天津大学

博士学位论文

基于机器学习的信息过滤和信息检索的模型和算法研究

姓名：张亮

申请学位级别：博士

专业：管理科学与工程

指导教师：李敏强

20070601

中文摘要

随着 Internet 技术的飞速发展，信息网络在人们的工作生活中具有越来越重要的地位。从网络上的海量信息中快速、高效地获取人们真正需要的信息资源，已成为信息社会中的一个关键问题。信息过滤和信息检索技术是解决这一问题的有效方法，具有重要的学术意义和应用价值。本文基于统计机器学习方法，重点研究了信息过滤和信息检索模型与求解算法。主要研究内容包括：

首先，介绍了信息过滤和信息检索的概念和意义，总结了它们的起步和发展情况。概括介绍了几种基于统计的机器学习方法的概念和特点以及它们在信息过滤和信息检索中的应用，作为本文的理论基础。

其次，介绍了协同过滤问题的几种常见方法，提出了应用于协同过滤的一种概率模型，称为真实偏好高斯混合模型。新模型引入了两个隐含变量，分别用于描述用户类和项目类，用户和项目依概率可以同时属于多个类中。模型中考虑了用户评分习惯以及项目的公众评价对用户－项目最终评价的综合影响。与传统协同过滤模型相比，新模型更符合用户评价的实际情况。

第三，研究了有限混合模型在大规模文本数据聚类问题中的应用，提出了用有限混合模型进行无监督文本聚类的一种规范的广义方法。它将模型选择，特征选择以及混合模型的参数估计纳入一个统一的框架。定义了一种改进的"特征显著性"方法，将特征对各混合成员的相关性作为隐变量引入混合模型，在估计模型参数的同时完成特征选择。发展了一种带特征选择的多项式混合模型，作为广义方法的实例做了详细的说明。

第四，采用基于图的方法研究半监督学习问题。主要思想是定义样本间基于密度距离的相似度，得到数据集的内在结构信息，并将其引入学习器加以利用。对半监督分类，定义了一种基于密度的距离来反映数据点间的相似度，在此基础上以一种 Laplacian 核方法来构造整个特征空间上的超分类面。对半监督聚类问题，提出了一种基于密度的约束扩展方法。根据样本点间基于密度的距离和相似度关系，对已知约束集进行扩展，扩展后的约束集包含了数据集的内在结构信息。

最后，对论文的主要研究工作进行总结，展望了今后的研究前景。

关键词： 协同过滤，无监督学习，半监督分类，半监督聚类，有限混合模型

ABSTRACT

With the rapid development of the Internet technologies, information networks play more and more important roles in people's routine work and daily life. To obtain information that people really need from the massive information quickly and efficiently has become a key problem in our information research society. There are two main approaches to solve this problem: information filtering (IF) and information retrieval (IR), which are of important academic interest and valuable applications. The main research work of this thesis is based on statistical machine learning methods, especially the IF/IR models and algorithms. The main contents are as follows:

First, a brief introduction to IF/IR is given, including the concept, structure and features as well as their origin and history. As the theory basis of this thesis, several statistical machine learning methods and their functions in IF/IR are also introduced.

Second, on the basis of introduction on several popular collaborative filtering approaches, this thesis presents a new probabilistic model for collaborative filtering, named real preference Gaussian mixture model. It has two latent variables corresponding to classes of user and item. Each user or item may be probabilistically clustered to more than one groups. And it also consists of user rating style and item public praise. The new model is more actual and practical than the other methods.

Third, another focus of this thesis is on using finite mixture models to cluster large scale document data. A generalized method for unsupervised text clustering is presented. It integrates the mixture model's model selection, feature selection and parameter estimation into a general framework. Moreover, a modified version of "feature significance" is proposed such that the features' revalence to the mixture components is introduced to the mixture model as a set of latent variables and the component-relative features are selected when estimating the model's parameters. As an example of the generalized framework, a multinomial mixture model with feature selection is discussed in detail.

Fourth, this thesis use graph-based methods to deal with semi-supervised learning problems. The main idea is to investigate the similarities between data examples by defining some density-based distance over the graph. The inner structure information of the dataset is then obtained and utilized to compute the classifier. On semi-supervised classification, a *k*NN density-based distance form is presented to re-weight the graph, then the Laplacian kernel method is introduced to build classifiers over the whole feature space. On semi-supervised clustering, a density-based constraint expansion method is proposed. The constraint set is expanded by the similarity of the data samples. Then the expanded constraint set contains the manifold information of the dataset, and can be used in all semi-supervised clustering algorithms.

Finally, the main research contents are summarized at the end of the thesis with an expectation for future study and research.

KEY WORDS: Collaborative Filtering, Unsupervised Learning, Semi-supervised Classification, Semi-supervised Clustering, Finite Mixture Models

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和取得的研究成果，除了文中特别加以标注和致谢之处外，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 天津大学 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了谢意。

学位论文作者签名：张亮　　签字日期：2007 年 6 月 18 日

学位论文版权使用授权书

本学位论文作者完全了解 天津大学 有关保留、使用学位论文的规定。特授权 天津大学 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，并采用影印、缩印或扫描等复制手段保存、汇编以供查阅和借阅。同意学校向国家有关部门或机构送交论文的复印件和磁盘。

（保密的学位论文在解密后适用本授权说明）

学位论文作者签名：张亮　　导师签名：李秋红

签字日期：2007 年 6 月 18 日　　签字日期：2007 年 6 月 18 日

第一章 信息过滤和信息检索概述

在当今信息时代，Internet 和计算机技术的高速发展极大地加快了社会信息总量的增长以及信息的传输速度。在任何一种接入 Internet 的终端上，人们可以方便地获取大量信息，通过各种方式与他人进行高效、实时的通信交流而不受国家、地域的限制。

Internet 为人们带来了海量的信息资源和便捷的信息获取方式。但与此同时，以文本、音频、视频等多种形式存在的大量信息良莠不齐、菁芜并存。在浩如烟海的信息资源中挖掘和寻找有用的信息，往往如大海捞针，人们极易陷入无价值的"信息沼泽"中而迷失方向。

信息本身具有越来越重要的价值，有价值的信息已成为社会财富的一种新形式。从海量信息中快速、高效地获取人们真正需要的信息，已成为信息社会中的一个关键问题。解决这一问题主要包括两大类方法，即信息过滤和信息检索。本文研究与之相关的模型和算法。

1.1 研究背景与意义

近年来，Internet 在我国迅速发展，信息网络在人们的工作生活中具有越来越重要的地位。据中国互联网信息中心调查[1]显示，截至 2006 年 6 月，中国网民总人数已超过 1 亿 2 千万人，超过千万的网民经常使用基于 Internet 的各种服务，如网上招聘，网络教育，在线影视及网络游戏等。电子商务、电子政务、网上社区、论坛和博客等基于 Internet 的信息网络已成为许多人生活中不可或缺的一部分。

蓬勃发展的 Internet 在给人们带来丰富信息资源的同时也带来了一些新的问题：

- Internet 自身所容纳的巨大的信息量，远远超过了任何个人用户自然状态下的信息处理能力。如果不借助一些有效的辅助信息获取的工具，普通用户几乎不可能只通过 web 浏览器从数以亿计的页面中寻找到真正需要的信息。
- 虽然 Internet 上的海量信息资源包罗万象，涵盖了人类科技文化的各个方面，但各页面的信息质量参差不齐，冗余雷同、陈旧过时甚至是虚假的信息占据了大量页面，严重妨碍了人们获取高质量信息资源。
- 随着存储和传输技术的快速发展，Internet 上的主要应用已由传统的 Telnet、电子邮件、电子公告板、文本 web 页面等的文本信息转为容量更大、更直观的多种信息形式，包括音频、视频、flash 动画、java 小程序等。非文本格式

的更丰富的信息资源对传统的基于文本的信息搜索系统提出了更高的要求。

- 由于语言、文化、各国政策法规等原因，目前 Internet 信息分布存在一定程度的不平衡现象。英文仍是主流语言，Internet 上的中文文化科技信息近年来有了较大增长，但总体而言，目前仍不能与英文信息的数量和质量相媲美。
- 目前 Internet 信息网络发展的一个重要趋势是从传统的综合性网络新闻媒体到新兴的个人信息源的转变。博客（blog）、即时通讯工具、网上论坛和社区等的迅速兴起使得信息资源的分布相比以往更加分散，为有用信息的获取带来了新的挑战。
- Internet 的发展历史一直闪耀着自由精神的光辉，在极大地促进了人类知识传播的同时，不可避免地伴随着一些有害信息的传播。很多网页充斥着色情、暴力、迷信的信息，一些网站上侵犯知识产权的内容，以及某些对社会、个人有潜在危害的信息。在现阶段，这些仍是困扰 Internet 用户特别是青少年人群的重大问题。

随着 Internet 信息总量的不断增长和以上各种问题的出现，研究人员提出了各种信息过滤和信息检索技术，为人们在信息海洋中指引方向，使人们能够迅速获取有价值的信息。近年来，信息过滤和信息检索技术被广泛应用于各种搜索引擎、推荐系统、垃圾和不良信息过滤系统中，它们将信息与人们的需求相匹配，使人们可以更好地利用信息资源。

由于信息技术在各产业以及社会生活中所起的巨大推动作用，我国高度重视这一科技新领域的发展。“十五”期间，863 计划特别设置了信息获取与处理技术主题以及信息安全技术主题；共包括近百项课题。2006 年，中共中央办公厅、国务院办公厅印发了《2006—2020 年国家信息化发展战略》，其中特别指出我国信息化发展的战略重点之一是加强信息资源的开发利用。

信息过滤和信息检索技术包括多方面的内容，其核心问题是机器学习算法和模型的研究。本文的研究工作受国家自然科学基金（项目号：70171002 和 70571057）和高等学校博士学科点专项科研基金（基金号：20020056047）资助，研究信息检索和信息过滤的算法和模型，目的是准确地对数据建模，设计高效的过滤和检索算法，帮助用户得到最需要的信息。

1.2 与课题相关的研究主题

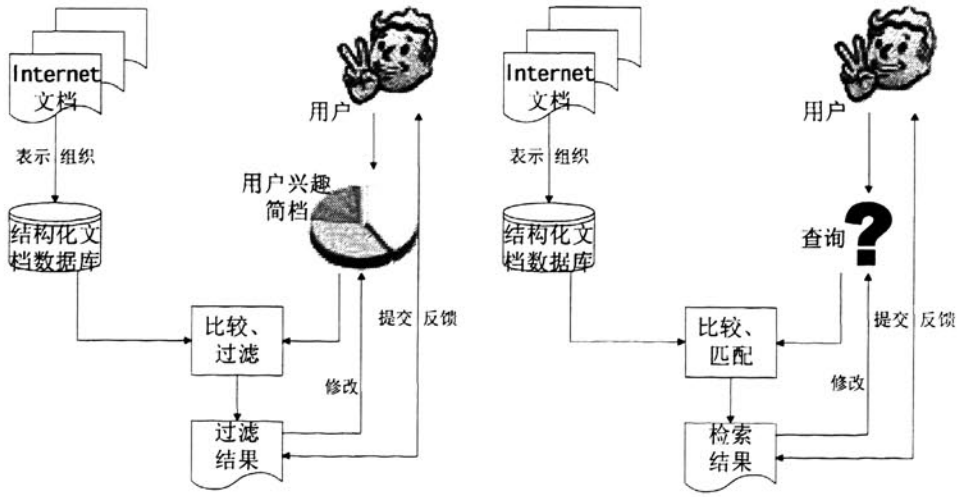
1.2.1 信息过滤和信息检索

信息过滤和检索技术分别起源于 20 世纪 70 年代和 60 年代，经过多年的发展，出现了很多成熟的模型和算法，并已成功地应用于各种商业搜索引擎、推荐系统和数字化图书馆。

