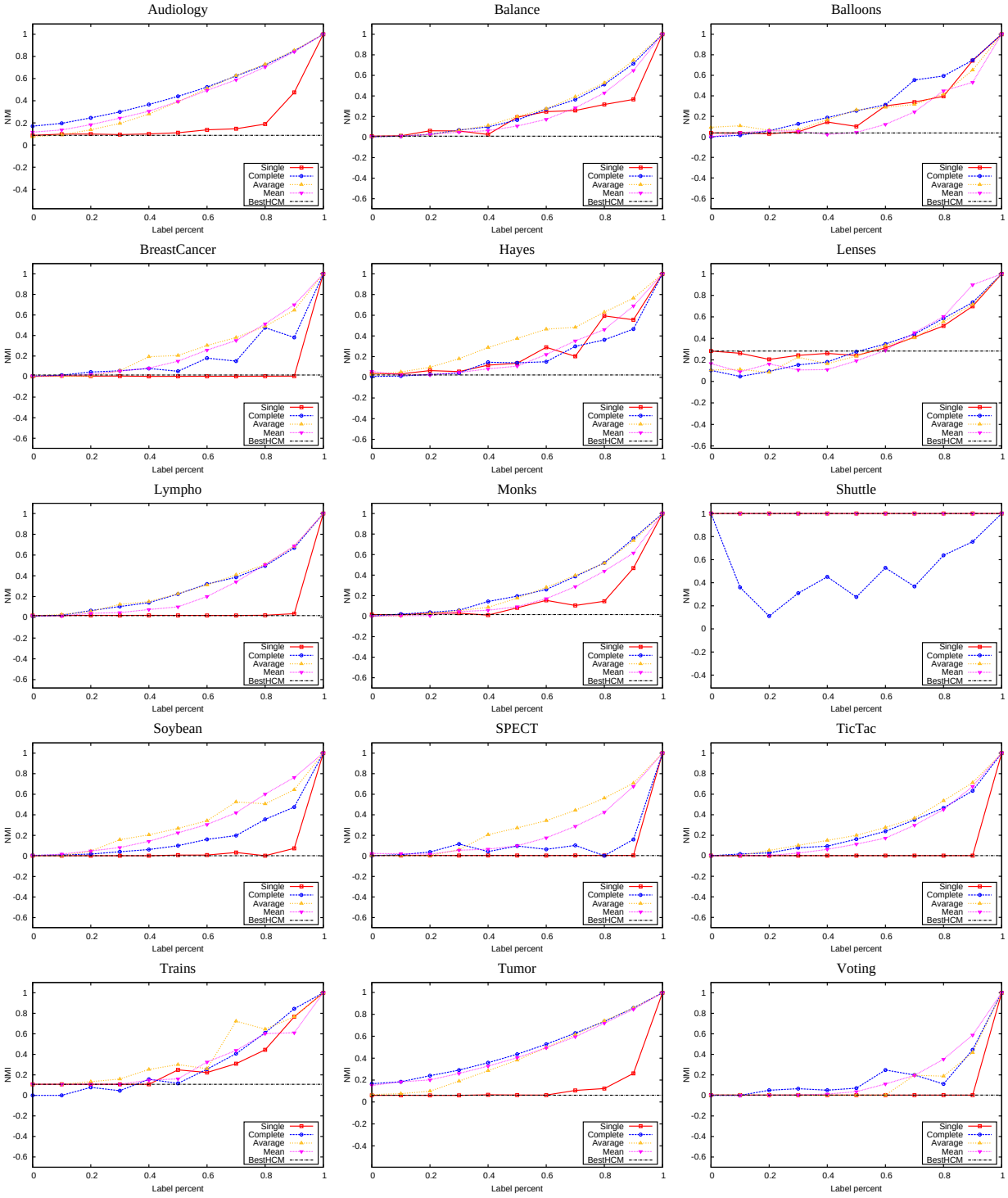


图 4.4　不同有类标数据比例Lpercent比较NMI

表 4.2　比较ACC:参数为Lpercent=0.50

Datasets	*SINGLE*		*COMPLETE*		*AVERAGE*		*MEAN*		*BestHCM*
	ACC±Std	ρ-Value	ACC±Std	ρ-Value	ACC±Std	ρ-Value	ACC±Std	ρ-Value	ACC±Std
Audiology	26.81±0.86	0.01✓	50.71±0.67	0.00✓	43.81±1.85	0.00✓	50.62±0.67	0.00✓	24.78±0.00
Balance	58.62±13.13	0.11-	66.43±1.26	0.00✓	69.60±0.99	0.00✓	54.66±0.68	0.00✓	46.40±0.00
Balloons	69.00±4.18	0.10-	76.00±8.22	0.04✓	77.00±10.37	0.06-	63.00±4.47	0.37-	65.00±0.00
BreastCancer	69.86±0.16	0.00×	64.69±8.23	0.15-	66.71±2.49	0.01×	76.43±1.12	0.00✓	71.33±0.00
Hayes	49.85±9.84	0.11-	53.64±7.47	0.02✓	62.58±1.15	0.00✓	57.27±1.38	0.00✓	40.91±0.00
Lenses	65.83±5.43	0.01✓	65.00±7.57	0.03✓	67.50±6.18	0.01✓	65.83±3.49	0.00✓	54.17±0.00
Lympho	54.46±0.37	0.07-	54.86±4.15	0.68-	71.76±3.49	0.00✓	61.35±1.46	0.00✓	54.05±0.00
Monks	56.45±12.18	0.89-	74.68±1.22	0.00✓	73.71±2.83	0.00✓	67.42±2.65	0.00✓	57.26±0.00
Shuttle	100.00±0.00	1.00-	73.33±20.55	0.04×	100.00±0.00	1.00-	100.00±0.00	1.00-	100.00±0.00
Soybean	86.56±0.30	0.37-	57.81±10.44	0.00×	90.79±1.00	0.00✓	87.88±0.30	0.00✓	86.42±0.00
SPECT	58.73±0.41	0.18-	60.67±10.93	0.67-	70.26±1.26	0.00✓	68.84±2.22	0.00✓	58.43±0.00
TicTac	65.45±0.00	1.00-	73.07±1.37	0.00✓	67.18±0.93	0.01✓	72.09±0.47	0.00✓	65.45±0.00
Trains	70.00±10.00	0.09-	68.00±8.37	0.10-	74.00±5.48	0.00✓	70.00±7.07	0.03✓	60.00±0.00
Tumor	25.90±0.57	0.83-	50.86±0.45	0.00✓	49.97±4.81	0.00✓	51.21±0.49	0.00✓	25.96±0.00
Voting	61.43±0.25	0.18-	64.28±9.65	0.57-	61.43±0.25	0.18-	62.90±0.66	0.01✓	61.61±0.00
Average	61.26		63.60		69.75		67.30		58.12
Win/Tie/Lose	2/12/1		8/5/2		11/3/1		13/2/0		

间的距离函数对聚类效果产生了有用的“指导”作用。

然后我们取有类标数据比例Lpercent为50%，计算10次得到一组准确率和一组归一化互信息，这里有类标的对象是随机选择，然后再和最优的层次聚类算法中四种不同集簇距离策略的10 次准确率和归一化互信息做置信率为0.05的T-test显著性检测，如表格 4.2和4.3所示。这里要指出的是，因为层次聚类不存在随机性，所以结果中最后一列的标准差都为0。

实验结果可以看出，对于聚类评价指标ACC和NMI，在有类标数据比例Lpercent为50%，Single策略与层次聚类的效果大致相同，Complete策略要略优于层次聚类，而Avergage 策略和Mean 策略则是比层次聚类有明显的优势。

4.3　小结

本章提出了一种基于分裂再组合的半监督聚类方法，它是通过对整个数据集先分裂再组合进行聚类，首先利用了无监督和有监督信息的等价关系，对属性集划分成一个个小的簇，然后再将这些小簇通过不同集簇间距离度量策略的层次聚类的方法组合得到最终的划

表 4.3　比较NMI:参数为Lpercent=0.50

Datasets	*SINGLE*		*COMPLETE*		*AVERAGE*		*MEAN*		*BestHCM*
	NMI±Std	ρ-Value	NMI±Std	ρ-Value	NMI±Std	ρ-Value	NMI±Std	ρ-Value	NMI±Std
Audiology	0.11±0.01	0.01✓	0.44±0.01	0.00✓	0.39±0.02	0.00✓	0.39±0.02	0.00✓	0.09±0.00
Balance	0.19±0.15	0.05-	0.17±0.02	0.00✓	0.20±0.02	0.00✓	0.11±0.01	0.00✓	0.01±0.00
Balloons	0.10±0.07	0.10-	0.25±0.16	0.04✓	0.26±0.23	0.10-	0.04±0.01	0.44-	0.04±0.00
BreastCancer	0.00±0.00	0.00×	0.05±0.09	0.43-	0.21±0.11	0.02✓	0.15±0.02	0.00✓	0.02±0.00
Hayes	0.14±0.11	0.09-	0.14±0.06	0.01✓	0.38±0.02	0.00✓	0.11±0.01	0.00✓	0.02±0.00
Lenses	0.24±0.09	0.37-	0.28±0.12	0.91-	0.24±0.12	0.44-	0.19±0.01	0.00×	0.28±0.00
Lympho	0.02±0.00	0.07-	0.22±0.05	0.00✓	0.23±0.03	0.00✓	0.10±0.02	0.00✓	0.02±0.00
Monks	0.08±0.16	0.41-	0.19±0.02	0.00✓	0.18±0.04	0.00✓	0.09±0.03	0.00✓	0.02±0.00
Shuttle	1.00±0.00	1.00-	0.28±0.41	0.02×	1.00±0.00	1.00-	1.00±0.00	1.00-	1.00±0.00
Soybean	0.01±0.01	0.37-	0.10±0.04	0.00✓	0.27±0.03	0.00✓	0.22±0.06	0.00✓	0.00±0.00
SPECT	0.00±0.00	0.18-	0.10±0.14	0.20-	0.27±0.03	0.00✓	0.10±0.02	0.00✓	0.00±0.00
TicTac	0.00±0.00	1.00-	0.16±0.02	0.00✓	0.20±0.01	0.00✓	0.11±0.01	0.00✓	0.00±0.00
Trains	0.25±0.14	0.09-	0.12±0.10	0.85-	0.30±0.09	0.01✓	0.16±0.14	0.44-	0.11±0.00
Tumor	0.06±0.00	0.25-	0.44±0.02	0.00✓	0.38±0.02	0.00✓	0.40±0.01	0.00✓	0.06±0.00
Voting	0.00±0.00	0.18-	0.07±0.16	0.39-	0.00±0.00	0.18-	0.04±0.00	0.00✓	0.00±0.00
Average	0.15		0.20		0.30		0.21		0.11
Win/Tie/Lose	1/13/1		9/5/1		11/4/0		11/3/1		

分。并比较了不同的分裂策略和组合策略。实验表明，该方法能随着带类标的比重不断加大，效果不断提升。

第5章 最小冗余最大相关半监督属性选择

无监督属性选择方法与有监督属性选择方法最主要的区别为:样本没有标记类别信息参与属性选择分析。无监督特征选择方法主要通过统计分析数据本身的信息来评价特征。而半监督特征选择则既要考虑有监督信息，又要考虑无监督信息，用这些信息共同作用参与属性选择。

本章基于mRMR算法，提出了一种最大相关最小冗余半监督属性选择，重新定义了属性的相关性和属性间的冗余，不仅考虑无监督信息，而且还单个属性对整个属性子集的作用。

5.1 最小冗余最大相关算法

半监督属性选择的问题描述为：假设数据集$X = X^L \bigcup X^U$，$A = \{a_1, a_2, \ldots, a_m\}$为$m$个属性的集合。特征选择的目的是找到一个属性子集$S$能“较好的”一致地描述属性全集$A$.