信息过滤和信息检索技术的提出，是为了解决 Internet 信息过载问题，方便用户在海量信息资源中迅速高效地寻找需要的信息。信息过滤和检索在概念上非常相近，二者之间没有明显的界限，从根本上说都基于机器学习方法，但它们还是有区别的两种不同技术。据文献[2]给出的定义，信息过滤是指向用户传递需要的信息的一系列过程。信息检索是将信息按一定的方式组织和存储起来，并根据信息用户的需要找出有关信息的过程。它们的主要区别如表 1.1 所示：

表 1.1　信息过滤和信息检索的区别

	信息过滤	信息检索
信息源	动态的无结构或半结构化数据	静态的结构化数据
目标	将不相关的信息经过滤去除	检索数据得到相关信息
需求表示	用户兴趣偏好简档（profile）	用户的查询
个性化程度	高	低
典型应用	推荐系统	搜索引擎



（a）信息过滤　　（b）信息检索

图 1.1　信息过滤和信息检索概念示意图

为更好地说明信息过滤和信息检索的工作过程，图 1.1 分别给出它们的一般形式的概念示意图。从图中可以看出，信息过滤和信息检索的概念结构非常相似。它们都包括对 Internet 原始文档的表示和组织，并将它们转换为结构化文档，保存在数据库中。用户的需求分别以用户兴趣简档或查询的形式表示。通过将用户需求与数据库中文档相匹配或过滤的方式，系统将符合用户需求的文档提交给用户，

并根据用户反馈给系统的评价修改用户兴趣简档或查询条件。

尽管大部分信息过滤和信息检索系统的功能结构与图 1.1 相似，但用于文档表示、检索和过滤、相关性定义等的思路和技术各有不同，且处于不断的发展过程中。下面分别介绍信息过滤和信息检索的主要发展历史。

（1） **信息过滤**：1958 年，Luhn[3]提出了“商业智能机器”的概念，用于实现“有选择的信息分发”，这是最早的信息过滤设想。1982 年，Denning[4]以“信息过滤”正式命名这一概念，并提出了一个电子邮件的内容过滤器的例子。1987 年，Malone 等[5]发表了一种信息过滤系统“Information Lens”，并提出了三类信息过滤模式：认知过滤，经济过滤和社会过滤。其中认知过滤即基于内容的过滤，社会过滤即协同过滤，成为最常见的信息过滤技术。近期的信息过滤研究工作包括：对过滤算法的研究[6,7,8,9]，对获取用户知识、兴趣和行为模式的研究[10,11,12]等等。

（2） **信息检索**：“信息检索”这一概念最早由 Mooers 于 1951 年提出[13]。1960 年代信息检索得到了重要发展，Salton 和他的学生提出了向量空间检索模型，并开发了 SMART 系统[14]，成为多年来信息检索领域最常用的实验平台和后来许多成熟信息检索系统的原型。在此基础上，1970 和 1980 年代信息检索技术进一步发展，产生了概率检索模型[15]和一些大型商用数据库检索系统，如 Dialog 和 MEDLINE 等。1990 年代至今，随着 Internet 的快速发展，信息检索技术更多地侧重于网络搜索引擎方面的应用，出现了 Yahoo!，Google 等成功的商业搜索引擎。1992 年开始，NIST 主办的文本检索会议（TREC）的召开，使信息检索技术的发展面向大规模文本集合和知识获取等新型应用。自然语言处理和机器学习与信息检索技术的融合是近期信息检索的一个重要研究趋势。

1.2.2 协同过滤

信息过滤的两种主要形式是基于内容的过滤和协同过滤。协同过滤（collaborative filtering）也称社会过滤（sociological filtering），它的基本假设是经常访问相似资源的用户兴趣相似，相似兴趣的用户又会访问相似的资源。通过对相似兴趣用户的判定，可以确定某个用户对某一未知资源是否感兴趣。在协同过滤问题中，信息资源通常称为“项目”，用户的兴趣由他对项目的“评价”或称“评分”来体现，拥有相同或相似兴趣的用户通常被称为互为“邻居”。协同过滤的任务就是根据用户评价寻找他的邻居，并根据邻居对项目的评价向用户推荐他可能感兴趣的项目，因此协同过滤系统也称推荐系统。

与基于内容的信息过滤相比，协同过滤具有以下优点：

- 特别适于处理难以进行内容过滤的信息，如音乐、电影、图像等。

- 重视用户对信息资源的评价，人的智慧和品位成为信息过滤的主要依据，可以在一定程度上避免计算机对信息内容相关性判断的误差。
- 能够推荐用户潜在感兴趣的信息，这一能力被称为“意外好运推荐”（serendipitous recommendations）。

因为协同过滤有这些优点，近年来受到研究人员的重视和关注。Goldberg 等[16]提出的邮件过滤系统是最早应用协同过滤技术的信息过滤系统。随之出现了 Usenet 新闻的过滤系统 Grouplens[9]，电影推荐系统 MovieLens[9]，音乐推荐系统 Ringo[7]，等协同过滤的研究实验平台。协同过滤技术在 Internet 信息提供商和在线商业领域得到了广泛应用，在商业应用方面比较成功的协同过滤推荐系统包括 Amazon/joyo、ebay、Internet 电影数据库（IMDB）、国内的淘宝网、豆瓣网等。

尽管协同过滤技术已比较成功地应用在各种商业推荐系统中，但由于这项技术还处于不断的发展中，仍有许多问题需要解决。Sarwar 等[17]指出了当前协同过滤面临的两项挑战：

- 如何提高协同过滤算法的可扩展性。大型信息推荐系统面临着上千万用户的推荐需求，每个用户的邻居用户都有可能达到百万级。如何处理庞大的用户群，实时地向用户推荐感兴趣的信息，是所有推荐系统都需要解决的问题。
- 如何提高对用户推荐的信息的质量。用户需要可信赖的推荐来寻找他们需要的信息。如果推荐系统的推荐不准确，使用户丧失了信任感，并可能不再使用这一系统。这会进一步降低系统的推荐效果，形成恶性循环。

这两个挑战在某种程度上是相互矛盾的。已有的协同过滤方法主要可分为三类：全局协同过滤[9,16]，基于项目的协同过滤[17]和基于模型的协同过滤[18,19,20]。全局协同过滤和基于项目的协同过滤在对某个用户的推荐项目时需要在整个数据集上进行计算，得到的推荐结果相对较准，但系统的可扩展能力较差。基于模型的方法一般是由数据集计算得到用户对各项目的兴趣偏好模型，再根据模型计算推荐结果。采用这种方法时，系统的可扩展性较好，但如果建立的模型与用户的实际偏好不符，会造成推荐的项目不准确。出于可扩展性考虑，由传统的全局协同过滤方式转向采用机器学习方法进行建模，在保证一定推荐准确率的前提下进行基于模型的协同过滤，已经成为推荐系统实际应用中的必然选择。

1.2.3 机器学习方法在信息检索中的应用

机器学习在信息检索和信息过滤领域有着广泛而重要的应用，而且在这两个相关领域上机器学习方法的研究内容有高度相似性和重合性。信息检索中的机器学习方法是本文的一项重点研究内容。

机器学习的研究内容是计算机“学习”的理论和方法，根据真实问题观测得到的数据集设计相应的模型和算法，使得方法能够在一定程度上近似反映真实问题的

性质，并能通过拟和能自动提高自身性能。常见的机器学习方法包括规则学习、基于统计的学习、神经网络、遗传算法、分析学习等。借助机器学习算法可以实现信息检索过程的自动化，如对文档建模和表示、对文档聚类和分类、对用户建模等。这些机器学习方法在信息检索中的应用使检索过程更加准确、高效。

信息检索的基本对象是文档，最常见的文档模型是"bag of words"即词向量模型（term vector model）[14,15]。文档集合由 n 维空间中的若干数据点组成，每个数据点即一个文档，由一个 n 维词向量表示，n 是词汇表的数量。一个词向量或者是代表在文档中出现的词的布尔值，或者是代表词的出现次数的数值（词频）。直观上，文档中某个词的出现次数越多，它与文章主题的关系越密切。词向量模型不考虑词在文档中的出现顺序，一些没有特殊意义且出现频率很高的词被作为停用词排除在词汇表之外。文档间的相关关系由文档向量间的距离决定，在计算距离时，词的权重通常以 Salton 等提出的"词频倒排文档频"（TFIDF）方法[14]确定，综合考虑了词频和词对不同文档的分辨能力。文档间的相似性度量一般采用欧氏度量或将文档向量单位化后采用余弦相似性度量。

文本信息检索的两个主要难点是特征的高维性与稀疏性。采用词向量模型表示的文档集往往伴随着一个高维词汇表，文档向量在词汇表上的分布非常稀疏。这就是机器学习中常见的"维数灾难"（curse of dimensionality）[21]在信息检索问题中的表现，它严重影响着聚类和分类算法的效率和性能。解决的方法大体可分为两类：特征选择和特征抽取（feature extraction）。特征选择（feature selection）是一种从数据集的大量特征中选择其中一部分具有类别区分能力的特征集的机器学习方法。特征抽取则构造从原始特征空间到低维空间的一个变换，将原始特征空间中包含的信息转移到低维空间中。注意机器学习问题中的"样本"和"特征"分别对应于信息检索问题中的"文档"和"词"。研究人员提出了多种用于文本的特征选择和特征抽取方法。特征选择方法主要包括文档频率、信息增益、互信息、CHI^2 统计、基于 PCA 的特征选择、基于边界的特征选择等。此外，Kohavi[22]的 wrapper 模型将特征选择与学习过程结合起来，以分类器性能作为特征选择算法的评价。特征抽取方法主要包括主成分分析（PCA）[23]，线性判别分析（LDA）[24]和独立成分分析（ICA）[25]等。

机器学习方法常用来对数据集进行回归分析、聚类和分类。在信息检索领域，机器学习方法最常见的应用是对文本的聚类和分类。下面分别介绍信息检索中的聚类和分类方法：

- 聚类方法根据样本间的某种距离将它们划分为内部具有高相似度的若干类，是一种无监督学习方法。利用聚类方法可以把大量文档划分成用户可迅速理解的聚集（cluster），使用户可以更快地把握大量文档中的信息内容，加快分析速度和辅助决策。常见的聚类方法包括层次方法[26,27]，划分（patitioning）方法[28]，基于网格（grid）的方法[29,30]，基于密度的方法[31]，基于模型的方

法[32]，以及神经网络和遗传算法等。

- 分类方法根据事先给定的分类体系和已标注好类别的训练样本，将文档划分到若干类，是一种有监督学习方法。依照文档的内容、主题或其他属性进行分类并加以索引，可以帮助用户更方便地查找需要的信息。除此之外，分类方法还可以被应用于区分垃圾邮件、有害信息等，因此在信息过滤问题中也有重要的应用价值。常见的文本分类方法包括决策树和规则方法[33]，最近邻方法，Naïve Bayes 方法，神经网络，支持向量机（SVM）以及 Boosting 方法[34]等。

1.2.4 有监督学习、无监督学习和半监督学习

上一节介绍了机器学习方法在信息检索问题中的应用，本节介绍如题目所示的三类基于统计的机器学习方法。

统计机器学习方法的提出源于 1969 年 Minsky 和 Papert 的著作《感知机》[35]，但它一直被用来对给定数据的拟和函数进行理论分析，直到 1990 年代中期才在实际应用中取得了突破性成就。Vapnik 于 1992 年开始提出了有限样本统计理论，并介绍了一种新算法——支持向量机（support vector machine，SVM）[36]。支持向量机在许多领域的成功应用促进了许多新型模型和算法的发展，从此统计机器学习不再只是一种理论分析的工具，还可以用于指导解决数据挖掘和信息检索领域中的很多实际问题。