在有监督学习中，最小冗余最大相关(mRMR)属性选择算法的目的是从属性空间中寻找与目标类别有最大相关性且相互之间具有最少冗余性的m个特征[71]，最大相关和最小冗余的定义为

$$\begin{aligned} \max D(S,d), D &= \frac{1}{|S|}\sum_{a_i \in S} MI(a_i, d) \\ \min R(S), R &= \frac{1}{|S|^2}\sum_{a_i, a_j \in A} MI(a_i, a_j) \end{aligned} \tag{5.1}$$

式子中的d代表类标。该算法定义了一个算子Φ联合相关性和冗余性,并叫做最小冗余最大相关(minimal-redundancy-maximal-relevance)[72]。

$$\max \Phi(D, R), \Phi = D - R \tag{5.2}$$

实际上，贪心式的搜索策略可以用来找到Φ的近似最优解。假设我们已经有属性子集S_{m-1}，当中包含$m-1$个属性。接下来的目标就是从$\{A - S_{m-1}\}$中寻找第m个属性最大

化算子Φ。因此贪心式算法优化的约束条件为:

$$\max_{a_j \in \{A - S_{m-1}\}} \left[MI(x_j, d) - \frac{1}{m-1} \sum_{a_i \in S_{m-1}} MI(a_j, a_i) \right] \tag{5.3}$$

5.2 半监督最小冗余最大相关SemiMRMR算法

5.2.1 相关性冗余性

因为mRMR算法为有监督的属性选择方法，对于半监督数据，我们结合有监督信息和无监督信息，重新定义了属性的相关性和属性间的冗余:

定义 5.1. 假设数据集$X = X^L \bigcup X^U$，$A = \{a_1, a_2, \ldots, a_m\}$为$m$ 个属性的集合，对于有类标数据X^L，d为决策属性，属性a_i 的相关性定义为

$$D(a_i) = MI^L(a_i, d) \tag{5.4}$$

MI^L代表只拿有类标数据的计算a_i和D的互信息。对于无类标数据X^U，假设我们已经有属性子集S_m，当中包含$m-1$个属性，不在S_m子集中的属性a_i的冗余性定义为

$$R(a_i, S_m) = MI^U(a_i, S_m) + \sum_{a_j \in S_m} \frac{MI(a_i, a_j)}{|S_m|} \tag{5.5}$$

MI^L代表只拿无类标数据的计算互信息。$R(a_i, S_m)$分两部分，第一部分是代选属性a_i与已经选出来的整个属性子集的S_m的冗余性，第二部分是代选属性a_i与已经选出来的属性子集的S_m中的单个属性a_j的平均冗余性。

因此贪心式算法优化的约束条件为:

$$\max_{a_j \in \{A - S_{m-1}\}} \left[MI^L(a_j, d) - MI^U(a_j, S_m - 1) - \frac{1}{|S_{m-1}|} \sum_{a_i \in S_{m-1}} MI(a_j, a_i) \right] \tag{5.6}$$

5.2.2 停止准则

mRMR算法要指定属性选择的个数，如果对于没有先验知识的数据集很难达到最好的效果。我们提出的SemiMRMR 算法基于计算已经找到的属性子集的信息和熵条件熵，首先我们定义信息熵和条件熵。

定义 5.2. 假设S为一个属性子集，d为决策属性，则相应的信息和条件熵为：

$$\begin{aligned} H(S) &= -\sum p(s_i)\log p(s_i) \\ H(d|S) &= -\sum_i \sum_j p(d_i, s_j)\log p(d_i|s_j) \end{aligned} \tag{5.7}$$

当属性子集所携带的信息量能和全集的信息相等，或者大体相当，我们就认为该属性子集可以作为一个属性选择。因此如果当前选择的属性子集为S_m，我们定义算法的停止准则为:

$$|H^U(S_m) - H^U(A)| < \alpha \vee |H^L(d|S_m) - H^L(d|A)| < \alpha \tag{5.8}$$

式子中$\alpha \geqslant 0$是一个松弛变量，控制停止准则条件的强弱。α越小，停止准则条件越强，所得的属性选择个数越多；α越大，停止准则条件越弱，所得的属性选择数目越小，算法流程如算法 5所示。

算法 5 最小冗余最大相关半监督属性选择算法SemiMRMR

输入: 数据集$X = X^U \cup X^L$，松弛变量α

输出: 特征选择S

1: $S = \varnothing$

2: $S = S \cup \max_{a_i \in A} MI^L(a_i, d)$

3: **while** $|H^U(S) - H^U(A)| \geqslant \alpha \wedge |H^L(d|S) - H^L(d|A)| \geqslant \alpha$ **do**

4: $\quad a_i = \max_{a_i \in \{A-S\}} \left[MI^L(a_i, d) - MI^U(a_i, S) - \frac{1}{|S|}\sum_{a_j \in S} MI(a_i, a_j) \right]$

5: $\quad S = S \cup \{a_i\}$

6: **end while**

7: 输出特征选择S

5.3 实验与分析

本章有13个来自UCI上的数据集用来做测试。数据集的详细情况见表 5.1。这里需要指出的，因为UCI上全部是符号属性的高维数据有限，因此，我们选取的数据集有连续性数值属性，需要进行离散化的预处理。我们用MDL(Minimum Description Length)[73]方法，有监督的离散化连续性数值数据。

评价标准我们分别从分类和聚类两个方面入手。对于分类问题，我们把属性选择出子集和属性全集分别送到某个算法中比较分类精度。对于聚类问题，我们把属性选择出子集和属性全集分别送到某个聚类算法中比较准确率ACC和归一化互信息NMI。

表 5.1　属性选择所用数据集信息

DataSet	*#Instances*	*#Attribute*	*#Classes*	*Abstract*
Colic	368	22	2	Horse Colic Data Set; Well documented attributes.
Credit	690	15	2	This data concerns credit card applications; Good mix of attributes.
Diabetes	768	8	2	From National Institute of Diabetes and Digestive and Kidney Diseases.
Heart-cleveland	303	13	5	Heart Disease Data Set.
Heart-hungarian	294	13	5	Heart Disease Data Set.
Heart-statlog	270	13	2	Heart Disease Data Set.
Hepatitis	155	19	2	From G.Gong: CMU; Includes cost data (donated by Peter Turney).
Ionosphere	351	34	2	Classification of radar returns from the ionosphere.
Musk2	707	166	2	Musk (Version 2) Data Set.
Promoters	106	57	2	E. Coli promoter gene sequences (DNA).
SPECT	267	22	2	Single Proton Emission Computed Tomography (SPECT) images.
Voting	435	16	2	1984 United Stated Congressional Voting Records.
WDBC	569	30	2	Breast Cancer Wisconsin (Diagnostic) Data Set.

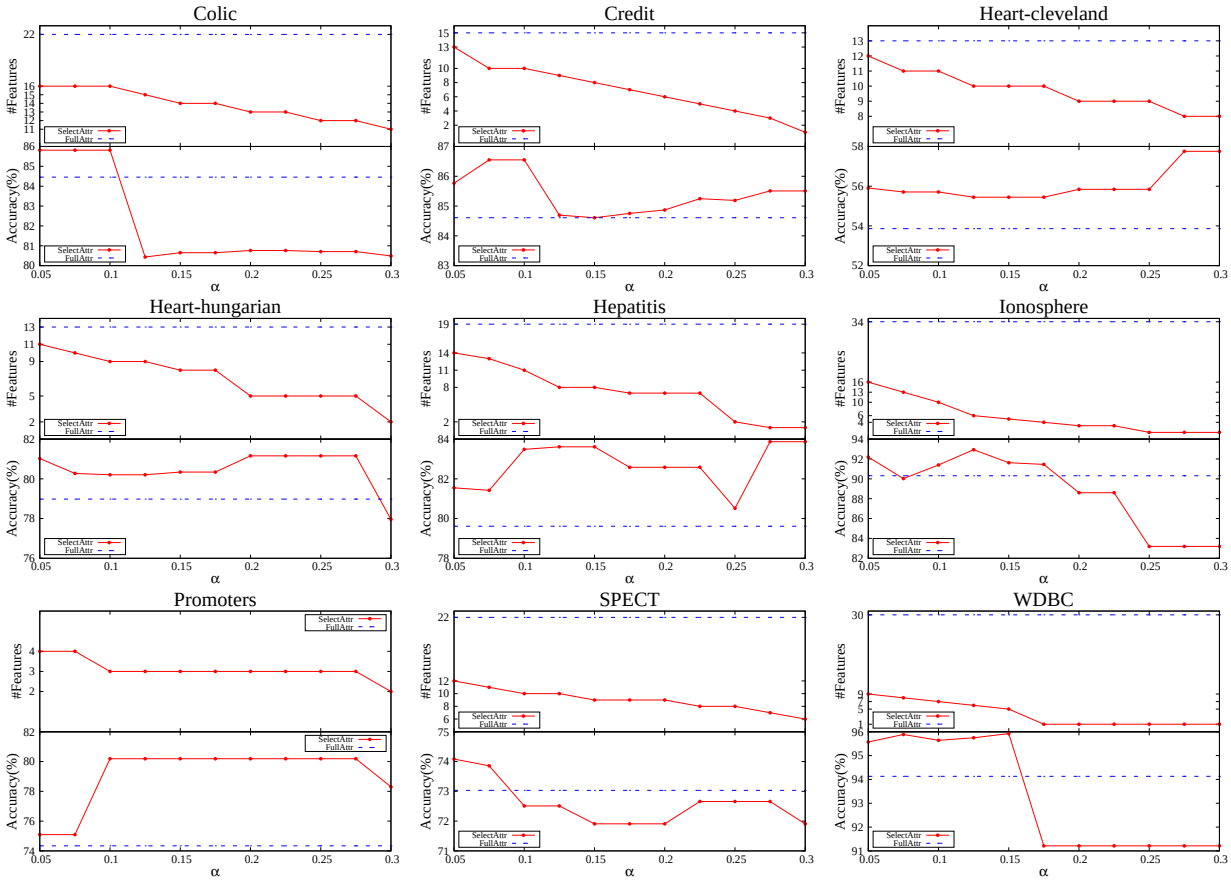


图 5.1　不同Alpha比较属性选择个数和CART的分类精度

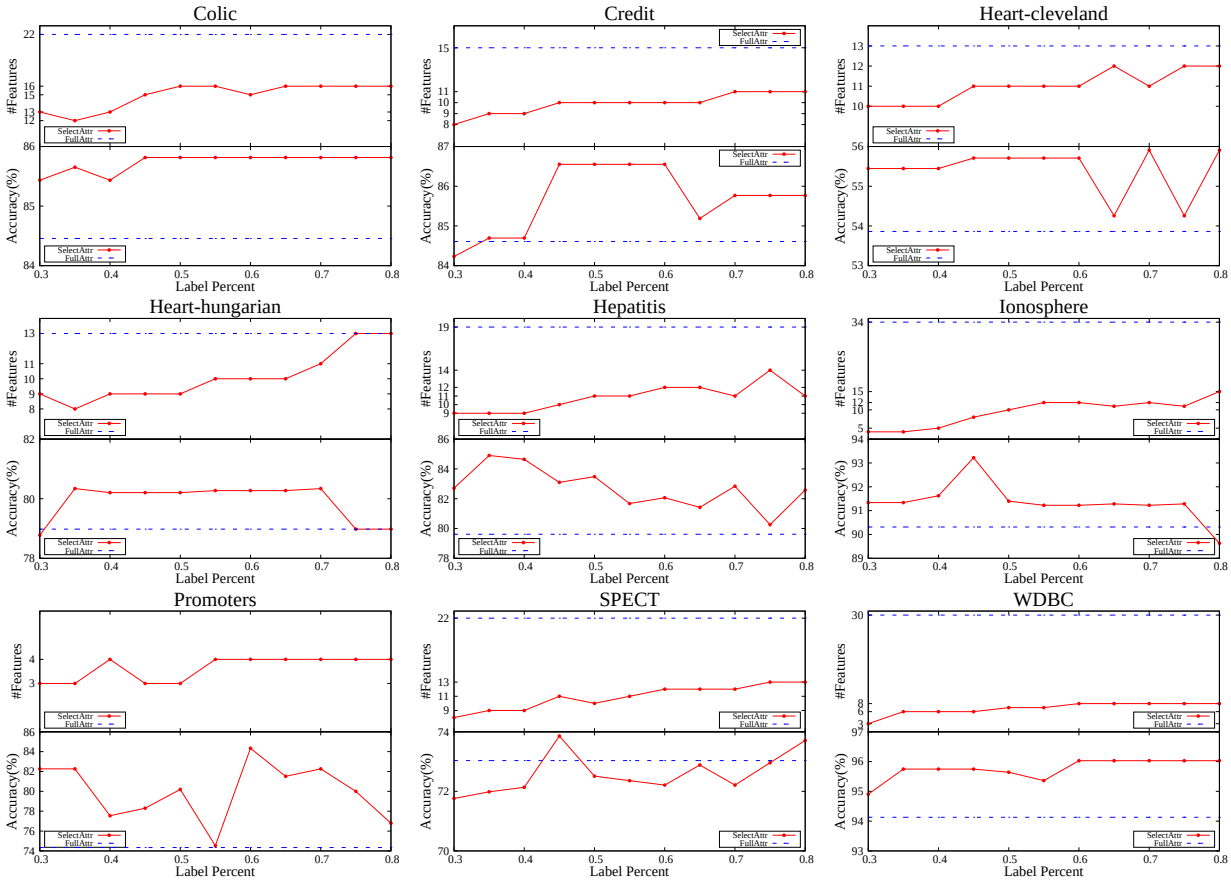


图 5.2　不同Lpercent比较属性选择个数和CART的分类精度

我们首先从松弛变量Alpha对算法进行分析。我们将Alpha分别取0.05到0.3之间，步进为0.025 的11个数，分类器我们选的是CART(Classification And Regression Tree)算法[74] 计算10次每次10折交叉验证得到平均值如图 5.1所示。图中分别显示了特征选择个数和分类精度。

我们再从有类标数据比例Lpercent对算法进行分析。我们将Lpercent分别取0到0.5之间，步进为0.05的11个数，这时Alpha取0.1，计算10次每次10折交叉验证得到平均值如图5.2所示。

当Alpha=0.1，Lercent=50%，我们提出的SemiMRMR算法所选的属性子集和属性全集分别通过不同的分类器得到分类精度，不同的聚类算法得到了准确率和归一化互信息作比较。分类器有NaiveBayes Classifier(NBC)，C4.5，JRip，PART和CART。聚类器有k-Modes，EM，Cobweb，FarthestFirst和CLOPE。

实验结果分别在表格 5.2， 5.3和 5.3中显示，其中，在表格 5.2有列出了属性选择个数。对于分类算法来说，属性选择后的数据集和全部属性的数据集在分类精度上可以保持大体相当，说明属性选择后能保持原来全集上的信息没有太大损失。从另一个方面，对于

表 5.2 算法SemiMRMR比较分类精度Accuracy参数：Alpha=0.1 Lercent=50%

Datasets	*#Features*		*NBC*		*C4.5*		*JRip*		*PART*		*CART*	
	SemiMRMR	*FullFeature*	*SemiMRMR*	*FullFeature*	*SemiMRMR*	*FullFeature*	*SemiMRMR*	*FullFeature*	*SemiMRMR*	*FullFeature*	*SemiMRMR*	*FullFeature*
			ACC±Std	ACC±Std	ACC±Std	ACC±Std	ACC±Std	ACC±Std	ACC±Std	ACC±Std	ACC±Std	ACC±Std
Colic	16	22	82.28±0.70	81.47±0.35	84.24±0.51	84.84±0.35	84.73±0.70	84.35±0.53	82.55±1.18	80.92±1.09	85.82±0.12	84.46±0.59
Credit	10	15	85.77±0.19	86.17±0.26	86.84±0.59	86.61±0.51	86.46±0.62	86.58±0.49	86.29±0.63	85.88±0.67	86.55±1.02	84.61±0.44
Diabetes	7	8	79.04±0.21	77.84±0.11	78.65±0.95	77.50±0.92	78.33±0.54	77.32±0.56	77.84±0.72	77.06±0.76	78.46±1.16	76.74±0.80
Heart-cleveland	11	13	57.62±1.03	56.30±1.11	55.78±1.49	53.00±2.42	53.07±0.72	53.93±0.50	54.79±0.52	54.46±2.07	55.71±1.33	53.86±1.18
Heart-hungarian	9	13	81.16±0.71	84.69±0.24	82.38±0.37	80.48±0.85	80.68±0.74	79.73±0.89	81.16±0.30	81.02±0.91	80.20±1.09	78.98±1.55
Heart-statlog	12	13	84.07±0.37	83.78±0.41	82.30±1.44	82.07±1.33	82.74±1.03	83.78±1.40	83.56±0.77	83.63±1.86	82.89±0.84	82.59±0.79
Hepatitis	11	19	88.13±0.58	84.00±0.54	83.35±1.90	80.26±1.91	81.81±2.16	80.00±2.28	85.03±1.54	82.32±0.87	83.48±0.98	79.61±0.35
Ionosphere	10	34	92.02±0.00	90.77±0.38	88.89±0.73	89.74±0.67	91.17±0.97	91.51±1.16	90.66±0.42	89.91±0.32	91.40±0.79	90.31±0.83
Musk2	61	166	88.20±0.40	85.26±0.27	90.69±0.18	90.83±1.00	90.21±0.62	90.01±1.05	90.69±0.47	90.69±0.46	90.16±0.28	90.27±0.50
Promoters	3	57	78.49±0.79	90.38±0.79	80.19±0.00	79.81±1.58	79.62±1.27	81.13±2.58	78.30±0.67	85.28±1.43	80.19±0.00	74.34±2.86
SPECT	10	22	69.74±0.67	68.84±0.62	71.69±0.94	69.96±1.33	72.51±0.33	72.06±1.49	66.44±2.46	68.01±2.46	72.51±0.21	73.03±0.37
Voting	3	16	95.63±0.00	90.30±0.19	95.63±0.00	96.37±0.34	95.63±0.00	95.45±0.44	95.63±0.00	95.36±0.70	95.63±0.00	95.54±0.13
WDBC	7	30	95.08±0.18	95.85±0.10	95.78±0.12	95.85±0.10	95.43±0.28	96.34±0.19	95.96±0.00	95.99±0.23	95.64±0.26	94.13±0.92
Average	13.08	32.92	82.86	82.74	82.80	82.10	82.49	82.48	82.22	82.35	82.97	81.42

表 5.3 算法SemiMRMR比较聚类精度ACC参数：Alpha=0.1 Lercent=50%

Datasets	*KModes*		*EM*		*Cobweb*		*FarthestFirst*		*CLOPE*	
	SemiMRMR	*FullFeature*	*SemiMRMR*	*FullFeature*	*SemiMRMR*	*FullFeature*	*SemiMRMR*	*FullFeature*	*SemiMRMR*	*FullFeature*
	ACC±Std	ACC±Std	ACC±Std	ACC±Std	ACC±Std	ACC±Std	ACC±Std	ACC±Std	ACC	ACC
Colic	62.23±10.85	58.10±8.19	77.45±0.00	66.85±0.00	72.61±3.87	66.25±2.75	63.26±8.97	57.12±8.89	79.08	74.18
Credit	72.55±10.48	75.45±11.76	86.67±0.00	73.33±0.00	83.86±0.91	71.28±2.49	65.22±5.66	60.03±8.59	80.58	79.71
Diabetes	60.05±7.24	60.05±7.24	66.41±0.00	66.41±0.00	65.73±1.19	65.73±1.19	57.14±8.83	57.14±8.83	76.30	67.06
Heart-cleveland	47.46±4.47	41.65±6.54	49.44±3.53	48.32±0.18	59.74±0.47	59.67±1.94	44.75±4.15	44.42±7.24	64.36	57.76
Heart-hungarian	82.31±1.42	82.31±1.42	84.08±0.15	84.08±0.15	81.70±1.69	81.70±1.69	77.35±5.61	77.35±5.61	80.27	77.21
Heart-statlog	82.67±0.41	82.67±0.41	80.37±0.00	80.74±0.00	80.89±1.25	79.41±3.24	68.89±12.17	65.63±14.34	72.22	74.07
Hepatitis	73.03±5.19	69.81±11.39	83.23±0.00	75.23±0.58	80.13±1.06	79.48±0.29	79.61±5.19	71.87±11.81	85.16	83.87
Ionosphere	87.35±1.02	88.77±0.96	90.03±0.00	89.17±0.00	89.12±1.22	79.43±2.87	85.36±3.11	78.35±2.38	95.73	95.73
Musk2	73.86±6.07	61.67±11.30	82.60±0.00	53.61±0.00	82.18±0.00	82.18±0.00	77.54±13.20	57.00±9.66	86.28	87.98
Promoters	62.08±7.53	61.32±10.05	61.89±5.88	57.17±5.99	73.96±16.70	69.43±3.10	62.64±3.10	58.11±6.21	100.00	100.00
SPECT	69.44±0.78	64.64±5.11	68.54±0.00	60.30±0.00	68.31±1.94	63.52±5.48	66.07±6.43	59.48±4.96	72.28	58.80
Voting	87.22±0.50	87.03±0.72	87.82±0.00	87.82±0.00	81.98±6.78	82.90±3.44	86.90±2.17	86.11±2.63	83.91	82.76
WDBC	94.48±0.10	92.79±0.48	95.25±0.00	94.55±0.00	92.76±1.15	83.23±2.46	84.08±13.55	88.51±8.37	97.19	89.63
Average	73.44	71.25	77.98	72.12	77.92	74.17	70.68	66.24	82.57	79.14