通常假定用于学习的数据集是观测到的样本集。在某些机器学习问题中，已知部分数据样本的标记信息，常常是由专家（监督者）手工标记得到，称为监督信息。有标记的样本中的一部分通常被用于学习器的训练过程，因此也称训练集。在对学习得到的模型进行性能评价时，常用样本的实际标记与预测得到的类别标记进行，这部分有标记的样本被称为测试集或评价集。一般来说，统计机器学习方法的训练集和测试集应不存在交集。

根据已知的用于指导学习的监督信息的不同，统计机器学习主要可分为有监督学习、无监督学习和半监督学习。下面分别介绍本文的研究中涉及的这三种机器学习方法。

- 有监督学习是在已知一些有监督信息的条件下，对全部数据集建立模型的学习方法。常见的有监督学习方法包括回归和分类，它们的区别是回归方法得到的结果是连续值，而分类结果是离散的类标记值。目前，对有监督学习算法及其应用的研究集中于概率模型方法、神经网络、支持向量机等几个方面。本文将协同过滤问题作为一种特殊的有监督学习问题，所采用的研究方法是对数据集建立概率模型，对用户评价的预测可视为在所有可能的评价集上的分类结果。

- 无监督学习和有监督学习的最大区别是没有事先已知的监督信息，只根据样本点的自然结构学习相应的模型。聚类是最常见的无监督学习方法，本文对无监督学习的研究集中于采用概率模型对文本进行无监督聚类。在基于概率模型的无监督聚类中，样本点集通常作为随机变量的集合，学习器在全部数据集上学习一个联合密度模型。最常见的模型是 Gaussian 混合模型，相应的参数估计方法是期望－最大化（expedition-maximization，EM）算法[37]。根据具体问题的不同，还可以采用多项式混合、von-Mises 混合、polya 混合等多种形式的混合模型。
- 半监督学习方法介于有监督学习和无监督学习之间，同时使用少量的有监督信息和大量无监督信息进行训练。半监督学习可分为两类：半监督分类和半监督聚类。除分类和聚类方法的区别外，两者的另一项主要区别是有监督信息的形式不同：前者采用的有监督信息是有标记样本，后者则是样本间是否属于同一类的成对约束关系（must-link 和 cannot-link）。对应于具体问题的不同，可采用各种适合的半监督分类和聚类方法。半监督分类的主要方法包括生成性模型[38]，自训练（self-training）[39]，协同训练（co-training）[40]、转导支持向量机（transductive SVM）[41]以及最小截（min-cut）[42]等基于图形的方法。半监督聚类包括相似性适应方法[43,44]和基于搜索的方法[45,46]。前者利用各种已有的聚类方法，将约束直接转化为样本间相似性度量上；后者改造聚类算法本身，搜索满足约束条件的“合适的”聚类。

1.3 研究内容和论文结构

本文主要对信息过滤和信息检索中的机器学习方法进行了研究，内容主要分为三部分：协同过滤方法、文本聚类和半监督学习。全文共分六章。

第一章介绍本文的研究背景、信息过滤和信息检索的概念、发展和研究现状。重点介绍了在信息过滤和信息检索中应用的各种机器学习方法，包括有监督学习、无监督学习和半监督学习。在此基础上，提出了本文的主要研究内容。

第二章主要讨论本文中应用的几种统计机器学习方法，包括混合模型和 EM 算法、图模型方法和支持向量机等。重点是对各种模型和算法的概念、应用环境、具体步骤和效果进行详细说明。这一章作为下面各章研究内容的理论基础而提出。

第三章研究基于概率模型的协同过滤方法，对现有各种协同过滤方法进行了分析，提出了一种同时考虑项目公众评价和用户偏好的协同过滤新模型，称为“真实偏好高斯混合模型”。

第四章研究文本数据集的无监督学习问题。提出了一种适应于无监督文本聚类、并结合模型选择和特征选择的广义方法，通过实验证明了该方法的有效性。

第五章的主要研究内容是半监督学习算法。对半监督分类问题，提出了一种基于密度的 Laplacian 核方法；对半监督聚类问题，提出了一种基于密度的约束扩展方法。

第六章是全文的总结，并对待研究的问题进行了展望。

整个论文的结构如图 1.2。

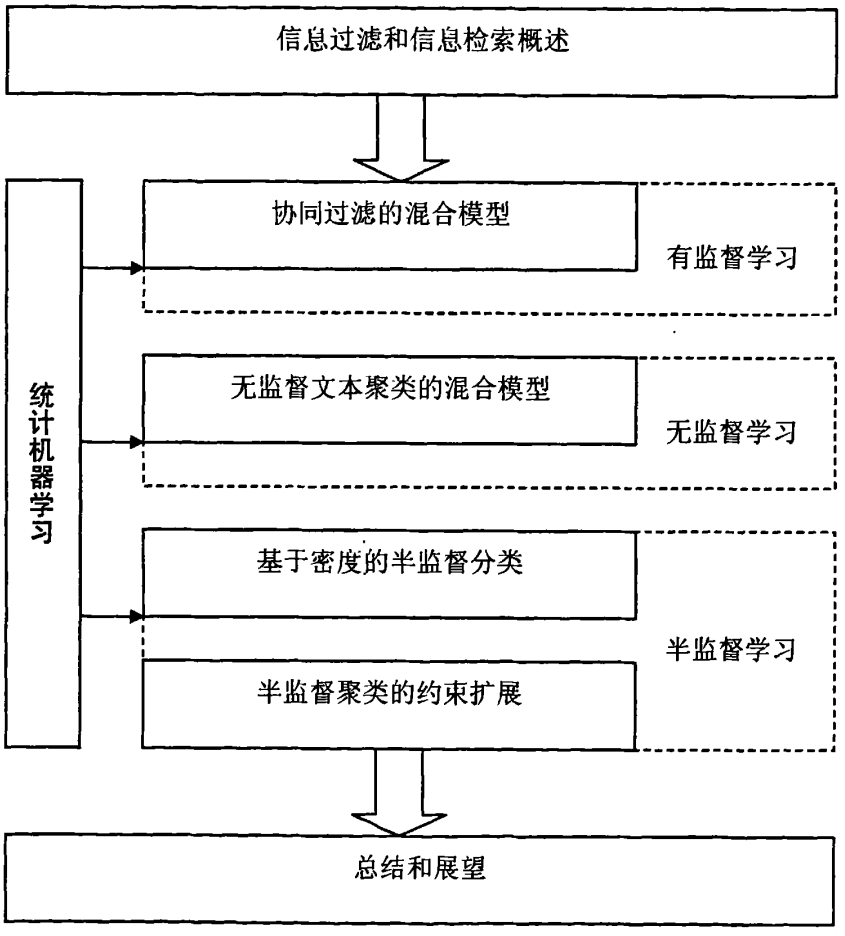


图 1.2 论文研究内容框架

1.4 论文创新点

论文的创新点主要体现在以下几个方面：

- 对协同过滤问题，认为传统方法存在两点不足：首先，没有区分用户偏好类型与项目主题类型的差异；其次，忽略了用户评分习惯以及项目公众评价对用户－项目最终评分的影响。提出了应用于协同过滤的一种概率模型，称为真实偏好高斯混合模型。模型中考虑了用户评分习惯以及项目的公众评价对用户－项目最终评价的综合影响。
- 对大规模文本数据聚类问题，认为基于概率模型的方法存在着两个主要问题：第一，在没有类标记的情况下，如何选择适合模型学习的相关特征集。第二，各种有限混合模型以及模型选择方法的应用，尚缺少有指导意义的一般性方法。对第一个问题，定义了一种改进的"特征显著性"方法，将特征对各混合成员的相关性作为隐变量引入混合模型，在估计模型参数的同时完成特征选择。对第二个问题，提出了用有限混合模型进行无监督文本聚类的一种广义方法，它将模型选择，特征选择以及混合模型的参数估计纳入一个统一的框架中。
- 对半监督分类问题，提出了一种基于密度的 Laplacian 核方法。在这种方法中，定义了一种基于密度的 *k*NN 距离来反映数据点间的距离，并在此基础上以一种 Laplacian 核方法来构造整个数据集上的超分类面，突破了传统基于图的机器学习方法的"转导"（只能处理现有数据点，无法处理新数据样本）限制。
- 对半监督聚类问题，认为现有方法很少利用数据集空间结构信息，有限的约束条件限制了聚类算法的性能。对这一问题，提出了一种基于密度的约束扩展方法 DCE。将数据集以图的形式表达，定义了一种基于密度的图形相似度。根据样本点间的距离和相似度关系，对已知约束集进行扩展，扩展后的约束集可用于各种半监督聚类算法。

第二章 统计机器学习

统计学和计算机科学在过去的几十年里一直各自走着不同的研究道路，虽然互有交流和帮助，但这两大领域的核心问题很少有交叉。然而，自 1990 年代以来，随着支持向量机在许多领域的成功应用，统计机器学习方法受到了计算机科学研究人员的高度重视。特别是在知识挖掘和信息获取领域，根据具体问题的不同，出现了各种形式的机器学习模型和算法。本章介绍几种信息过滤和信息检索领域中常见的统计机器学习方法，重点是问题的定义、模型的概念结构和算法步骤，并以此作为进一步研究的理论基础。

2.1 有限混合模型和 EM 算法

有限混合模型[47,48,49,50]是一种常见的概率模型，它对单变量和多变量数据具有强大的统计建模能力，被广泛应用于模式识别、计算机视觉、信号和图象分析、文本建模和聚类等多种领域。有限混合模型假定观测样本是由一系列随机源之一产生的，学习的过程就是推导模型的参数及判断产生样本的源。对应于聚类问题，有限混合模型的建模和推导就是求解样本所在类的属性和类标记的过程。与启发式聚类方法（如 *k*-means）或者层次聚类方法（汇聚聚类和划分聚类）相比，有限混合模型聚类方法可以将如何选择聚类数量以及如何判别模型有效性表示为一种规范的形式[32]。

使有限混合模型与观测数据相符合的一种常用方法是期望－最大化（expectation-maximization，EM）算法[37]，这种算法被用来估计混合模型参数。EM 算法特别适于对含有无法观测的隐变量的模型的参数估计问题，是处理这一类型问题的一种经典算法。然而，EM 算法本身也存在一些缺陷：它是一种贪心算法，只能达到局部最优；对初始值敏感；以及需要预先确定混合成员的数量。如何选择混合成员的数量通常被称为有限混合模型的“模型选择”问题，它是模型准确度与复杂度的一种折衷：过多的成员会使模型过于复杂，产生“过拟和”（over-fitting）问题；过少的成员使模型无法准确反映真实情况，称为“不足拟和”（under-fitting）问题。

下面对有限混合模型和 EM 算法进行简要介绍。

2.1.1 有限混合模型

首先给出有限混合模型的定义：

记 $\mathbf{X}=(X_1,...,X_d)^T$ 为一个 d 维随机变量，$\mathbf{x}=(x_1,...,x_d)^T$ 为 $\mathbf{X}$ 的一个观测样本。称 $\mathbf{X}$ 服从于一个成员数为 k 的有限混合分布，如果它的概率密度函数为：

$$p(\mathbf{x}\mid\boldsymbol{\theta})=\sum_{m=1}^{k}\alpha_m p(\mathbf{x}\mid\boldsymbol{\theta}_m) \tag{2.1}$$

其中 $\alpha_1,...,\alpha_k$ 是混合概率，假定混合模型可表示为某种函数的形式，每个向量 $\boldsymbol{\theta}_m$ 是对应于第 m 个成员的参数集。$\boldsymbol{\theta}=\{\boldsymbol{\theta}_1,...,\boldsymbol{\theta}_k,\alpha_1,...,\alpha_k\}$ 是混合模型参数的完全集。由于 α_m 是概率，因此必须满足：

$$\alpha_m\geq 0,\ \ m=1,...,k,\ \ 且\sum_{m=1}^{k}\alpha_m=1 \tag{2.2}$$

在对有限混合模型进行参数估计时，通常是根据对数似然函数求它的最大似然估计或最大后验估计：

给定 n 个独立同分布的样本集合 $\mathcal{X}=\{\mathbf{x}^{(1)},...,\mathbf{x}^{(n)}\}$，$k$ 成员混合模型的对数似然函数为：

$$\log p(\mathcal{X}\mid\boldsymbol{\theta})=\log\prod_{i=1}^{n}p(\mathbf{x}^{(i)}\mid\boldsymbol{\theta})=\sum_{i=1}^{n}\log\sum_{m=1}^{k}\alpha_m p(\mathbf{x}^{(i)}\mid\boldsymbol{\theta}_m) \tag{2.3}$$

混合模型的最大似然（maximum likelihood，ML）估计为：

$$\hat{\boldsymbol{\theta}}_{\mathrm{ML}}=\arg\max_{\boldsymbol{\theta}}\{\log p(\mathcal{X}\mid\boldsymbol{\theta})\} \tag{2.4}$$

混合模型的最大后验（maximum a posteriori，MAP）估计为：

$$\hat{\boldsymbol{\theta}}_{\mathrm{MAP}}=\arg\max_{\boldsymbol{\theta}}\{\log p(\mathcal{X}\mid\boldsymbol{\theta})+\log p(\boldsymbol{\theta})\} \tag{2.5}$$

其中 $p(\boldsymbol{\theta})$ 是参数的先验概率。在对有限混合模型进行参数估计时，它的最大似然和最大后验都必须满足公式（2.2）的约束条件。

2.1.2 EM 算法

对有限混合模型进行 ML 或 MAP 估计的常见方法是期望一最大化（EM）算法[37,49]。EM 算法是包括期望步和最大化步的一系列迭代过程，可以得到 $\log p(\mathcal{X}\mid\boldsymbol{\theta})$ 或 $[\log p(\mathcal{X}\mid\boldsymbol{\theta})+\log p(\boldsymbol{\theta})]$ 的局部最大值。文献[50,51]详细研究了 Gaussian 混合模型中 EM 算法的收敛性。文献[52]证明了 EM 算法属于一类近似点算法（proximal point algorithm，PPA），可采用 PPA 方法对 EM 的迭代过程进行加速。

EM 算法用于对不完全数据进行参数估计。对于有限混合模型的数据集 $\mathcal{X}$，缺失数据为各数据点的类标记 $\mathcal{C}=\{\mathbf{c}^{(1)},...,\mathbf{c}^{(n)}\}$，表示每个样本由哪个混合成员生成。每个标记都是一个 0/1 指示向量 $\mathbf{c}^{(i)}=(c_1^{(i)},...,c_k^{(i)})$，其中的一个分量 $c_m^{(i)}=1$，其他均为 0，表示样本 $\mathbf{x}^{(i)}$ 由第 m 个混合成员生成。将 $\mathcal{Y}=\{\mathcal{X},\mathcal{C}\}$ 作为完全数据集，模型

的对数似然函数为：

$$\log p(\mathcal{X},\mathcal{C}\mid\boldsymbol{\theta})=\sum_{i=1}^{n}\sum_{m=1}^{k}c_m^{(i)}\log\left[\alpha_m p(\mathbf{x}^{(i)}\mid\boldsymbol{\theta}_m)\right] \tag{2.6}$$

迭代运行 EM 算法的两个步骤，产生一系列参数的估计值$\{\hat{\boldsymbol{\theta}}(t),t=0,1,2,...\}$，直到达到收敛条件。下面给出 EM 算法的描述。

算法 2.1 对有限混合模型进行参数估计的 EM 算法

输入：数据集$\mathcal{X}=\{\mathbf{x}^{(1)},...,\mathbf{x}^{(n)}\}$

输出：参数估计值$\{\hat{\boldsymbol{\theta}}(t),t=0,1,2,...\}$

步骤：运行以下期望步和最大化步直至收敛。

期望步（E-step）：根据当前参数估计值$\hat{\boldsymbol{\theta}}(t)$，以及观测数据$\mathcal{X}$和隐藏数据$\mathcal{C}$计算完全数据集对数似然函数$\log p(\mathcal{X},\mathcal{C}\mid\boldsymbol{\theta})$的条件期望。由于

$$p(\mathcal{X},\mathcal{C}\mid\boldsymbol{\theta})=p(\mathcal{C}\mid\mathcal{X},\boldsymbol{\theta})p(\mathcal{X}\mid\boldsymbol{\theta}) \tag{2.7}$$

其中$p(\mathcal{X}\mid\boldsymbol{\theta})$是常量。因此将$\mathcal{C}$视为随机变量，$\mathcal{C}$和$\hat{\boldsymbol{\theta}}(t)$作为已知常量，计算条件期望$\mathcal{W}=E[\mathcal{C}\mid\mathcal{X},\hat{\boldsymbol{\theta}}(t)]$，并将它加入$\log p(\mathcal{X},\mathcal{C}\mid\boldsymbol{\theta})$，得到如下形式的$Q$函数：

$$Q(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}(t))=E\left[\log p(\mathcal{X},\mathcal{C}\mid\boldsymbol{\theta})\mid\mathcal{X},\hat{\boldsymbol{\theta}}(t)\right]=\log p(\mathcal{Y},\mathcal{W}\mid\boldsymbol{\theta}) \tag{2.8}$$

由于$\mathcal{C}$是 0/1 向量，它的各分量的条件期望为：

$$w_m^{(i)}=E\left[c_m^{(i)}\mid\mathcal{X},\hat{\boldsymbol{\theta}}(t)\right]=P(c_m^{(i)}=1\mid\mathbf{x}^{(i)},\hat{\boldsymbol{\theta}}(t))=\frac{\hat{\alpha}_m p(\mathbf{x}^{(i)}\mid\hat{\boldsymbol{\theta}}_m(t))}{\sum_{j=1}^{k}\hat{\alpha}_j p(\mathbf{x}^{(i)}\mid\hat{\boldsymbol{\theta}}_j(t))} \tag{2.9}$$

公式（2.9）由 Bayes 定理得到。其中α_m是$c_m^{(i)}=1$的先验概率，$w_m^{(i)}$是已知观测数据$\mathbf{x}^{(i)}$的条件下$c_m^{(i)}=1$的后验概率。

最大化步（M-step）：根据如下公式更新参数的估计值：

最大后验估计：

$$\hat{\boldsymbol{\theta}}(t+1)=\arg\max_{\boldsymbol{\theta}}\{Q(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}(t))+\log p(\boldsymbol{\theta})\} \tag{2.10}$$

最大似然估计：

$$\hat{\boldsymbol{\theta}}(t+1)=\arg\max_{\boldsymbol{\theta}}Q(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}(t)) \tag{2.11}$$

算法结束

应用 EM 算法对混合模型进行参数估计面临的两个主要问题是：1，混合成员数量k的确定；2，各参数的初始化。对第一个问题，主要通过某种模型选择条件$\mathrm{Cr}(\hat{\boldsymbol{\theta}}(k),k)$确定，其中$\hat{\boldsymbol{\theta}}(k)$是成员个数为$k$时模型的参数。常见的准则包括 Bayes

信息准则[53,54,55,56]，Akaike 信息准则[57]等。对第二个问题，常见的方法包括采用多个随机初始点并取相应似然函数值的最大值的方法[58]，先经聚类得到初始点的方法[50,59]，以及模拟退火方法[60,61]等。

2.2 图模型

图模型是以有向图或无向图的形式定义的一种概率模型，是图论与概率论的结合。图中的节点表示随机变量，节点之间的连接关系表示它们的联合/条件概率分布函数。图模型方法在设计机器学习算法以及对其中的不确定性和复杂性进行分析时起着越来越重要的作用。

统计学、系统工程、信息论和模式识别等领域中研究的大部分多变量概率系统都可以视为一般形式的图模型的特例[62]。在统一的图模型框架下，一个领域中的特殊技术可以方便地转化至另一个领域中，使各种机器学习技术得到更广泛的应用。本节介绍一般形式的图模型定义，以及信息过滤和信息检索中常见的层次 Bayesian 图模型。

2.2.1 图模型介绍

图模型的两种主要形式是有向图模型和无向图模型，它们的主要区别是图中的边是否有方向性。在表示概率模型时，通常采用有向无环的图模型。

图 2.1 最简单的图模型，左为有向图，右为无向图

记 $G=(V,E)$ 为一个图，其中 V 是节点集合，E 是边集。图中的一条边由一对节点构成，可以是有向的或无向的。图 2.1 是两个最简单的图模型的例子，只有两个节点 A、B 和连接它们的一条边。其中左图为有向图，表示条件概率，有向边无箭头的方向表示条件，有箭头的方向表示原因；右图为无向图，表示事件的联合概率。

图 2.2 是一个比图 2.1 稍微复杂的图模型，根据图中各节点之间的条件概率关系，可知它们的联合概率关系为：

$$P(A,B,C,D)=P(D\mid C)P(C\mid A,B)P(A)P(B) \qquad (2.12)$$

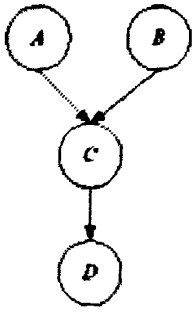


图 2.2 一个稍微复杂的图模型

图模型中常用"面板"（plate）来表示多个结构和性质相同的节点。图 2.3 是面板的一个例子，左图中的面板中的$\{Z_k\}_1^K$表示右图中的K个节点。面板常用来表示若干独立同分布的向量。

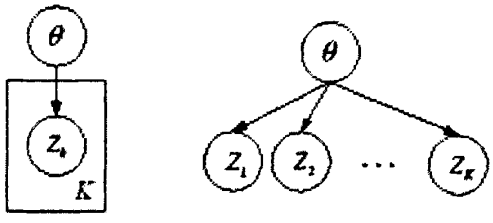


图 2.3 图模型中的面板：左图等价于右图

2.2.2 层次 Bayesian 模型

在信息过滤和信息检索任务中，基于 Bayesian 统计的学习是最常见的机器学习方法，在信息过滤和信息检索问题的各个方面都有重要的应用。首先，信息检索系统可视为对不确定的用户需求的学习过程，Bayesian 层次模型可以表现这种不确定知识，并具有丰富的概率表达能力，特别适于根据关于用户的先验信息学习用户需求。其次，文档表示过程中涉及到的分类和聚类方法中也常常采用层次 Bayesian 模型来对文档建模。第三，在大部分信息过滤和信息检索系统的性能评价和分析问题中，都可以用基于 Bayesian 层次模型的决策论方法来对模型的复杂性和表达能力加以平衡。

信息过滤和信息检索中的重要任务之一是对文档建模。应用层次 Bayesian 模型对文本建模时，通常采用 Salton 等提出的"bag-of-words"模型。这一模型将文档中的词作为随机变量，只考虑词的出现次数，而不考虑词在文档中的位置信息，这主要是受庞大的文本数据集带来的计算量的限制。

文本模型中的一个重要概念是文档的"主题"，它是指文档主要包括哪些方面的内容，例如新闻文章中的财经、政治、文化、科技、体育等主题分类。主题之间

可能存在某种层次关系，如科技可以包括生命科学、能源技术、信息技术等若干方面。同一文档也可以同时涵盖多个不同的主题。在文本的层次 Bayesian 模型中，文档主题常作为隐变量出现。这样，文档的聚类和分类等学习方法就转化为概率模型中隐变量的参数估计问题，可采用相应的机器学习方法实现。

下面介绍一种著名的层次 Bayesian 模型：Hoffman 提出的概率潜在语义分析（probabilistic latent semantic analysis，PLSA）[63,64]。

首先给出数据集的定义。假定文档集是 $D=\{d_1,...,d_N\}$，词汇表为 $W=\{w_1,...,w_M\}$。忽略词在文档中的位置和出现顺序，数据集可表示为一个 $N\times M$ 矩阵 $\mathbf{N}=(n(d_i,w_j))_{ij}$，其中 $n(d,w)\in\mathbf{N}$ 表示词 w 在文档 d 中的出现次数。$\mathbf{N}$ 的行向量就是"bag-of-words"文档向量。