表 5.4 算法SemiMRMR比较聚类归一化互信息NMI参数：Alpha=0.1 Lercent=50%

Datasets	*KModes*		*EM*		*Cobweb*		*FarthestFirst*		*CLOPE*	
	SemiMRMR	*FullFeature*	*SemiMRMR*	*FullFeature*	*SemiMRMR*	*FullFeature*	*SemiMRMR*	*FullFeature*	*SemiMRMR*	*FullFeature*
	NMI±Std	NMI±Std	NMI±Std	NMI±Std	NMI±Std	NMI±Std	NMI±Std	NMI±Std	NMI	NMI
Colic	0.110±0.073	0.075±0.049	0.205±0.000	0.109±0.000	0.149±0.042	0.084±0.043	0.043±0.085	0.063±0.063	0.096	0.082
Credit	0.197±0.113	0.234±0.131	0.429±0.000	0.158±0.000	0.351±0.017	0.135±0.029	0.082±0.058	0.051±0.068	0.087	0.098
Diabetes	0.069±0.022	0.069±0.022	0.089±0.000	0.089±0.000	0.049±0.011	0.049±0.011	0.022±0.018	0.022±0.018	0.062	0.053
Heart-cleveland	0.210±0.014	0.210±0.022	0.242±0.007	0.224±0.003	0.178±0.016	0.179±0.026	0.164±0.018	0.143±0.054	0.151	0.139
Heart-hungarian	0.157±0.013	0.157±0.013	0.181±0.010	0.181±0.010	0.130±0.012	0.130±0.012	0.119±0.038	0.119±0.038	0.140	0.139
Heart-statlog	0.343±0.022	0.343±0.022	0.281±0.000	0.287±0.000	0.269±0.036	0.252±0.055	0.147±0.154	0.136±0.164	0.129	0.167
Hepatitis	0.150±0.016	0.159±0.052	0.300±0.000	0.187±0.005	0.175±0.059	0.143±0.042	0.103±0.079	0.077±0.109	0.109	0.175
Ionosphere	0.468±0.021	0.478±0.044	0.507±0.000	0.475±0.000	0.474±0.042	0.272±0.054	0.393±0.063	0.236±0.041	0.173	0.168
Musk2	0.139±0.063	0.051±0.033	0.206±0.000	0.026±0.000	0.093±0.023	0.039±0.027	0.168±0.084	0.040±0.023	0.078	0.092
Promoters	0.066±0.082	0.072±0.092	0.051±0.053	0.024±0.026	0.251±0.256	0.110±0.032	0.053±0.023	0.032±0.030	0.151	0.150
SPECT	0.105±0.013	0.066±0.043	0.104±0.000	0.028±0.000	0.079±0.023	0.050±0.046	0.073±0.046	0.016±0.015	0.064	0.068
Voting	0.474±0.016	0.465±0.034	0.486±0.000	0.486±0.000	0.375±0.098	0.376±0.068	0.460±0.047	0.439±0.074	0.320	0.276
WDBC	0.678±0.004	0.611±0.017	0.711±0.000	0.680±0.000	0.614±0.047	0.385±0.063	0.429±0.199	0.510±0.192	0.216	0.258
Average	0.244	0.230	0.292	0.227	0.245	0.169	0.174	0.145	0.137	0.143

聚类算法来说，属性选择能明显提高聚类效果，从聚类评价指标ACC和NMI来看，都要比在属性全集上做聚类有较明显的优势。

5.4 小结

本章基于mRMR算法，提出了一种最大相关最小冗余半监督属性选择，重新定义了属性的相关性和属性间的冗余，不仅考虑无监督信息，而且还单个属性对整个属性子集的作用。而并提出了一种新的停止准则来控制属性选择的个数。实验结果表明，在分类预测精度和聚类效果不降低的前提下，SemiMRMR能有效降低数据集的维度。

第6章 基于耦合依赖度的半监督属性选择

近年来，粗糙集理论（Rough Set Theory）[75–80]已经成为一种能够有效的处理复杂系统中信息和数据，它是一种处理不确定和不精确问题的新型的数学工具。粗糙集理论中的属性选择（属性约简）算法是粗糙集理论的核心内容之一。在粗糙集理论中，一个属性集的约简或者相对约简并不唯一，而求属性集的全部约简或者最小约简是一个NP-hard的问题，因此我们常常使用启发式的约简算法来获取属性约简，这里说的一个属性约简也就是属性选择。通过定义属性的一个重要度函数或差别函数，依照此函数在启发式算法的框架下来获取信息系统或者决策系统的属性选择。但是传统的重要度函数都是基于决策属性也就是类标来度量，在半监督属性约简中，还有相当部分数据不带类标，在这种情况下，粗糙集属性选择方法效果就会下降。本章拓展了传统的依赖度，不仅可以度量对决策属性（类别）的相关程度，而且可以度量属性之间的冗余程度。

6.1 粗糙集概述

在粗糙集理论中，Pawlak使用知识表达系统，主要有两种类型：一类是信息系统（信息表）；一类是决策系统（决策表）。知识表达系统被看成是一个关系数据表，关系表的行对应要研究的对象，关系表的列对应对象的属性，对象信息通过指定对象的各属性值来表达。定义 6.1给出了它的形式化定义[81]。

定义 6.1. 一个信息系统（信息表）可被定义为一个四元组$IS =< U, A, V, f >$，其中，

U：表示对象的非空有限集合，称为论域；

A：表示属性的非空集合；

V：表示全体属性的值域，$V = \bigcup_{a \in A} V_a$，$V_a$表示属性$a \in A$的值域；

f：表示$U \times A \to V$的一个映射，称为信息函数。$\forall x \in U$，$a \in A$，$f(a, x) \in V_a$表示对象x在属性a上的一个取值。

定义 6.2. 一个决策系统（决策表）可被定义为一个四元组$DS =< U, A \cup D, V, f >$，其中，

U：表示对象的非空有限集合，称为论域；

A：表示条件属性的非空集合；D表示决策属性的非空集合，$C \cap D = \varnothing$；

V：表示全体条件属性和决策属性的值域，$V = \{V_a | a \in A\} \cup \{V_d | d \in D\}$；

f：表示$U \times A \to V$的一个映射，称为信息函数.$\forall x \in U$，$a \in A \cup D$，$f(a, x) \in V_a$表示对象x在属性a上的一个取值。

这里的U实质上表示上文中过的数据集X，D代表数据集的类标。接下来我们介绍不可分辨关系，上近似、下近似、正域和依赖度。

定义 6.3. 在知识表达系统$S =< U, A, V, f >$中，根据属性集合$B \subseteq A$，不可分分辨关系（即等价关系）定义为：

$$IND(B) = \{(x, y) | \forall b \in B, f(b, x) = f(b, y)\} \tag{6.1}$$

论域的等价关系构成了论域的划分，记为$U/IND(B)$(简记为U/B)。对象$x \in U$的等价类定义为:

$$[x]_{IND(B)} = [x]_B = \{y | (x, y) \in IND(B)\} \tag{6.2}$$

定义 6.4. 假设给定一个知识表达系统$S =< U, A, V, f >$，$\forall X \subseteq U$和$B \subseteq A$，子集X关于B的下近似和上近似分别是

$$\begin{aligned} \underline{B}(X) =& \{x | \forall x \in U, [x]_B \subseteq X\} \\ \overline{B}(X) =& \{x | \forall x \in U, [x]_B \cap X \neq \varnothing\} \end{aligned} \tag{6.3}$$

集合$BN_B(X) = \overline{B}(X) - \underline{B}(X)$ 称为X关于B的边界域；$POS_B(X) = \underline{B}(X)$称为$X$关于$B$的正域。

下近似$\underline{B}(X)$或正域$POS_B(X)$是由那些根据$IND(B)$肯定属于集合X的论域U中元素组成的集合；上近似$\overline{B}(X)$是由那些根据$IND(B)$肯定属于或者可能属于集合X的论域U中元素组成的集合；边界域$BN_B(X)$是由那些根据$IND(B)$既不能判定肯定属于X也不能判定肯定不属于X的论域U中元素组成的集合。

6.2 基于依赖度的属性约简算法

在决策系统中，基于条件属性对决策属性依赖度的属性约简算法是通过Pawlak属性重要度为启发函数，所以，我们首先定义决策表中的属性的依赖度和重要度。

定义 6.5. 给定一个决策系统$DS =< U, A \cup D, V, f >$，$B \subseteq A$，定义

$$\gamma_B(D) = \frac{POS_B(D)}{|U|} = \frac{\left|\bigcup_{X \in U/D} \underline{B}(X)\right|}{|U|} \tag{6.4}$$

为D依赖于P的程度，记作$P \Rightarrow_k D$

定义 6.6. 给定一个决策系统$DS =< U, A \cup D, V, f >$，$\forall B \subseteq A$，$\forall b \in B$，定义条件属性$b$对条件属性集$B$相对于决策属性$D$的重要度为：

$$SIG(b, B, D) = \gamma_{B \cup \{b\}}(D) - \gamma_B(D) = \frac{|POS_{B \cup \{b\}}(D)| - |POS_B(D)|}{|U|} \tag{6.5}$$

其中，$\gamma_B(D)$是定义 6.5中的依赖度。

定义了属性的重要度度量之后，基于Pawlak属性重要度的属性约简算法描述如下:

算法 6 基于Pawlak属性重要度的决策表属性约简算法

输入: 决策系统$DS =< U, A \cup D, V, f >$

输出: 条件属性A相对于决策决策属性D的一个相对约简$B \in RED_D(A)$

1: $B = \varnothing$
2: $a_i = \text{argmax}_{a_i \in A-B}\, SIG(a_i, B, D)$
3: $B = B \cup \{a_i\}$
4: **while** $POS_B(D) \neq POS_A(D)$ **do**
5: $\quad a_i = \text{argmax}_{a_i \in A-B}\, SIG(a_i, B, D)$
6: $\quad B = B \cup \{a_i\}$
7: **end while**
8: **return** $B \in RED_D(A)$

6.3 基于耦合依赖度的属性选择算法DaulPOS

定义 6.7. 假设给定一个知识表达系统$S =< U, A, V, f >$，$a_i, a_j \subseteq A$，耦合依赖度定义：

$$daulDep(a_i, a_j) = \frac{POS_{a_i}(a_j) + POS_{a_j}(a_i)}{2|U|} \tag{6.6}$$

定义 6.8. 假设给定一个决策系统$DS =< U, A \cup D, V, f >$，$a_i, a_j \subseteq A$，$D$为决策属性，则相关性定义为：

$$Rel(a_i, d) = daulDep^L(a_i, d) \tag{6.7}$$

属性间冗余性定义为：

$$Red(a_i, a_j) = daulDep(a_i, a_j) \tag{6.8}$$

定义 6.9. 假设给定数据集$X = X^L \bigcup X^U$，对于有类标数据X^L，可以定义某个属性a_i对决策属性的一致率为

$$Con(a_i) = \frac{\sum_{x \in X^L} \sum_{y \in X^L} Cons_i(x, y)}{|X^L| \cdot |X^L|} \tag{6.9}$$

$Cons$为两个对象的属性a_i与类标的一致度量，表示为：

$$Cons_i(x, y) \begin{cases} 0 & a_i(x) = a_i(y) \oplus d(i) = d(j) \\ 1 & otherwise \end{cases} \tag{6.10}$$

$a_i(x) = a_i(y)$代表x和y在属性a_i下取同一个值，$d(i) = d(j)$代表类标相同，$\oplus$ 为异或符号，该式子表示属性a_i和类标一致的程度。

定义属性的相关性和属性的冗余性后，我们考虑属性选择的搜索策略，我们通过属性与类标的一致度量对属性进行排序，按照这个顺序，对于单个待考察的属性a_i，如果属性的相关性大于其对于已选属性中每个属性的冗余性，我们就将它加入的属性选择子集中，反之，就将它丢弃。知道遍历完所有的属性，算法停止。整个DaulPOS算法流程如算法 7所示。

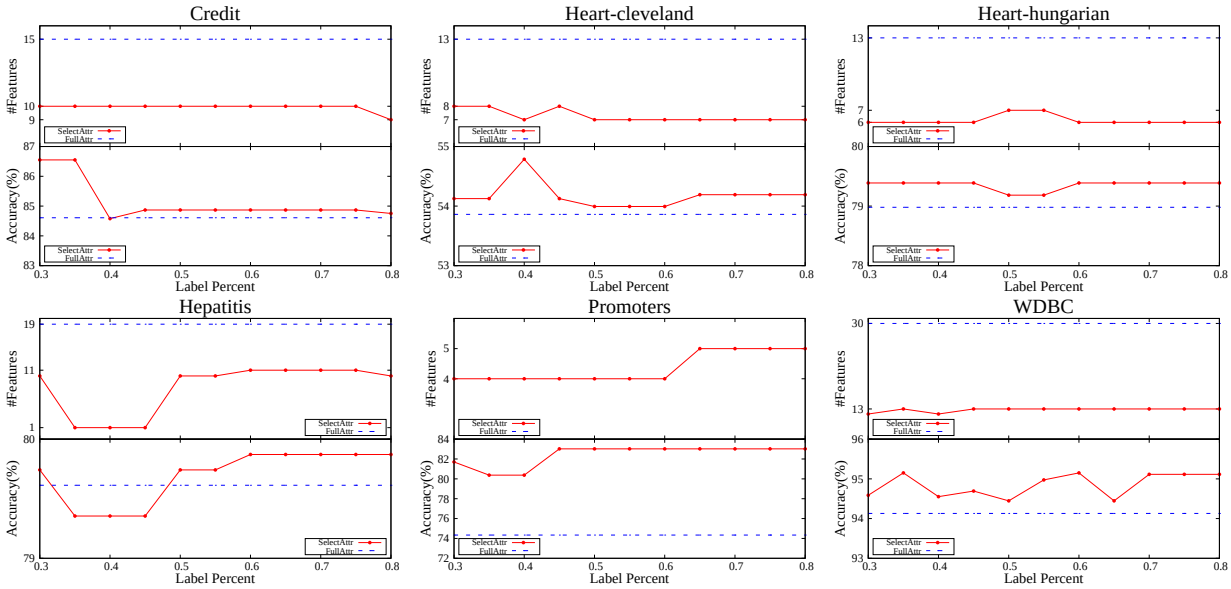


图 6.1　不同Lpercent比较特征选择个数和CART的分类精度

算法 7 基于耦合依赖度的半监督属性选择算法DaulPOS

输入： 数据集$X = X^U \cup X^L$，属性全集为A

输出： 特征选择S

1: $S = \varnothing$，$Rest = A$

2: $S = S \cup \text{argmax}_{a_i \in A} Cons(a_i)$

3: $Rest = Rest - S$

4: **while** $Rest \neq \varnothing$ **do**

5: $a_i = \text{argmax}_{a_i \in Rest} Cons(a_i)$

6: **if** $Rel(a_i, d) >= Red(a_i, a_j)$，$\forall a_j \in S$ **then**

7: $S = S \cup a_i$

8: **end if**

9: $Rest = Rest - a_i$

10: **end while**

11: 输出特征选择S

6.4　实验与分析

本章的13个测试数据集与上一章相同，数据集的详细情况见表 3.1。评价标准也是分别从分类和聚类两个方面入手。对于分类问题，我们把属性选择出子集和属性全集分别送

到某个算法中比较分类精度。对于聚类问题，我们把属性选择出子集和属性全集分别送到某个聚类算法中比较准确率ACC和归一化互信息NMI。

我们对有类标数据比例Lpercent对算法进行分析。我们将Lpercent分别取0到0.5之间，步进为0.05的11个数，计算10次每次10折交叉验证得到平均值如图 5.2所示。

表 6.1 算法DualPOS比较分类精度Accuracy

Datasets	*#Features*		*NBC*		*C4.5*		*JRip*		*PART*		*CART*	
	DualPOS	*FullFeature*	*DualPOS*	*FullFeature*	*DualPOS*	*FullFeature*	*DualPOS*	*FullFeature*	*DualPOS*	*FullFeature*	*DualPOS*	*FullFeature*
			ACC±Std	ACC±Std	ACC±Std	ACC±Std	ACC±Std	ACC±Std	ACC±Std	ACC±Std	ACC±Std	ACC±Std
Colic	6	22	81.47±0.23	81.47±0.35	80.49±0.62	84.84±0.35	81.25±0.72	84.35±0.53	78.97±0.89	80.92±1.09	80.49±0.52	84.46±0.59
Credit	10	15	87.01±0.17	86.17±0.26	86.99±0.47	86.61±0.51	85.71±0.58	86.58±0.49	85.51±0.61	85.88±0.67	84.87±0.24	84.61±0.44
Diabetes	5	8	77.89±0.25	77.84±0.11	75.70±0.77	77.50±0.92	77.01±0.47	77.32±0.56	75.52±0.62	77.06±0.76	74.77±0.79	76.74±0.80
Heart-cleveland	7	13	58.48±1.03	56.30±1.11	52.67±1.47	53.00±2.42	53.60±0.55	53.93±0.50	55.84±1.03	54.46±2.07	53.99±0.64	53.86±1.18
Heart-hungarian	7	13	83.88±0.19	84.69±0.24	79.86±1.39	80.48±0.85	80.34±1.06	79.73±0.89	80.20±1.21	81.02±0.91	79.18±1.30	78.98±1.55
Heart-statlog	8	13	81.33±0.33	83.78±0.41	82.44±0.89	82.07±1.33	83.70±1.17	83.78±1.40	84.59±2.00	83.63±1.86	81.70±1.75	82.59±0.79
Hepatitis	10	19	85.16±0.00	84.00±0.54	80.52±1.40	80.26±1.91	79.23±2.16	80.00±2.28	81.68±1.34	82.32±0.87	79.74±0.98	79.61±0.35
Ionosphere	9	34	91.51±0.13	90.77±0.38	89.34±0.59	89.74±0.67	89.91±0.72	91.51±1.16	89.86±0.38	89.91±0.32	87.75±0.94	90.31±0.83
Musk2	11	166	87.81±0.25	85.26±0.27	89.84±0.49	90.83±1.00	89.28±0.46	90.01±1.05	90.69±0.42	90.69±0.46	89.70±0.65	90.27±0.50
Promoters	4	57	90.57±0.94	90.38±0.79	81.89±1.40	79.81±1.58	81.70±2.27	81.13±2.58	86.04±1.81	85.28±1.43	83.02±1.49	74.34±2.86
SPECT	9	22	71.99±0.41	68.84±0.62	72.06±0.50	69.96±1.33	72.28±1.52	72.06±1.49	71.01±2.19	68.01±2.46	72.73±0.89	73.03±0.37
Voting	13	16	90.16±0.19	90.30±0.19	96.32±0.28	96.37±0.34	96.00±0.31	95.45±0.44	95.77±0.38	95.36±0.70	96.00±0.53	95.54±0.13
WDBC	13	30	95.89±0.44	95.85±0.10	94.31±0.24	95.85±0.10	94.73±0.51	96.34±0.19	95.82±0.47	95.99±0.23	94.45±0.55	94.13±0.92
Average	8.62	32.92	83.32	82.74	81.73	82.10	81.90	82.48	82.42	82.35	81.42	81.42

表 6.2 算法DualPOS比较聚类精度ACC

Datasets	*KModes*		*EM*		*Cobweb*		*FarthestFirst*		*CLOPE*	
	DualPOS	*FullFeature*	*DualPOS*	*FullFeature*	*DualPOS*	*FullFeature*	*DualPOS*	*FullFeature*	*DualPOS*	*FullFeature*
	ACC±Std	ACC±Std	ACC±Std	ACC±Std	ACC±Std	ACC±Std	ACC±Std	ACC±Std	ACC	ACC
Colic	62.23±10.85	58.10±8.19	77.45±0.00	66.85±0.00	72.61±3.87	66.25±2.75	63.26±8.97	57.12±8.89	79.08	74.18
Credit	72.55±10.48	75.45±11.76	86.67±0.00	73.33±0.00	83.86±0.91	71.28±2.49	65.22±5.66	60.03±8.59	80.58	79.71
Diabetes	60.05±7.24	60.05±7.24	66.41±0.00	66.41±0.00	65.73±1.19	65.73±1.19	57.14±8.83	57.14±8.83	76.30	67.06
Heart-cleveland	47.46±4.47	41.65±6.54	49.44±3.53	48.32±0.18	59.74±0.47	59.67±1.94	44.75±4.15	44.42±7.24	64.36	57.76
Heart-hungarian	82.31±1.42	82.31±1.42	84.08±0.15	84.08±0.15	81.70±1.69	81.70±1.69	77.35±5.61	77.35±5.61	80.27	77.21
Heart-statlog	82.67±0.41	82.67±0.41	80.37±0.00	80.74±0.00	80.89±1.25	79.41±3.24	68.89±12.17	65.63±14.34	72.22	74.07
Hepatitis	73.03±5.19	69.81±11.39	83.23±0.00	75.23±0.58	80.13±1.06	79.48±0.29	79.61±5.19	71.87±11.81	85.16	83.87
Ionosphere	87.35±1.02	88.77±0.96	90.03±0.00	89.17±0.00	89.12±1.22	79.43±2.87	85.36±3.11	78.35±2.38	95.73	95.73
Musk2	73.86±6.07	61.67±11.30	82.60±0.00	53.61±0.00	82.18±0.00	82.18±0.00	77.54±13.20	57.00±9.66	86.28	87.98
Promoters	62.08±7.53	61.32±10.05	61.89±5.88	57.17±5.99	73.96±16.70	69.43±3.10	62.64±3.10	58.11±6.21	100.00	100.00
SPECT	69.44±0.78	64.64±5.11	68.54±0.00	60.30±0.00	68.31±1.94	63.52±5.48	66.07±6.43	59.48±4.96	72.28	58.80
Voting	87.22±0.50	87.03±0.72	87.82±0.00	87.82±0.00	81.98±6.78	82.90±3.44	86.90±2.17	86.11±2.63	83.91	82.76
WDBC	94.48±0.10	92.79±0.48	95.25±0.00	94.55±0.00	92.76±1.15	83.23±2.46	84.08±13.55	88.51±8.37	97.19	89.63
Average	73.44	71.25	77.98	72.12	77.92	74.17	70.68	66.24	82.57	79.14

当Lercent=50%，我们提出的DaulPOS算法所选的属性子集和属性全集分别通过不同的分类器得到分类精度，不同的聚类算法得到了准确率和归一化互信息作比较。分类器有NaiveBayes Classifier(NBC)，C4.5，JRip，PART和CART。聚类器有k-Modes，EM，Cobweb，FarthestFirst和CLOPE。

表 6.3 算法DualPOS比较聚类归一化互信息NMI

Datasets	*KModes*		*EM*		*Cobweb*		*FarthestFirst*		*CLOPE*	
	DualPOS	*FullFeature*	*DualPOS*	*FullFeature*	*DualPOS*	*FullFeature*	*DualPOS*	*FullFeature*	*DualPOS*	*FullFeature*
	NMI±Std	NMI±Std	NMI±Std	NMI±Std	NMI±Std	NMI±Std	NMI±Std	NMI±Std	NMI	NMI
Colic	0.110±0.073	0.075±0.049	0.205±0.000	0.109±0.000	0.149±0.042	0.084±0.043	0.043±0.085	0.063±0.063	0.096	0.082
Credit	0.197±0.113	0.234±0.131	0.429±0.000	0.158±0.000	0.351±0.017	0.135±0.029	0.082±0.058	0.051±0.068	0.087	0.098
Diabetes	0.069±0.022	0.069±0.022	0.089±0.000	0.089±0.000	0.049±0.011	0.049±0.011	0.022±0.018	0.022±0.018	0.062	0.053
Heart-cleveland	0.210±0.014	0.210±0.022	0.242±0.007	0.224±0.003	0.178±0.016	0.179±0.026	0.164±0.018	0.143±0.054	0.151	0.139
Heart-hungarian	0.157±0.013	0.157±0.013	0.181±0.010	0.181±0.010	0.130±0.012	0.130±0.012	0.119±0.038	0.119±0.038	0.140	0.139
Heart-statlog	0.343±0.022	0.343±0.022	0.281±0.000	0.287±0.000	0.269±0.036	0.252±0.055	0.147±0.154	0.136±0.164	0.129	0.167
Hepatitis	0.150±0.016	0.159±0.052	0.300±0.000	0.187±0.005	0.175±0.059	0.143±0.042	0.103±0.079	0.077±0.109	0.109	0.175
Ionosphere	0.468±0.021	0.478±0.044	0.507±0.000	0.475±0.000	0.474±0.042	0.272±0.054	0.393±0.063	0.236±0.041	0.173	0.168
Musk2	0.139±0.063	0.051±0.033	0.206±0.000	0.026±0.000	0.093±0.023	0.039±0.027	0.168±0.084	0.040±0.023	0.078	0.092
Promoters	0.066±0.082	0.072±0.092	0.051±0.053	0.024±0.026	0.251±0.256	0.110±0.032	0.053±0.023	0.032±0.030	0.151	0.150
SPECT	0.105±0.013	0.066±0.043	0.104±0.000	0.028±0.000	0.079±0.023	0.050±0.046	0.073±0.046	0.016±0.015	0.064	0.068
Voting	0.474±0.016	0.465±0.034	0.486±0.000	0.486±0.000	0.375±0.098	0.376±0.068	0.460±0.047	0.439±0.074	0.320	0.276
WDBC	0.678±0.004	0.611±0.017	0.711±0.000	0.680±0.000	0.614±0.047	0.385±0.063	0.429±0.199	0.510±0.192	0.216	0.258
Average	0.244	0.230	0.292	0.227	0.245	0.169	0.174	0.145	0.137	0.143