PLSA 是一种分析方法，它对应的模型称为相位模型（aspect model），将它们统称为 PLSA 模型方法。这种模型中，每个观测样本（d 或 w）都对应着一个未观测的分类隐变量 $z\in Z=\{z_1,...,z_K\}$。定义 $D\times W$ 上的联合概率模型为如下混合：

$$P(d,w)=P(d)P(w|d)=\sum_{z\in Z}P(w|z)P(z|d)P(d) \tag{2.13}$$

考虑到潜在变量的数量 K 通常比 M 或 N 小得多，相位模型还可以表示为另一种形式：

$$P(d,w)=\sum_{z\in Z}P(z)P(d|z)P(w|z) \tag{2.14}$$

这两种形式的图模型描述如图 2.4 所示。

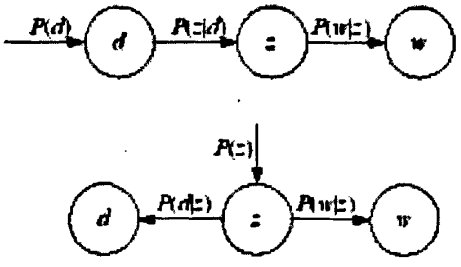


图 2.4 PLSA（相位模型）的图模型

上图对应于公式（2.13），下图对应于公式（2.14）

模型的学习过程称为 PLSA 方法，通过 EM 算法迭代估计模型的最大似然函数，进行参数估计。算法的期望步为：

$$P(z|d,w)=\frac{P(z)P(d|z)P(w|z)}{\sum_{z'\in Z}P(z')P(d|z')P(w|z')} \tag{2.15}$$

算法的最大化步为：

$$P(w|z) \propto \sum_{d \in D} n(d,w) P(z|d,w) \tag{2.16}$$

$$P(d|z) \propto \sum_{w \in W} n(d,w) P(z|d,w) \tag{2.17}$$

$$P(z) \propto \sum_{d \in D} \sum_{w \in W} n(d,w) P(z|d,w) \tag{2.18}$$

除了 PLSA 方法，另一种著名的层次 Bayesian 图模型是 Blei 等提出的潜在 Dirichlet 分配（latent Dirichlet allocation，LDA）[62,65]。这种模型与 PLSA 相似，区别是在 LDA 中假定主题的分布与一个 Dirichlet 先验有关。在信息检索和协同过滤问题中 PLSA 和 LDA 都有突出的表现[62]。

2.3 支持向量机

Vapnik 提出的统计学习理论[66]的基本思想是结构化风险最小化原理，具有更好的泛化能力，在理论上要优于传统的经验风险最小化原理。建立在 VC 维理论和结构风险最小化原理上的支持向量机（support vector machine，SVM）方法[36]，已在模式识别、回归问题、分类问题、预测和评价等许多领域得到了成功应用。本节介绍 SVM 方法的基本形式和一种适于半监督学习的 TSVM 方法[41]。

2.3.1 SVM

SVM 的提出是为了解决特征空间上数据点的二值分类问题。对一个学习任务 $P(\mathbf{x},y)=P(y|\mathbf{x})P(\mathbf{x})$，SVM 学习器的任务是建立一个决策函数 $f:\mathbf{x}\to\{-1,+1\}$，$f=L(S_{train})$，其中 $S_{train}=\{(\mathbf{x}_i,y_i)\}_1^L$ 是包含 L 个样本的训练集。

根据结构风险最小化原理，最小化如下分类决策函数的经验风险：

$$R=\frac{1}{l}\sum_{i=1}^{L}|f(\mathbf{x}_i)-y_i| \tag{2.19}$$

其中 l 是归一化常数。SVM 计算具有最大分类间隔的最优分类超平面，并允许一定的分类误差。图 2.5 显示了在一个二值平面分类问题中，支持向量机计算得到的最优分类超平面。其中决定分类边界的样本点被称为支持向量。

分类超平面等分两个类的凸包（convex hull）间最短距离，可表示为以下约束最小化问题：

$$\begin{aligned} &Minimize: \frac{1}{2}\mathbf{w}^T\mathbf{w} \\ &Subject\ to: y_i(\mathbf{w}\mathbf{x}+b)\geq 1 \end{aligned} \tag{2.20}$$

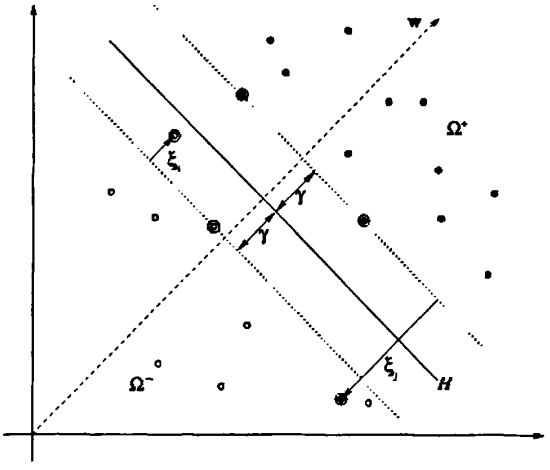


图 2.5 支持向量机计算的最优分类超平面

对线性不可分问题，可以引入松弛变量$\{\xi_i\}_1^L$，允许一定的分类误差。公式(2.20)变为：

$$\begin{aligned} &Minimize: \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{L}\xi_i \\ &Subject\ to: y_i(\mathbf{w}\cdot\varphi(\mathbf{x}_i)+b) \geq 1-\xi_i, \xi_i \geq 0 \end{aligned} \tag{2.21}$$

其中参数 C 用来调整误差惩罚参数$\{\xi_i\}_1^L$，$\varphi(\cdot)$是一个非线性函数，作用是将原始特征空间映射到一个高维空间。公式（2.21）中的最小化问题相当于最小化学习器的 VC 维，其约束条件控制了经验风险。

引入拉格朗日乘子$\{\alpha_i\}_1^L$，并定义

$$\mathbf{w}(\alpha) = \sum_{i=1}^{L}\alpha_i y_i \mathbf{x}_i \tag{2.22}$$

公式（2.21）的对偶问题可表示为：

$$\begin{aligned} &Maximum: \sum_{i=1}^{L}\alpha_i - \frac{1}{2}\mathbf{w}^T(\alpha)\cdot\mathbf{w}(\alpha) \\ &subject\ to: \alpha_i \geq 0,\ \sum_{i=1}^{L}\alpha_i y_i = 0 \end{aligned} \tag{2.23}$$

公式（2.23）的求解是一个约束二次优化问题，可采用采用序列最小优化（sequential minimal optimization，SMO）[67]等优化算法来解决。

2.3.2 TSVM

TSVM[41]与 SVM 的最大区别是在学习过程中利用了测试集$S_{test} = \{\mathbf{x}_i^*\}_1^U$的信息，是一种半监督学习方法。TSVM 学习器的决策函数为$f = L(S_{train}, S_{test})$，其中$S_{train} = \{(\mathbf{x}_i, y_i)\}_1^L$是包含 L 个样本的训练集。

对线性可分问题，分类超平面同时对训练集和测试集求解最大划分边界：

$$\begin{aligned} &Minimize\ Over\ (y_1^*, y_2^*, ..., y_U^*, \mathbf{w}, b): \frac{1}{2}\mathbf{w}^T\mathbf{w} \\ &Subject\ to: y_i(\mathbf{w}\mathbf{x}+b) \geq 1 \\ &\qquad y_j^*(\mathbf{w}\mathbf{x}_j^*+b) \geq 1 \end{aligned} \tag{2.24}$$

对于线性不可分的数据，类似 SVM，TSVM 可表示为如下优化问题：

$$\begin{aligned} &Minimize\ Over\ (y_1^*, y_2^*, ..., y_U^*, \mathbf{w}, b, \xi_1, ..., \xi_L, \xi_1^*, ..., \xi_U^*): \\ &\qquad \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{L}\xi_i + C^*\sum_{j=1}^{U}\xi_j^* \\ &Subject\ to: y_i(\mathbf{w}\cdot\varphi(\mathbf{x}_i)+b) \geq 1-\xi_i, \xi_i \geq 0 \\ &\qquad y_j^*(\mathbf{w}\varphi(\mathbf{x}_j^*)+b) \geq 1-\xi_j^*, \xi_j^* \geq 0 \end{aligned} \tag{2.25}$$

其中 C^*是训练集样本的调整因子，与 C 类似。以下是 TSVM 的训练算法：

算法 2.2 转导支持向量机（TSVM）

输入：训练集 $S_{train} = \{(\mathbf{x}_i, y_i)\}_1^L$，测试集 $S_{test} = \{\mathbf{x}_i^*\}_1^U$，参数 C 和 C^*，测试样本中正例估计值 num_+

输出：预测测试样本的预测标记 $y_1^*, y_2^*, ..., y_U^*$

步骤：1，用标准 SVM 算法对有标记样本进行初始学习，得到初始分类器。

2，用初始分类器对无标记样本分类，对判别函数值输出值最大的 num_+个无标记样本取正标记，其余的取负。并指定一个临时影响因子 C_{tmp}^*。

3，对所有样本重新训练，对新得到的分类器，按一定的规则交换一对标记值不同的测试样本的标记值，使得优化问题（2.25）中的目标函数值获得最大下降。反复执行这一步骤，直到找不出符合交换条件的样本对为止。

4，均匀地增加临时影响因子 C_{tmp}^* 的值并返回到步骤 3，当 $C_{tmp}^* \geq C^*$ 时停止，输出结果。

算法结束

2.4 本章小结

本章介绍了信息过滤和信息检索领域常见的几种统计学习方法，包括有限混合模型和 EM 算法、图模型机器学习方法和层次 Bayesian 模型，以及支持向量机。较为详细地介绍了这些方法的问题定义、模型结构和算法步骤等。

第三章 协同过滤的混合模型

信息过滤中的一项重要技术是协同过滤，它能根据用户的兴趣偏好作出个性化的信息推荐，在电子商务、科技文献、网页检索等领域得到了广泛的应用。协同过滤的基本任务是根据相似的偏好匹配用户，以推荐用户可能会喜欢的项目。本章首先分析了协同过滤传统方法存在的不足，然后提出了应用于协同过滤的一种概率模型，称为真实偏好高斯混合模型。新模型的突出特点是考虑了用户评分习惯以及项目的公众评价对用户－项目最终评价的综合影响。

3.1 协同过滤概述

与信息检索系统不同，推荐系统不由用户查询驱动检索信息，而是主动向用户提供推荐。推荐系统因其可以针对每个用户提供个性化的服务而被广泛应用，它采用的技术包括基于内容的过滤和协同过滤。

基于内容的过滤采用与传统信息检索相似的技术。用户需要清晰、明确地定义和描述他的兴趣偏好，推荐系统将项目的特征与用户兴趣资料比较，将最相关的项目推荐给用户。基于内容的过滤在实际应用中受到一些限制：第一，难于获得用户兴趣资料的准确表达形式。在电子商务应用等许多情况下，用户凭感觉选择商品，很难说清自己的兴趣。第二，对于文本信息缺乏的项目（如音乐、电影、视频片段等），目前尚无令人满意的基于内容的分析方法。第三，基于内容的方法能推荐与用户以往选择相似的项目，但无法找到与以往选择不同但有潜在新意的项目。

作为基于内容过滤的补充，协同过滤技术使用记录用户偏好、行为或表现的数据库，根据其他用户对项目的评价预测某一用户的兴趣。协同过滤不需要项目的具体内容信息，只依靠用户对项目的评价数据。协同过滤的基本思想是，要查找某个用户感兴趣的内容，可以先找到和他有相似兴趣的其他人，然后推荐这些人喜欢的项目。目前已有的协同过滤算法包括基于存储的方法（memory-based algorithm）和基于模型的方法（model-based algorithm）。