实验结果分别在表格 6.1， 6.3和 6.3中所示。首先可以看到，我们提出的DualPOS能明显的降低属性空间的大小，对于分类效果来看可以得出与上一章相同的结论，属性选择后的数据集和全部属性的数据集在分类精度上可以保持大体相当，这表明所选的属性在很大程度上保持了原有的信息。对于聚类效果评价指标，都要比在属性全集上做聚类有较明显的优势。

接下来，我们本文所提的方法进行一个综合的比较，图 **??**和图 6.5分别表示了SemiMRMR(SM)算法和DualPOS(DP)算法在经典的分类器上和聚类器上训练所用的时间。从图中我们可以清楚的观察到经过本文提出的算法后，不管分类和聚类，训练所用的时间都有明显的减少。图 6.6列出了SM算法和DP算法在本文提出的半监督聚类算法所用的时间。我们可以得出SM和DP属性选择算法能提高本文提出的半监督聚类算法的时间效率。

图 6.7和图 6.8分别表示了本文的两种半监督聚类算法经过了SM和DP算法降维后聚类的ACC和NMI，从表中可以看出降维后的数据大都能保持属性全集上的聚类效果，并且在某些数据集上要优于半监督聚类算法在属性全集上的ACC和NMI。

6.5 小结

在本章中，我们将粗糙集理论中传统的依赖度，拓展到了半监督领域，称作耦合依赖度。耦合依赖度不仅可以度量条件属性对决策属性的依赖程度，还能度量条件属性间的

表 6.4 比较属性选择时间和各个分类算法所用时间

Datasets	*FeatureSelection*		*NBC*			*C4.5*			*JRip*			*PART*			*CART*		
	SM	*DP*	*SM*	*DP*	*FF*	*SM*	*DP*	*FF*	*SM*	*DP*	*FF*	*SM*	*DP*	*FF*	*SM*	*DP*	*FF*
Colic	0.933	0.284	0.014	0.003	0.007	0.033	0.012	0.026	0.070	0.045	0.095	0.058	0.020	0.065	1.316	0.796	2.144
Credit	1.808	0.547	0.007	0.007	0.008	0.043	0.033	0.042	0.150	0.128	0.179	0.061	0.068	0.088	1.312	2.189	2.518
Diabetes	0.689	0.185	0.006	0.006	0.006	0.022	0.020	0.029	0.150	0.125	0.176	0.043	0.040	0.063	1.010	1.196	1.600
Heart-cleveland	0.226	0.089	0.004	0.003	0.004	0.021	0.017	0.023	0.085	0.050	0.108	0.065	0.039	0.071	0.737	0.659	0.969
Heart-hungarian	0.149	0.078	0.003	0.003	0.004	0.009	0.011	0.016	0.023	0.037	0.049	0.017	0.020	0.033	0.419	0.549	0.809
Heart-statlog	0.201	0.067	0.003	0.003	0.003	0.015	0.012	0.016	0.066	0.046	0.063	0.035	0.022	0.029	0.519	0.454	0.589
Hepatitis	0.117	0.030	0.002	0.002	0.003	0.006	0.007	0.010	0.024	0.021	0.027	0.010	0.011	0.024	0.238	0.243	0.388
Ionosphere	1.806	0.422	0.004	0.003	0.007	0.012	0.010	0.024	0.082	0.100	0.183	0.020	0.021	0.055	1.114	1.405	4.531
Musk2	257.825	10.201	0.023	0.008	0.057	0.136	0.032	0.308	0.430	0.185	1.562	0.396	0.076	0.999	7.292	2.121	29.238
Promoters	0.612	0.083	0.001	0.001	0.004	0.002	0.003	0.013	0.011	0.015	0.059	0.005	0.005	0.028	0.105	0.101	1.086
SPECT	0.397	0.133	0.004	0.003	0.004	0.015	0.016	0.033	0.027	0.037	0.060	0.038	0.033	0.075	0.520	0.486	1.066
Voting	0.382	0.177	0.003	0.005	0.005	0.005	0.014	0.018	0.022	0.064	0.051	0.007	0.024	0.028	0.182	0.653	0.789
WDBC	2.906	0.984	0.009	0.006	0.011	0.012	0.018	0.037	0.056	0.146	0.191	0.016	0.028	0.051	0.605	1.752	3.263
Average	20.619	1.022	0.006	0.004	0.010	0.026	0.016	0.046	0.092	0.077	0.215	0.059	0.031	0.124	1.182	0.970	3.769

表 6.5 比较属性选择时间和各个聚类算法所用时间

Datasets	*FeatureSelection*		*KModes*			*EM*			*Cobweb*			*FarthestFirst*			*CLOPE*		
	SM	*DP*	*SM*	*DP*	*FF*	*SM*	*DP*	*FF*	*SM*	*DP*	*FF*	*SM*	*DP*	*FF*	*SM*	*DP*	*FF*
Colic	0.933	0.284	0.013	0.008	0.016	0.107	0.098	0.203	0.218	0.216	0.421	0.003	0.003	0.003	0.098	0.297	0.534
Credit	1.808	0.547	0.009	0.013	0.020	0.215	0.206	0.238	0.298	0.522	0.722	0.003	0.003	0.004	0.180	0.872	0.948
Diabetes	0.689	0.185	0.007	0.008	0.009	0.108	0.150	0.178	0.177	0.237	0.255	0.002	0.002	0.002	0.142	0.519	0.361
Heart-cleveland	0.226	0.089	0.006	0.005	0.009	0.188	0.156	0.319	0.087	0.089	0.134	0.002	0.002	0.003	0.027	0.135	0.112
Heart-hungarian	0.149	0.078	0.004	0.006	0.007	0.052	0.224	0.281	0.059	0.105	0.119	0.002	0.003	0.003	0.039	0.091	0.096
Heart-statlog	0.201	0.067	0.004	0.004	0.005	0.071	0.064	0.073	0.074	0.071	0.090	0.001	0.001	0.001	0.040	0.098	0.048
Hepatitis	0.117	0.030	0.002	0.003	0.005	0.031	0.031	0.060	0.035	0.039	0.074	0.001	0.001	0.001	0.015	0.042	0.037
Ionosphere	1.806	0.422	0.005	0.007	0.032	0.087	0.106	0.301	0.180	0.192	0.786	0.001	0.001	0.004	0.585	1.502	5.691
Musk2	257.825	10.201	0.071	0.036	0.222	1.047	0.491	2.553	1.644	0.949	6.468	0.015	0.007	0.040	0.292	1.216	5.146
Promoters	0.612	0.083	0.001	0.013	0.013	0.027	0.174	0.268	0.021	0.238	0.348	0.000	0.001	0.002	0.055	0.657	1.124
SPECT	0.397	0.133	0.004	0.004	0.010	0.103	0.091	0.259	0.062	0.074	0.177	0.001	0.001	0.002	0.030	0.071	0.118
Voting	0.382	0.177	0.006	0.009	0.010	0.071	0.117	0.120	0.098	0.216	0.246	0.001	0.002	0.003	0.050	0.105	0.138
WDBC	2.906	0.984	0.008	0.015	0.034	0.147	0.174	0.405	0.275	0.523	1.022	0.002	0.003	0.006	0.362	2.250	2.218
Average	20.619	1.022	0.011	0.010	0.030	0.173	0.160	0.405	0.248	0.267	0.836	0.003	0.002	0.006	0.147	0.604	1.275

表 6.6 比较属性选择时间和各个半监督聚类算法所用时间

Datasets	$W-Voting$			$S-W-Voting$			$Semi-HC-Avg$			$Semi-HC-Mean$		
	SM	*DP*	*FF*	*SM*	*DP*	*FF*	*SM*	*DP*	*FF*	*SM*	*DP*	*FF*
Colic	0.147	0.137	0.198	0.145	0.105	0.214	0.033	0.063	0.052	0.338	0.430	0.488
Credit	0.320	0.266	0.348	0.268	0.251	0.348	0.084	0.176	0.278	1.562	4.548	7.671
Diabetes	0.228	0.230	0.254	0.242	0.226	0.254	0.076	0.082	0.112	0.320	1.475	1.501
Heart-cleveland	0.102	0.087	0.130	0.112	0.080	0.138	0.016	0.009	0.016	0.058	0.058	0.135
Heart-hungarian	0.090	0.099	0.117	0.115	0.103	0.135	0.006	0.009	0.009	0.023	0.065	0.047
Heart-statlog	0.084	0.065	0.090	0.086	0.064	0.082	0.007	0.017	0.015	0.040	0.055	0.090
Hepatitis	0.039	0.041	0.078	0.046	0.037	0.075	0.003	0.002	0.004	0.014	0.016	0.037
Ionosphere	0.102	0.111	0.370	0.111	0.100	0.386	0.018	0.025	0.062	0.184	0.277	0.677
Musk2	1.078	0.364	2.998	1.094	0.366	2.994	0.268	0.138	0.596	8.966	5.426	19.924
Promoters	0.018	0.166	0.208	0.016	0.166	0.202	0.000	0.004	0.002	0.006	0.006	0.008
SPECT	0.058	0.056	0.150	0.058	0.068	0.150	0.012	0.008	0.018	0.072	0.082	0.146
Voting	0.106	0.136	0.166	0.104	0.134	0.162	0.020	0.032	0.042	0.086	0.296	0.610
WDBC	0.192	0.244	0.468	0.180	0.236	0.486	0.044	0.114	0.128	0.588	2.444	2.814
Average	0.197	0.154	0.429	0.198	0.149	0.433	0.045	0.052	0.103	0.943	1.168	2.627

表 6.7 算法SemiMRMR和DualPOS比较半监督聚类精度Accuracy

Datasets	$W-Voting$			$S-W-Voting$			$Semi-HC-Avg$			$Semi-HC-Mean$		
	SemiMRMR	*DualPOS*	*FullAttr*	*SemiMRMR*	*DualPOS*	*FullAttr*	*SemiMRMR*	*DualPOS*	*FullAttr*	*SemiMRMR*	*DualPOS*	*FullAttr*
	ACC±Std	ACC±Std	ACC±Std	ACC±Std	ACC±Std	ACC±Std	ACC	ACC	ACC	ACC	ACC	ACC
Colic	77.88±1.18	79.40±0.59	74.24±6.64	77.99±1.07	78.59±1.29	74.08±6.27	64.67	65.49	65.76	70.65	68.75	66.85
Credit	80.09±2.94	81.80±1.44	82.20±0.45	77.62±1.49	81.57±2.26	81.86±0.63	71.59	72.61	72.32	61.88	60.87	60.14
Diabetes	65.16±6.72	60.36±8.78	60.36±8.78	64.66±6.64	59.40±7.29	59.40±7.29	66.02	66.80	66.80	73.96	71.09	71.22
Heart-cleveland	45.81±9.78	50.63±5.83	43.70±4.73	45.08±10.73	52.54±4.41	43.89±5.03	56.11	55.78	55.12	67.33	64.69	61.06
Heart-hungarian	79.93±3.83	82.04±1.76	82.04±1.76	80.00±3.67	82.99±0.59	82.99±0.59	44.90	35.37	35.37	49.32	46.60	46.60
Heart-statlog	83.70±1.87	82.22±0.00	82.22±0.00	82.89±2.30	82.22±0.00	82.22±0.00	69.26	68.52	68.89	60.37	63.70	67.41
Hepatitis	75.35±0.71	75.74±0.74	74.84±2.41	75.23±2.64	75.61±1.15	73.68±3.72	76.13	63.87	78.06	82.58	83.23	80.65
Ionosphere	81.42±6.18	88.55±0.62	88.03±0.20	81.71±6.63	88.83±0.13	87.92±0.16	70.09	70.09	70.09	68.38	70.09	67.81
Musk2	74.77±4.13	75.70±1.86	53.35±0.06	78.81±3.21	78.22±1.70	53.41±0.08	86.99	59.97	59.55	82.46	83.17	82.46
Promoters	54.34±2.72	58.49±11.26	56.79±8.71	54.72±4.62	57.36±8.86	57.55±7.46	71.70	71.70	71.70	74.53	65.09	68.87
SPECT	59.40±3.58	69.96±0.56	63.22±0.89	59.85±4.31	69.51±0.82	63.22±1.97	71.91	70.79	71.91	72.28	72.28	68.16
Voting	84.74±3.14	86.67±0.00	86.11±0.60	84.74±3.14	86.67±0.00	86.53±0.76	61.61	64.60	65.98	61.38	64.60	61.84
WDBC	90.86±0.00	94.45±0.10	92.97±0.48	90.69±0.39	94.45±0.10	93.32±0.00	67.14	68.01	68.01	64.67	63.62	63.80
Average	73.34	75.85	72.31	73.38	76.00	72.31	67.55	64.12	65.35	68.45	67.52	66.68

表 6.8 算法SemiMRMR和DualPOS比较半监督聚类归一化互信息NMI

Datasets	$W-Voting$			$S-W-Voting$			$Semi-HC-Avg$			$Semi-HC-Mean$		
	SemiMRMR	*DualPOS*	*FullAttr*	*SemiMRMR*	*DualPOS*	*FullAttr*	*SemiMRMR*	*DualPOS*	*FullAttr*	*SemiMRMR*	*DualPOS*	*FullAttr*
	NMI±Std	NMI±Std	NMI±Std	NMI±Std	NMI±Std	NMI±Std	NMI	NMI	NMI	NMI	NMI	NMI
Colic	0.220±0.015	0.234±0.006	0.193±0.057	0.223±0.012	0.218±0.024	0.188±0.054	0.137	0.231	0.222	0.113	0.092	0.072
Credit	0.301±0.039	0.312±0.032	0.319±0.010	0.263±0.008	0.307±0.049	0.311±0.014	0.244	0.295	0.271	0.037	0.031	0.027
Diabetes	0.064±0.027	0.063±0.032	0.063±0.032	0.058±0.026	0.066±0.018	0.066±0.018	0.184	0.235	0.235	0.141	0.105	0.106
Heart-cleveland	0.197±0.045	0.220±0.059	0.185±0.037	0.179±0.051	0.240±0.024	0.194±0.052	0.283	0.216	0.281	0.281	0.268	0.260
Heart-hungarian	0.163±0.038	0.161±0.023	0.161±0.023	0.159±0.026	0.187±0.026	0.187±0.026	0.135	0.130	0.130	0.244	0.126	0.126
Heart-statlog	0.361±0.040	0.319±0.000	0.319±0.000	0.345±0.050	0.319±0.000	0.319±0.000	0.147	0.206	0.194	0.034	0.059	0.092
Hepatitis	0.199±0.020	0.196±0.018	0.195±0.003	0.177±0.045	0.181±0.040	0.190±0.014	0.015	0.200	0.006	0.115	0.133	0.086
Ionosphere	0.306±0.137	0.493±0.014	0.441±0.007	0.316±0.144	0.498±0.011	0.437±0.005	0.305	0.305	0.305	0.075	0.085	0.062
Musk2	0.145±0.026	0.156±0.014	0.027±0.000	0.169±0.030	0.178±0.015	0.027±0.000	0.202	0.164	0.143	0.124	0.139	0.127
Promoters	0.007±0.007	0.058±0.118	0.035±0.054	0.012±0.016	0.037±0.072	0.035±0.060	0.261	0.261	0.261	0.185	0.068	0.107
SPECT	0.014±0.011	0.107±0.006	0.034±0.004	0.019±0.012	0.100±0.012	0.035±0.011	0.310	0.236	0.310	0.136	0.131	0.086
Voting	0.460±0.032	0.491±0.000	0.448±0.026	0.460±0.032	0.491±0.000	0.452±0.028	0.059	0.071	0.092	0.028	0.048	0.031
WDBC	0.537±0.000	0.676±0.004	0.617±0.017	0.532±0.012	0.676±0.004	0.629±0.000	0.199	0.276	0.263	0.045	0.039	0.041
Average	0.229	0.268	0.234	0.224	0.269	0.236	0.191	0.217	0.209	0.120	0.102	0.094

冗余程度。对于两种属性选择方法，我们提出了两种不同的搜索策略。我们做了两方面实验，与上一章相同，我们分别用了不同分类器和聚类算法验证我们的方法，结果表明在分类预测精度和聚类效果不降低的前提下，DaulPOS能有效降低数据集的维度。另一方面，我们将上文中提到半监督聚类方法与两种半监督属性选择方法结合，并从时间和聚类效果做了验证。

第7章 总结与展望

7.1 主要工作总结

本文对半监督聚类和属性选择进行了深入的研究和探讨，发现符号属性数据在半监督学习下研究的价值，并提出了相应的解决方案。本文的主要工作包括:

1) 结合聚类集成思想，基于权值投票的半监督聚类集成方法，对于每个聚类成员，我们分别计算有监督的权重和无监督的权重，共同投票产生最终的聚类结果。并提出了四种不同的投票策略。并做了相应的实验验证其有效性。

2) 通过对整个数据集先分裂再组合，提出基于分裂再组合的半监督聚类方法，并比较了不同的分裂策略和组合策略。实验表明，该方法能随着带类标的比重不断加大，效果不断提升。

3) 基于mRMR算法，提出了一种最大相关最小冗余半监督属性选择，重新定义了属性的相关性和属性间的冗余，不仅考虑无监督信息，而且还考虑单个属性对整个属性子集的作用。并提出了一种新的停止准则。

4) 拓展了粗糙集中的依赖度定义，使其不仅能度量与类别相关性，而且能度量两个属性之间的冗余性。并提出了属性一致性度量的概念和对应的属性选择算法。

7.2 研究展望

本文对符号属性数据的半监督聚类和属性选择做了一定程度上的探索，但是还有一些工作有待进一步研究，包括:

1) 基于权值投票的半监督聚类集成方法中结合有监督部分权值和无监督部分权值有一个比例系数α，需要一定的先验知识。

2) 基于分裂再组合的半监督聚类方法在时间代价相对较大，以后考虑换种集簇组合方法。

3) 半监督最小冗余最大相关算法SemiMRMR中，同样存在松弛变量α，需要进一步研究。

4) 耦合依赖度DaulPOS算法中，可以引入一个参数λ预先控制待选属性子集的长度。

参考文献

[1] T.M. Mitchell著, 曾华军等译. 机器学习[M]. 北京: 机械工业出版社, 2003.

[2] E. Mjolsness, D. DeCoste. Machine learning for science: state of the art and future prospects[J]. Science, 2001, 293(5537):2051–2055.

[3] 管仁初, 梁艳春. 半监督聚类算法的研究与应用[D]. [博士学位论文], 吉林大学, 2010.

[4] B. Mirkin. Clustering for data mining: a data recovery approach[M]. Chapman & Hall/CRC, 2005.

[5] M. Plasse, N. Niang, G. Saporta, A. Villeminot, L. Leblond. Combined use of association rules mining and clustering methods to find relevant links between binary rare attributes in a large data set[J]. Computational Statistics & Data Analysis, 2007, 52(1):596–613.

[6] P. Chopra, J. Kang, J. Yang, H.J. Cho, H.S. Kim, M.G. Lee. Microarray data mining using landmark gene-guided clustering[J]. BMC bioinformatics, 2008, 9(1):92.

[7] V.S. Tseng, C.P. Kao. Efficiently mining gene expression data via a novel parameterless clustering method[J]. Computational Biology and Bioinformatics, IEEE/ACM Transactions on, 2005, 2(4):355–365.

[8] M. Li, L. Zhang. Multinomial mixture model with feature selection for text clustering[J]. Knowledge-Based Systems, 2008, 21(7):704–708.

[9] Y. Li, C. Luo, S.M. Chung. Text clustering with feature selection by using statistical data[J]. Knowledge and Data Engineering, IEEE Transactions on, 2008, 20(5):641–652.

[10] E. Alpaydin著,范明,咎红英等译. 机器学习导论[M]. 北京: 机械工业出版社, 2009.

[11] J. He, A.H. Tan, C.L. Tan. Modified art 2a growing network capable of generating a fixed number of nodes[J]. Neural Networks, IEEE Transactions on, 2004, 15(3):728–737.

[12] O. Chapelle, B. Schölkopf, A. Zien, et al. Semi-supervised learning[M]. MIT press Cambridge, MA, 2006.

[13] 高滢, 刘大有, 齐红. 一种半监督置均值多关系数据聚类算法[J]. 软件学报, 2008, 19:2814–2821.

[14] S. Zhong. Semi-supervised model-based document clustering: A comparative study[J]. Machine Learning, 2006, 65(1):3–29.

[15] Kiri Wagstaff, Claire Cardie, Seth Rogers, Stefan Schrödl. Constrained k-means clustering with background knowledge. In Proceedings of the Eighteenth International Conference on Machine Learning. San Francisco, CA, USA, 2001:577–584.

[16] D. Huang, W. Pan. Incorporating biological knowledge into distance-based clustering analysis of microarray gene expression data[J]. Bioinformatics, 2006, 22(10):1259–1268.

[17] H. Chang, D.Y. Yeung. Locally linear metric adaptation with application to semi-supervised clustering and image retrieval[J]. Pattern Recognition, 2006, 39(7):1253–1264.

[18] Sugato Basu, Arindam Banerjee, Raymond J. Mooney. Semi-supervised clustering by seeding. In Proceedings of the Nineteenth International Conference on Machine Learning. San Francisco, CA, USA, 2002:27–34.

[19] Dan Klein, Sepandar D. Kamvar, Christopher D. Manning. From instance-level constraints to space-level constraints: Making the most of prior knowledge in data clustering. In Proceedings of the Nineteenth International Conference on Machine Learning. San Francisco, CA, USA, 2002:307–314.

[20] 王玲, 薄列峰, 焦李成. 密度敏感的半监督谱聚类[J]. 软件学报, 2007, 18(10):2412–2422.

[21] H. Ralambondrainy. A conceptual version of the k-means algorithm[J]. Pattern Recognition Letters, 1995, 16(11):1147–1157.