基于存储的方法利用整个数据集的全部数据计算对某个用户的推荐。常见的基于存储的方法如最近邻方法一般有两个步骤。首先，对某个待推荐的活动用户，算法寻找与他兴趣相似的用户，即所谓的邻居用户。然后，根据这些邻居用户对各项目的评价或偏好，对活动用户进行预测，计算推荐结果。还有一种基于项目的协同过滤方法[17,68]，计算项目之间而不是用户之间的相关度，作为预测的依据。

基于存储的方法实现简单，不需要复杂、费时的建模过程，因而在各种推荐系统中得到广泛应用，如 GroupLens[9]，Ringo[7]等。

基于模型的方法不需要在每次推荐时计算数据库中的全部数据，而是事先定义并学习一个概率模型，计算用户对项目评价的期望。基于模型的方法离线完成建模等大部分的计算，能保证在线计算推荐时可以在常数时间内得出结果，从而解决了基于存储的方法在线计算量随数据规模一同增大、可扩展性差的问题。现有的几种基于模型的方法有 Bayesian 网络[18]，Bayesian 聚类[18]，个性诊断[19]，潜在语义分析[63]，潜在 Dirichlet 分配[65]等。

目前，无论基于存储的方法还是基于模型的方法，都基于一个基本的假设，即兴趣相近的用户对同一项目有相似的评价。这样，协同过滤算法的本质问题就是如何准确地找出与活动用户兴趣相近的用户。在基于存储的方法中，这一问题主要表现为如何定义用户相关性，如何确定用于计算推荐的邻居用户的规模。在基于模型的方法中，问题主要表现在模型的概念结构，用户或项目聚类算法等方面。更具体地，基于模型的方法有必要考虑如下问题：第一，模型的概念结构应反映用户对项目评价的实际情况。用户对项目评分的高低受多种因素影响，包括用户对项目主题/内容的感兴趣程度，用户的评分习惯，项目本身的品质等。一个好的协同过滤模型应能体现它们的影响。第二，模型中用户或项目的聚类结果应具有一定的语义意义。一些模型用单一的潜在变量来同时对用户和项目聚类。这种聚类方法中，一个潜在变量很难同时体现用户与用户、项目与项目、用户与项目之间的关系。一个合理的想法是对用户和项目分别聚类，将用户按照相似的兴趣和偏好聚类，将项目按照相似的主题和内容聚类。第三，模型中的聚类方法应具有一定灵活性。许多协同过滤模型将用户和项目进行“硬”聚类，每个用户或项目只属于单一的一个类，这就很难描述同一用户可能具有的多种兴趣以及同一项目的多种不同属性。比较合理的做法是“软”聚类，每个用户或项目依概率属于多个类，用户在某类上概率值的大小反映了用户对这一类的偏好程度，项目亦然。本文提出了一种新的概率模型——真实偏好高斯混合模型（real preference Gaussian mixture model，RPGMM），从以上三个方面入手研究协同过滤问题。RPGMM 假定评分由用户对项目主题和内容的真实偏好，用户的评分习惯，以及项目的公众评价三个因素决定。用户和项目被分别建模，用两个隐变量表示，同一用户或项目可以依概率属于多个类。

本章首先介绍先前研究者的工作，包括几种基于存储和基于模型的协同过滤算法。然后介绍真实偏好高斯混合模型，主要介绍了模型的概念结构和用于估计模型参数的算法。最后给出一些实验和性能评价，包括不同参数的选取对算法预测结果的影响，以及新模型与几种传统协同过滤模型的对比实验。

3.2 问题的一般模型与算法

3.2.1 协同过滤的符号表示

协同过滤环境下，一般使用用户评分数据库进行预测和推荐。用户评分数据库是一个如表 3.1 所示的 $n\times m$ 稀疏矩阵 $\mathbf{M}$，它的行向量是用户集合 $U=\{u_1,u_2,...,u_n\}$，列向量是项目集合 $Y=\{y_1,y_2,...,y_m\}$，用户对部分项目做出评价，广义评分值集合为 $V_0=\{\perp,v_1,v_2,...,v_k\}$，其中 $\perp$ 表示未评价。有效评分值集合为 $V=\{v_1,v_2,...,v_k\}$。三元组集合 $D=\{\langle u_1,v_1,y_1\rangle,...,\langle u_N,v_N,y_N\rangle\}$ 表示数据库中的所有 N 个有效评分。令 $v_{u,y}$ 表示用户 u 对项目 y 的评分，$\mathbf{v}_u$ 表示用户 u 对各项目的评分值向量（其中用户未评价的项目取值 0），$\mathbf{v}_y$ 表示各用户对项目 y 的评分值向量。

表 3.1　用户－项目评分数据库（满分为 5 分）

项目 / 用户	可可西里	功夫	指环王	唐伯虎点秋香	英雄	十面埋伏	天下无贼
王宏	5	2	5	$\perp$	3	$\perp$	4
田津	2	$\perp$	3	$\perp$	1	$\perp$	$\perp$
李航	$\perp$	1	5	4	$\perp$	4	2
张亮	$\perp$	5	5	5	1	1	$\perp$

研究人员多采用 EachMovie 电影评分数据集，不失一般性，本文假定在电影评分的情景下研究协同过滤问题。这里，用户表现为电影观众，项目表现为电影，用户对项目的评价表现为观众电影的评分值或投票值。这一情景下得到的模型和方法可以被方便地应用到电子商务等其他领域。

3.2.2 基于存储的方法

基于存储的方法的基本思想是计算活动用户与数据库中其他用户的相似程度，得到最相似的邻居用户，根据活动用户的平均评分和邻居用户的加权评分得出预测结果。

用户 u 对项目 y 的预测评分 $p_{u,y}$ 为所有邻居用户 i 的评分加权和：

$$p_{u,y} = \bar{\mathbf{v}}_u + \kappa \sum_{i=1}^{n} w(u,i)(v_{i,y} - \bar{\mathbf{v}}_i) \tag{3.1}$$

其中 n 是协同过滤数据库中与 u 相似度不为 0 的用户数，$w(u,i)$ 是用户 u 和 i 的相似度（权重），$\bar{\mathbf{v}}_u$ 是用户 u 的有效评分平均值，κ 是令相似度绝对值之和为单位值的归一化因子。

计算用户之间相似度的算法中有代表性的是向量空间相似性[18]和 Pearson 相关系数算法[9]

（1） **向量空间相似性**

用户评分向量之间的相似性定义为向量之间的夹角余弦：

$$w(u,i) = \cos(\mathbf{v}_u, \mathbf{v}_i) = \frac{\mathbf{v}_u \cdot \mathbf{v}_i}{\|\mathbf{v}_u\| \cdot \|\mathbf{v}_i\|} \tag{3.2}$$

（2） **Pearson 相关系数**

用户评分向量之间的相似性定义为它们之间的 Pearson 相关系数：

$$w(u,i) = \frac{\sum_{y} (v_{u,y} - \bar{\mathbf{v}}_u)(v_{i,y} - \bar{\mathbf{v}}_i)}{\sqrt{\sum_{y} (v_{u,y} - \bar{\mathbf{v}}_u)^2 \sum_{y} (v_{i,y} - \bar{\mathbf{v}}_i)^2}} \tag{3.3}$$

3.2.3 基于模型的方法

这里介绍三个有代表性的基于模型的方法：贝叶斯聚类[18]，个性诊断[19]，以及概率潜在语义分析[63,20]。

（1） **贝叶斯聚类**

贝叶斯聚类（Bayesian Clustering）模型也称为多项式混合模型，假定用户可以按照他们的共同兴趣和偏好被分为多个群组或类，同一组中的用户对项目有相似的评分。给定一个用户 u 的类 C_u，朴素贝叶斯假定每一类对不同项目的偏好（评分）相互独立。用户所属类和此用户对项目评分的联合概率为：

$$P(C_u = c, \mathbf{v}_u) = P(C_u = c) \prod_{y \in Y} P(v_{u,y} \mid C_u = c) \tag{3.4}$$

用户 u 评分类型的概率为：

$$P(\mathbf{v}_u) = \sum_{c} P(C_u = c, \mathbf{v}_u) = \sum_{c} P(C_u = c) \prod_{y \in Y} P(v_{u,y} \mid C_u = c) \tag{3.5}$$

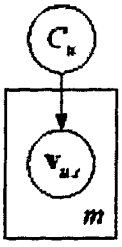


图 3.1 贝叶斯聚类模型，C_u是用户 u 的类

用户属于各类的概率 $P(C_u = c)$，以及条件概率 $P(v_{u,y} \mid C_u = c)$ 都可以由用户评分的训练集上应用期望最大化（EM）算法[37]得到。

（2） 个性诊断

个性诊断（Personality Diagnosis）模型假定用户 u 对项目 y 的评分服从于高斯分布，均值为真实评分 $v_{u,y}^{true}$，标准差 σ 是个自由变量，可根据数据集特点选取合适的值。评分值 $v_{u,y}$ 的条件分布为：

$$P(v_{u,y} = v \mid v_{u,y}^{true}) \propto \exp\left[-\frac{(v - v_{u,y}^{true})^2}{2\sigma^2}\right] \tag{3.6}$$

活动用户 a 与其他用户具有相同个性的概率为：

$$P(\mathbf{v}_a^{true} = \mathbf{v}_u \mid \mathbf{v}_a) \propto \prod_y P(v_{a,y} \mid v_{a,y}^{true} = v_{u,y}) P(\mathbf{v}_a^{true} = \mathbf{v}_u) \tag{3.7}$$

活动用户 a 对项目 y 的预测评分概率为：

$$P(v_{a,y} = v \mid \mathbf{v}_a) \propto \sum_u P(v_{a,y} = v \mid \mathbf{v}_u) P(\mathbf{v}_a^{true} = \mathbf{v}_u \mid \mathbf{v}_a) \tag{3.8}$$

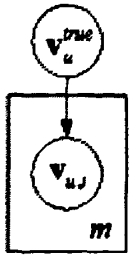


图 3.2 个性诊断模型

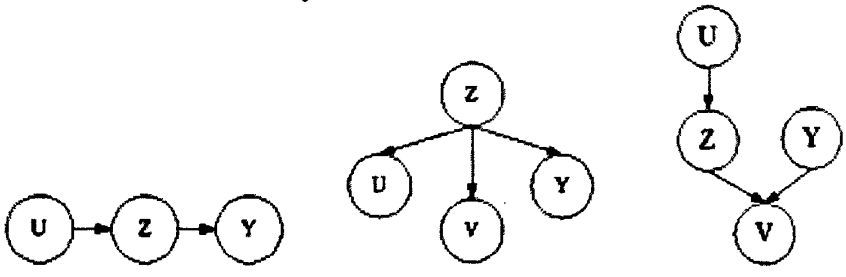
$\mathbf{v}_u^{true}$ 是用户的“真实”个性。实际评分值为 $v_{u,y}$，互相独立并服从于高斯分布

个性诊断算法与贝叶斯聚类方法的区别是它对每个用户建立一个模型。用户评价的项目越多，模型越准确，但在用户评价数量较少时，这种方法的性能比较差。

（3） 概率潜在语义分析

前一章介绍了概率潜在语义分析（probabilistic latent semantic analysis，PLSA）的一般形式，它是一种用统计的方法建立起来的概率潜在空间模型。模型通过引入若干隐藏（潜在）变量探求用户与项目之间的相关关系，每个隐变量都代表一个“主题”或“用户态度”，用户评分数据库中的每个$\langle u,v,y\rangle$三元组都与隐变量相关联。给定某个隐变量，用户和项目相互独立。用户和项目依概率在潜在语义空间的各维（隐变量）上分布，根据这些分布概率，模型对用户和项目分别进行软聚类，聚类结果被用于预测评分和推荐。

PLSA 最先被应用在信息检索领域，对<文本文档，关键词>二元组进行分析建模。将 PLSA 应用于协同过滤的评分预测问题时，模型需要包括评分信息，已提出的 PLSA 协同过滤模型有相位模型[20]和高斯概率潜在语义分析（Gaussian probabilistic latent semantic analysis，GPLSA）[69]。



（a）不含评分信息的基本模型 （b）含有评分信息的相位模型 （c）GPLSA

图 3.3 三种形式的 PLSA 的图模型

含有评分信息的相位模型的图模型表示见图 3.3(b)，$\langle u,v,y\rangle$的联合分布为：

$$P(u,v,y)=\sum_{z}P(z)P(u\mid z)P(y\mid z)P(v\mid z) \tag{3.9}$$

其中$z\in Z=\{z_1,z_2,...,z_k\}$是隐变量，$P(z)$是$\langle u,v,y\rangle$在类$z$上的先验概率，$P(u\mid z)$、$P(y\mid z)$、$P(v\mid z)$分别是是用户、项目和评分与类变量相关的条件概率。

GPLSA 的图模型见图 3.3(c)，预测用户 u 对项目 y 的评分 v 的条件概率，需要计算下面的高斯混合模型：

$$P(v\mid u,y)=\sum_{z}P(v\mid y,z)P(z\mid u)=\sum_{z}P(z\mid u)P(v;\mu_{y,z},\sigma_{y,z}) \tag{3.10}$$

$$P(v:\mu,\sigma)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(v-\mu)^2}{2\sigma^2}\right] \tag{3.11}$$

相位模型和 GPLSA 模型都应用 EM 算法迭代计算模型中的各参数，得到对应于各隐变量的用户群组和项目群组。离线进行 PLSA 的建模过程后，在线预测评分可以在常数时间完成。与贝叶斯聚类和个性诊断模型直接对用户聚类的方法不同，PLSA 模型的特点是每个$\langle u,v,y\rangle$三元组都与一个隐变量 z 相关，用户和项目按照在隐变量上的概率被聚类，聚类结果具有一定的语义意义。

3.2.4 现有协同过滤算法的问题

首先，现有的协同过滤算法和模型没有考虑用户的评分习惯对用户评分行为的影响。不同用户对分值的理解各不相同，评分习惯也不尽相同，一些人比较宽松，另一些人比较严格，致使同样的分值对不同用户而言意义相差很大。即使人为限定了每一分值的含义，仍无法使所有用户对每一档评分的意义达成共识。

其次，没有考虑项目公众评价对用户评分行为的影响。项目的公众评价反映了在大多数用户眼中项目本身的品质，它与项目的内容、主题等没有直接的关系。举例来说，电影《指环王》的公众评价较高，可以认为它的品质较高。这样，某个并不特别偏好奇幻类电影的观众可能也会给出一个比较高的评分，用户给出这个评分，更多的是受电影品质和公众评价的影响，而不是此观众对该电影的主题、内容等的兴趣偏好。

第三，没有将用户与项目分别建模。PLSA 模型只用一个隐变量反映用户的兴趣偏好与项目的内容、主题。但实际情况是，根据用户的兴趣偏好的聚类与根据项目主题内容的聚类结果往往不完全一致。例如，某个观众是演员周星驰的影迷，因而偏好《唐伯虎点秋香》，但他对其他古装戏剧电影并不感兴趣；而另一个观众并不特别偏好周星驰的喜剧表演，他仅仅因为喜欢古装戏剧电影而给予《唐伯虎点秋香》较高评价。用两个隐变量分别对用户和项目进行聚类，可以更准确地反映用户兴趣与项目主题之间的这种差异。

3.3 真实偏好高斯混合模型

本节提出一种应用于协同过滤的新的概率模型，称为真实偏好高斯混合概率模型（real preference Gaussian mixture model，RPGMM）。这种模型的特点是同时考虑了用户评分习惯和项目公众评价对的最终评分值影响，并以两个隐变量对用

户和项目分别建模，用户和项目依概率可同时属于多个类。本节首先描述 RPGMM 的概念模型，然后提出用于实际计算的化简模型以及相应的训练和评分预测算法。

3.3.1 真实偏好高斯混合模型

针对现有协同过滤模型的不足，综合考虑用户评分习惯和项目公众评价对用户最终评分的影响，本文在 PLSA 模型基础上进行了两方面扩展。首先，增加两个隐变量 W_r 和 W_p，分别表示用户评分习惯的类型和项目公众评价的类型。例如，用户评分习惯可以被分为偏高、偏低、中庸三档，项目公众评价也可做类似区分。然后，将 PLSA 模型中的一个隐变量 Z 扩展为两个 Z_u 和 Z_y，它们分别对应于用户按兴趣/偏好的聚类和项目按主题/内容的聚类。自然地，本文假定用户评分习惯与他的兴趣偏好相互独立，项目的公众评价与它的主题内容相互独立。进一步，本文在模型中引入代表用户群组对项目群组的真实偏好的隐变量 X，用户对项目的最终评分由 W_r，W_p 和 X 共同决定。

在预测评分时，有强制预测和自由预测两种方式[20]。强制预测是指为用户提供一个特定的项目以及对它的预测评分。自由预测则允许用户自由选择他感兴趣的项目，需要同时预测用户可能会选择的项目以及他对这个项目的评分。二者对应的模型的形式也稍有区别。自由预测的情形，混合模型的条件概率为：

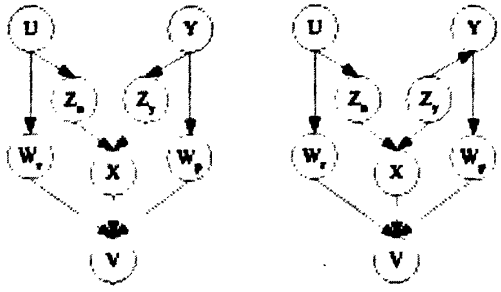
$$P(v,y|u)=\sum_{z_u,z_y,w_r,w_p,x} P(z_u|u)P(z_y)P(y|z_y)P(x|z_u,z_y)$$
$$P(w_r|u)P(w_p|y)P(v|x,w_r,w_p) \qquad (3.12)$$

强制预测的情形，模型可写成：

$$P(v|u,y)=\sum_{z_u,z_y,w_r,w_p,x} P(z_u|u)P(z_y|y)P(x|z_u,z_y)$$
$$P(w_r|u)P(w_p|y)P(v|x,w_r,w_p) \qquad (3.13)$$

其中 $P(z_u|u)$ 是用户 u 在某个特定用户类 z_u 中的概率，$P(z_y|y)$ 是项目 y 在某个项目类 z_y 中的概率。$P(z_y)$ 和 $P(y|z_y)$ 是 z_y 的先验概率和项目 y 由 z_y 产生的概率。$P(x|z_u,z_y)$ 是用户类 z_u 对项目类 z_y 的偏好程度的概率。$P(w_r|u)$ 是用户 u 属于用户评分习惯类 w_r 的概率，$P(w_p|y)$ 是项目 y 属于项目公众评价类 w_p 的概率。

$P(v|x,w_r,w_p)$是给定用户评分习惯类w_r，项目公众评价类w_p，以及真实偏好类x三个条件，最终评分值为v的概率。



（a）强制预测　　（b）自由预测

图 3.4　RPGMM 的图模型

在计算$P(x|z_u,z_y)$时，使用如下高斯模型：

$$P(x|z_u,z_y)=P(x;\mu_{z_u,z_y},\sigma_{z_u,z_y}) \tag{3.14}$$

$$P(x:\mu,\sigma)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \tag{3.15}$$

将公式（3.14）代入公式（3.12）和公式（3.13），得到的新模型是一个高斯混合模型。而且，在这个模型中，隐变量 X 表示去除了用户评价习惯和项目公众评价后的用户群组对项目群组的“真实”偏好，因此称新模型为真实偏好高斯混合模型（RPGMM）。

RPGMM 与 PLSA 模型相比，主要有以下区别：首先，RPGMM 考虑了用户评价习惯和项目公众评价对最终评分值的影响，模型透过表面评分值探求其中隐的用户对项目的真实偏好程度。而 PLSA 模型直接以评分值作为用户对项目偏好程度的度量，无法避免先天的误差。其次，RPGMM 用两个隐变量对用户和项目分别聚类建模，与 PLSA 模型用户和项目共用一个隐变量的方式相比，可以更清晰地得到按兴趣和偏好区分的用户群组和按内容和主题区分的项目群组，因而更具有语义意义。

3.3.2 化简的模型

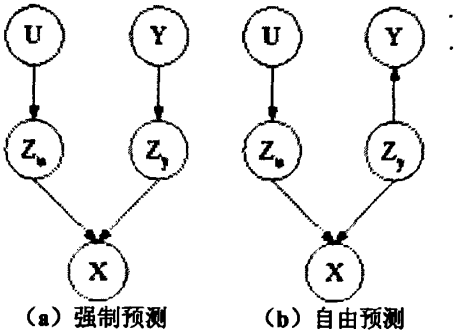
从公式（3.12）和公式（3.13）可以看出，RPGMM 计算评分的概率时需要对各隐变量的所有取值求和。应用 EM 算法迭代计算求解模型的各参数时，计算复

杂度将是$O(I\cdot k_1\cdot k_2\cdot w_1\cdot w_2\cdot k)$，其中$I$是迭代次数，$k_1,k_2,w_1,w_2,k$为隐变量$Z_u$，$Z_y$，$W_r$，$W_p$，$X$各自的变量个数。而且，由于模型的参数过多，在进行迭代优化时很难保证各参数都收敛到最优解。显然这种形式的模型过于复杂而不适于进行实际建模计算。

进一步研究 RPGMM，不难发现用户评分习惯类型和项目公众评价类型信息分别由某用户对各个项目，以及所有用户对某一项目的评分值体现。本文把用统计方法从用户评分数据库中得到的用户评分习惯类型和项目公众评价类型作为先验信息化简模型，去掉隐变量W_r和W_p。化简后的 RPGMM 的图形模型如图 3.5 所示，以下本文研究的 RPGMM 就是化简的 RPGMM。RPGMM 不直接使用用户对项目的评分值，而需要预先去除用户评分习惯和项目公众评价的影响，得到用户对项目的"真实偏好"，即模型中的隐变量 X。强制预测和自由预测的情形下，RPGMM 预测真实偏好的概率模型可写成如下形式：

强制预测：
$$P(x|u,y)=\sum_{z_u,z_u}P(u|z_u)P(y|z_y)P(x|z_u,z_y) \tag{3.16}$$

自由预测：
$$P(x,y|u)=\sum_{z_u,z_u}P(z_u|u)P(z_y)P(y|z_y)P(x|z_u,z_y) \tag{3.17}$$



（a）强制预测　　（b）自由预测

图 3.5　化简的 RPGMM 的图模型

和其他基于模型的协同过滤算法一样，RPGMM 算法分训练和预测两部分。在训练阶段，模型的各条件概率参数被迭代计算得到最优值；在预测阶段，利用先前建立的模型预测测试集的用户评分。

3.3.2 对评分值的预处理和后处理

对评分值的预处理，就是从原始的用户评分值中去除用户评分习惯和项目公

众评价的影响，得到真实偏好的过程。后处理则相反，是将用户的评分习惯和项目的公众评价的影响加到预测出的真实偏好上，得到"实际"评分的预测值。这样，RPGMM 算法的三个步骤为：预处理，模型的训练和预测，后处理。