[22] D.H. Fisher. Knowledge acquisition via incremental conceptual clustering[J]. Machine learning, 1987, 2(2):139–172.

[23] Y. Reich, S.J. Fenves. The formation and use of abstract concepts in design. In Concept formation: knowledge and experience in unsupervised learning. Citeseer, 1991.

[24] Kiri Wagstaff, Claire Cardie. Clustering with instance-level constraints. In Proceedings of the Seventeenth International Conference on Machine Learning. 2000:1103–1110.

[25] Z. Huang. A fast clustering algorithm to cluster very large categorical data sets in data mining. In Workshop on Research Issues on Data Mining and Knowledge Discovery. 1997.

[26] S. Guha, R. Rastogi, K. Shim. Rock: A robust clustering algorithm for categorical attributes[J]. Information systems, 2000, 25(5):345–366.

[27] D. Barbará, Y. Li, J. Couto. Coolcat: an entropy-based algorithm for categorical clustering. In Proceedings of the eleventh international conference on Information and knowledge management. ACM, 2002:582–589.

[28] A.L. Blum, P. Langley. Selection of relevant features and examples in machine learning[J]. Artificial intelligence, 1997, 97:245–271.

[29] R.B. Bhatt, M. Gopal. On fuzzy-rough sets approach to feature selection[J]. Pattern Recognition Letters, 2005, 26(7):965–975.

[30] R. Kohavi, G.H. John. Wrappers for feature subset selection[J]. Artificial intelligence, 1997, 97(1):273–324.

[31] Huan Liu, Rudy Setiono. A Probabilistic Approach to Feature Selection - A Filter Solution. In International Conference on Machine Learning. 1996:319–327.

[32] X. Wang, J. Yang, X. Teng, W. Xia, R. Jensen. Feature selection based on rough sets and particle swarm optimization[J]. Pattern Recognition Letters, 2007, 28(4):459–471.

[33] L. Yu, H. Liu. Efficient feature selection via analysis of relevance and redundancy[J]. The Journal of Machine Learning Research, 2004, 5:1205–1224.

[34] A.L. Blum, R.L. Rivest. Training a 3-node neural network is np-complete[J]. Neural Networks, 1992, 5(1):117–127.

[35] Z. Wu, C. Li. Feature selection using transductive support vector machine. In Proc. NIPS 2003 Workshop Feature Selection. 2003.

[36] Thorsten Joachims. Transductive inference for text classification using support vector machines. In Proceedings of the Sixteenth International Conference on Machine Learning. San Francisco, CA, USA, 1999:200–209.

[37] Z. Zhao, H. Liu. Semi-supervised feature selection via spectral analysis. In Proceedings of the 7th SIAM International Conference on Data Mining, Minneapolis, MN. 2007:1151–1158.

[38] Y. Cheng, Y. Cai, Y. Sun, J. Li. Semi-supervised feature selection under logistic i-relief framework. In Pattern Recognition, 2008. ICPR 2008. 19th International Conference on. IEEE, 2008:1–4.

[39] Y. Yaslan, Z. Cataltepe. Co-training with relevant random subspaces[J]. Neurocomputing, 2010, 73(10):1652–1661.

[40] A. Blum, T. Mitchell. Combining labeled and unlabeled data with co-training. In Proceedings of the eleventh annual conference on Computational learning theory. ACM, 1998:92–100.

[41] D. Zhang, S. Chen, Z.H. Zhou. Constraint score: A new filter method for feature selection with pairwise constraints[J]. Pattern Recognition, 2008, 41(5):1440–1451.

[42] D. Sun, D. Zhang. Bagging constraint score for feature selection with pairwise constraints[J]. Pattern Recognition, 2010, 43(6):2106–2118.

[43] Thomas G. Dietterich. Ensemble methods in machine learning. In Proceedings of the First International Workshop on Multiple Classifier Systems. London, UK, UK: Springer-Verlag, 2000:1–15.

[44] M. Seeger. Learning with labeled and unlabeled data. Technical report, Technical report, University of Edinburgh, 2001.

[45] L.K. Hansen, P. Salamon. Neural network ensembles[J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 1990, 12(10):993–1001.

[46] L. Breiman. Bagging predictors[J]. Machine learning, 1996, 24(2):123–140.

[47] R.E. Schapire. A brief introduction to boosting. In International Joint Conference on Artificial Intelligence. 1999, volume 16:1401–1406.

[48] L. Breiman. Random forests[J]. Machine learning, 2001, 45(1):5–32.

[49] Alexander Strehl, Joydeep Ghosh, Claire Cardie. Cluster ensembles - a knowledge reuse framework for combining multiple partitions[J]. Journal of Machine Learning Research, 2002, 3:583–617.

[50] Aristides Gionis, Heikki Mannila, Panayiotis Tsaparas. Clustering aggregation[J]. ACM Trans. Knowl. Discov. Data, 2007, 1(1):1–30.

[51] 杨草原, 刘大有, 杨博, 池淑珍, 金弟. 聚类集成方法研究[J]. 计算机科学, 2011, 38(2):166–170.

[52] 阳琳, 王文渊. 聚类融合方法综述[J]. 计算机应用研究, 2005, 22(012):8–10.

[53] Zhi-Hua Zhou, Wei Tang. Clusterer ensemble[J]. Know.-Based Syst., 2006, 19(1):77–83.

[54] A. Topchy, A.K. Jain, W. Punch. Combining multiple weak clusterings. In Data Mining, 2003. ICDM 2003. Third IEEE International Conference on. 2003:331 – 338.

[55] L.I. Kuncheva, S.T. Hadjitodorov. Using diversity in cluster ensembles. In Systems, man and cybernetics, 2004 IEEE international conference on. IEEE, 2004, volume 2:1214–1219.

[56] S.T. Hadjitodorov, L.I. Kuncheva, L.P. Todorova. Moderate diversity for better cluster ensembles[J]. Information Fusion, 2006, 7(3):264–275.

[57] Y. Yang, M.S. Kamel. An aggregated clustering approach using multi-ant colonies algorithms[J]. Pattern Recognition, 2006, 39(7):1278–1289.

[58] B. Minaei-Bidgoli, A. Topchy, W.F. Punch. Ensembles of partitions via data resampling. In Information Technology: Coding and Computing, 2004. Proceedings. ITCC 2004. International Conference on. IEEE, 2004, volume 2:188–192.

[59] A. Topchy, B. Minaei-Bidgoli, A.K. Jain, W.F. Punch. Adaptive clustering ensembles. In Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on. IEEE, 2004, volume 1:272–275.

[60] 罗会兰, 孔繁胜. 聚类集成关键技术研究[D]. [博士学位论文], 浙江大学, 2007.

[61] S. Dudoit, J. Fridlyand. Bagging to improve the accuracy of a clustering procedure[J]. Bioinformatics, 2003, 19(9):1090–1099.

[62] Ana L. N. Fred. Finding consistent clusters in data partitions. In Proceedings of the Second International Workshop on Multiple Classifier Systems. London, UK, UK: Springer-Verlag, 2001:309–318.

[63] A.L.N. Fred, A.K. Jain. Data clustering using evidence accumulation. In Pattern Recognition, 2002. Proceedings. 16th International Conference on. IEEE, 2002, volume 4:276–280.

[64] A. Topchy, A.K. Jain, W. Punch. Clustering ensembles: Models of consensus and weak partitions[J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2005, 27(12):1866–1881.

[65] Y. Yang, MS Kamel, F. Jin. Art-based clustering aggregation. In Granular Computing, 2006 IEEE International Conference on. IEEE, 2006:482–485.

[66] A. Frank, A. Asuncion. UCI machine learning repository, 2010.

[67] W. Xu, X. Liu, Y. Gong. Document clustering based on non-negative matrix factorization. In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval. ACM, 2003:267–273.

[68] L. Lovász, M.D. Plummer. Matching theory[M]. American Mathematical Society, 2009.

[69] R. Sibson. Slink: an optimally efficient algorithm for the single-link cluster method[J]. The Computer Journal, 1973, 16(1):30–34.

[70] D. Defays. An efficient algorithm for a complete link method[J]. The Computer Journal, 1977, 20(4):364–366.

[71] H. Peng, F. Long, C. Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy[J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2005, 27(8):1226–1238.

[72] C. Ding, H. Peng. Minimum redundancy feature selection from microarray gene expression data[J]. Journal of bioinformatics and computational biology, 2005, 3(02):185–205.

[73] Usama M. Fayyad, Keki B. Irani. Multi-interval discretization of continuousvalued attributes for classification learning. In Thirteenth International Joint Conference on Articial Intelligence. Morgan Kaufmann Publishers, 1993, volume 2:1022–1027.

[74] L. Breiman, J. Friedman, R. Olshen, C. Stone. Classification and Regression Trees[M]. Monterey, CA: Wadsworth and Brooks, 1984.

[75] Z. Pawlak. Rough Sets: Theoretical Aspects of Reasoning about Data. Volume 9 of System Theory, Knowledge Engineering and Problem Solving[M]. Kluwer Academic Publishers, Dordrecht, The Netherlands, 1991.

[76] Z. Pawlak. Rough sets. In Proceedings of the 1995 ACM 23rd annual conference on Computer science. ACM, 1995:262–264.

[77] Z. Pawlak. Vagueness and uncertainty: a rough set perspective[J]. Computational intelligence, 1995, 11(2):227–232.

[78] Z. Pawlak. Rough set theory and its applications to data analysis[J]. Cybernetics & Systems, 1998, 29(7):661–688.

[79] Z. Pawlak, A. Skowron. Rough set rudiments[J]. Bulletin of International Rough Set Society, 1999, 3(4):181–185.

[80] Z. Pawlak, A. Skowron. Rudiments of rough sets[J]. Information sciences, 2007, 177(1):3–27.

[81] 许青, 代建华. 不完备信息系统中的粗糙集理论与方法[D]. [硕士学位论文], 浙江大学, 2012.

发表文章目录

[1] Jianhua Dai, **Wentao Wang**, Qing Xu, Haowei Tian. Uncertainty measurement for interval-valued decision systems based on extended conditional entropy[J]. Knowledge-Based Systems, 2011(27):443450

[2] Jianhua Dai, Qing Xu, **Wentao Wang**, Haowei Tian, Conditional Entropy for Incomplete Decision Systems and Its Application in Data Mining[J]. International Journal of General Systems,2012(41):713728.

[3] Jianhua Dai, Qing Xu, **Wentao Wang**. A Comparative Study on Strategies of Rule Induction for Incomplete Data Based on Rough Set Approach[J]. International Journal of Advancements in Computing Technology, 2011(3):176183.

[4] 代建华, **王文涛**. 区间值信息系统的粒计算模型与方法, 《粒计算研究丛书: 云模型与粒计算》, 科学出版社, 2012, 第5章:94107.

致　　谢

在本硕士论文即将完成之际，我谨向所有帮助、支持过我的人表示崇高的敬意和诚挚的感谢。

首先要衷心的感谢我的导师，代建华副教授。代老师无论是在学习、科研，还是在生活、工作中都给予我莫大的支持和鼓励。代老师严谨求实的治学态度、一丝不苟的钻研精神、深厚的学术功底、敏锐的洞察力和平易近人的为师风范都让我受益匪浅。由衷祝愿代老师今后工作顺心，科研工作更上一层楼。

感谢我的师弟田浩炜，刘亮，感谢寝室室友，同学给我的帮助，祝愿大家前途似锦，工作、学习一帆风顺。

最后感谢我的爸爸妈妈，在我最困难的时候一直支持我，安慰我，祝愿你们永远健康，幸福。

署名:

当前日期:　　年　　月　　日