（1） 预处理

在给出预处理算法之前，参考文献[70]中计算用户对项目偏好值的方法，本文提出两个假定：第一，u 给予评分的所有项目中，低于 v 的比例越高，u 对 y 的真实偏好值越大。例如：某用户的评分习惯是对每个项目都给予较低评价，此用户对某个项目的评价值不高，但相对于他对其他项目的评价来说较高，也可认为他比较偏好这个项目。第二，y 得到的所有评分中，低于 v 的比例越高，y 对 u 的吸引力越大（u 对 y 的真实偏好值越大）。消除电影公众评价影响。以电影评价为例：多数人对某动画片评价不高，但某儿童可能会给予高分评价，这反应了他对"动画片"一类电影比较高的真实偏好。

根据这两个假定，提出以下算法，由实际评分值 $v_{u,y}$ 计算真实偏好 $x_{u,y}$：

$$x_{u,y}=\sum_{p=0}^{1} p\cdot P(p\mid u,v,y) \tag{3.18}$$

其中 $p=pref(u,y)$ 表示用户 u 是否偏好项目 y，定义如下：

$$p=pref(u,y)=\begin{cases}1, \text{如果用户 } u \text{ 偏好项目 } i\\ 0, \text{反之}\end{cases} \tag{3.19}$$

$P(p\mid u,v,y)$ 表示用户 u 偏好项目 y 的概率，即 RPGMM 中 u 对 y 的"真实偏好"概率。它的计算公式为：

$$P(p\mid u,v,y)=\alpha P(p_u\mid u,v)+(1-\alpha)P(p_y\mid v,y) \tag{3.20}$$

其中 $P(p_u\mid u,v)$ 是用户 u 对他给予评分 v 的所有项目的"个体偏好"概率，它去除了用户评分习惯的影响。$P(p_y\mid v,y)$ 是对项目 y 给予评分 v 的所有用户的"公众偏好"概率，它去除了项目公众评价的影响。$0<\alpha<1$ 是比例系数，根据数据集中个体偏好和公众偏好对真实偏好的影响程度设定。这样，真实偏好就是个体偏好和公众偏好的加权和。

为了去除用户评分习惯的影响得到个体偏好，一个简单的想法是统计不大于 v 的评分数在用户 u 所有评分中所占比例，概率形式为 $P(v_u\le v\mid u)$；以及用 u 做出的值为 v 的评分数在他的所有评分中所占的比例，概率形式为 $P(v_u=v\mid u)$。前者值越大，说明用户越偏好这个项目；后者值越大，说明用户越不偏好这个项目。个体偏好概率的计算公式如下：

$$P(p_u \mid u,v) = P(v_u \le v \mid u) - \frac{1}{2}P(v_u = v \mid u) = \frac{\sum_{v' \le v} c_u(v') - \frac{1}{2}c_u(v)}{\sum_{v'' \in V} c_u(v'')} \tag{3.21}$$

其中 $c_u(v)$ 是用户 u 做出的值为 v 的评分的数量。

公众偏好与个体偏好计算方法类似，其计算公式如下：

$$P(p_y \mid v,y) = P(v_y \le v \mid y) - \frac{1}{2}P(v_y = v \mid y) = \frac{\sum_{v' \le v} c_y(v') - \frac{1}{2}c_y(v)}{\sum_{v'' \in V} c_y(v'')} \tag{3.22}$$

其中 $P(v_y \le v \mid y)$ 表示不大于 v 的评分数在所有对项目 y 的评分中所占比例，$P(v_y = v \mid y)$ 表示值为 v 的评分数在所有对项目 y 的评分中所占比例。$c_y(v)$ 是对项目 y 的值为 v 的评分的数量。

注意经过预处理计算得到的隐变量 X 与未经化简的 RPGMM 中的 X 略有区别。前者是连续变量，$0 \le x \le 1$ $(x \in X)$ 表示用户对项目的真实偏好程度。后者是离散变量，和 V 一样分为 k 档，表示用户对项目的真实评分。

（2） 后处理

后处理实际是预处理的逆过程，由预测得到的真实偏好 $x_{u,y}$ 计算实际评分的预测值 $v_{u,y}$。计算公式如下：

$$P(p \mid u,y) = p \cdot x_{u,y} \tag{3.23}$$

$$v_{u,y} = \left\lfloor \alpha f_u(P(p \mid u,y)) + (1-\alpha) f_y(P(p \mid u,y)) + 0.5 \right\rfloor \tag{3.24}$$

其中 $f_u(t)$ 是用户 u 所有评分由低到高排成的序列中第 $\left\lfloor t \cdot \sum_v c_u(v) \right\rfloor$ 项的评分值。$f_y(t)$ 是对项目 y 的所有评分由低到高排成的序列中第 $\left\lfloor t \cdot \sum_v c_y(v) \right\rfloor$ 项的评分值。

3.3.3 训练过程

经过预处理，由用户对项目的实际评分 V 得到真实偏好 X，RPGMM 被化简为图 3.5 所示的含有两个隐变量的混合模型。算法的训练过程任务是估计模型的参

数。以下以强制预测的情形为例，说明参数估计过程。

算法的输入数据为 $D'=\{\langle u_1,x_1,y_1\rangle,...,\langle u_N,x_N,y_N\rangle\}$，包括 N 个观测三元组 $\langle u,x,y\rangle$。数据集 D'的对数相似度（L）为：

$$\text{强制预测：}\quad L=\sum_{<u,x,y>}\log P(x|u,y) \tag{3.25}$$

$$\text{自由预测：}\quad L=\sum_{<u,x,y>}\log P(x,y|u) \tag{3.26}$$

其中 Y_u 是用户 u 评价的项目集。

对 L 进行归一化并取反，得到反映模型描述实际数据准确程度的归一化负对数相似度（NNL）：

$$\text{强制预测：}\quad NNL=-\frac{1}{N}\sum_{<u,x,y>}\log P(x|u,y) \tag{3.27}$$

$$\text{自由预测：}\quad NNL=-\frac{1}{N}\sum_{<u,x,y>}\log P(x,y|u) \tag{3.28}$$

根据 RPGMM 的结构，定义完全数据模型为五元组 $\langle u,x,y,z_u,z_y\rangle$，其中 z_u 和 z_y 是隐变量。将公式（3.16）、（3.17）分别代入（3.27）、（3.28），可得：

$$\text{强制预测：}\quad NNL=-\frac{1}{N}\sum_{<u,x,y,z_u,z_y>}\Big[\log P(u|z_u)+\log P(y|z_y)+\log P(x|z_u,z_y)\Big] \tag{3.29}$$

$$\text{自由预测：}\quad NNL=-\frac{1}{N}\sum_{<u,x,y,z_u,z_y>}\Big[\log P(z_u|u)+\log P(z_y)+\log P(y|z_y)+\log P(x|z_u,z_y)\Big] \tag{3.30}$$

引入一个隐变量概率分布 $Q(z_u,z_y|u,x,y)$，它满足

$$\sum_{z_u,z_y}Q(z_u,z_y|u,x,y)=1 \tag{3.31}$$

来描述隐变量与三元组 $\langle u,x,y\rangle$ 的条件概率关系。根据 NNL 得到最大似然估计函数（MLF）：

强制预测：$$MLF=-\frac{1}{N}\sum_{<u,x,y>}\sum_{z_u,z_y}Q^*(z_u,z_y\mid u,x,y)$$

$$\left[\log P(u\mid z_u)+\log P(y\mid z_y)+\log P(x\mid z_u,z_y)\right] \quad (3.32)$$

自由预测：$$MLF=-\frac{1}{N}\sum_{<u,x,y>}\sum_{z_u,z_y}Q^*(z_u,z_y\mid u,x,y)$$

$$\left[\log P(z_u\mid u)+\log P(z_y)+\log P(y\mid z_y)+\log P(x\mid z_u,z_y)\right] \quad (3.33)$$

其中$Q^*(z_u,z_y\mid u,x,y)$是似然函数最大值对应的隐变量概率分布。

本文使用期望最大化（EM）算法[37]最大化混合模型的对数似然函数。EM 算法通过期望一最大化两步过程反复迭代直到收敛。在期望步，根据贝叶斯公式，计算如下后验概率：

强制预测：$$Q^*(z_u,z_y\mid u,x,y)=P(z_u,z_y\mid u,x,y)$$

$$=\frac{P(u\mid z_u)P(y\mid z_y)P(x\mid z_u,z_y)}{\sum_{z_u',z_y'}P(u\mid z_u')P(y\mid z_y')P(x\mid z_u',z_y')} \quad (3.34)$$

自由预测：$$Q^*(z_u,z_y\mid u,x,y)=P(z_u,z_y\mid u,x,y)$$

$$=\frac{P(z_u\mid u)P(z_y)P(y\mid z_y)P(x\mid z_u,z_y)}{\sum_{z_u',z_y'}P(z_u'\mid u)P(z_y')P(y\mid z_y')P(x\mid z_u',z_y')} \quad (3.35)$$

将公式（3.34）、（3.35）分别代入（3.32）、（3.33），*MLF* 可被写成：

强制预测：$$MLF=-\frac{1}{N}\sum_{<u,x,y>}\sum_{z_u,z_y}\frac{P(u\mid z_u)P(y\mid z_y)P(x\mid z_u,z_y)}{\sum_{z_u',z_y'}P(u\mid z_u')P(y\mid z_y')P(x\mid z_u',z_y')}$$

$$\cdot\left[\log P(u\mid z_u)+\log P(y\mid z_y)+\log P(x\mid z_u,z_y)\right] \quad (3.36)$$

自由预测：$$MLF=-\frac{1}{N}\sum_{<u,x,y>}\sum_{z_u,z_y}\frac{P(z_u\mid u)P(z_y)P(y\mid z_y)P(x\mid z_u,z_y)}{\sum_{z_u',z_y'}P(z_u'\mid u)P(z_y')P(y\mid z_y')P(x\mid z_u',z_y')}$$

$$\cdot\left[\log P(z_u\mid u)+\log P(z_y)+\log P(y\mid z_y)+\log P(x\mid z_u,z_y)\right] \quad (3.37)$$

在 EM 算法的最大化步，更新模型的参数：

$$P(u|z_u)=\frac{\sum_{z_1}\sum_{(u',x,y):u'=u}P(z_u,z_y|u,x,y)}{\sum_{z_y}\sum_{(u,x,y)}P(z_u,z_y|u,x,y)} \tag{3.38}$$

$$P(z_u|u)=\frac{\sum_{(u',x,y):u'=u}\sum_{z_y}P(z_u,z_y|u,x,y)}{\sum_{(u',x,y):u'=u}\sum_{z_u,z_y}P(z_u,z_y|u,x,y)} \tag{3.39}$$

$$P(z_y)=\frac{\sum_{(u,x,y)}\sum_{z_u}P(z_u,z_y|u,x,y)}{\sum_{(u,x,y)}\sum_{z_u,z_y}P(z_u,z_y|u,x,y)}=\frac{\sum_{(u,x,y)}\sum_{z_u}P(z_u,z_y|u,x,y)}{N} \tag{3.40}$$

$$P(y|z_y)=\frac{\sum_{z_u}\sum_{(u,x,y'):y'=y}P(z_u,z_y|u,x,y)}{\sum_{z_u}\sum_{(u,x,y)}P(z_u,z_y|u,x,y)} \tag{3.41}$$

计算$P(x|z_u,z_y)$时，采用公式（3.14）的高斯模型，模型参数的更新公式为：

$$\mu_{z_u,z_y}=\frac{\sum_{(u,x,y)}xP(z_u,z_y|u,x,y)}{\sum_{(u,x,y)}P(z_u,z_y|u,x,y)} \tag{3.42}$$

$$\sigma_{z_u,z_y}=\frac{\sum_{(u,x,y)}(x-\mu_{z_u,z_y})^2P(z_u,z_y|u,x,y)}{\sum_{(u,x,y)}P(z_u,z_y|u,x,y)} \tag{3.43}$$

由于在算法的预处理阶段已将用户评分习惯和项目公众评价的影响消除，评分值$0\le x_{u,y}\le 1$，所以公式（3.14）的高斯模型不需要将评分值归一化。

为避免模型的统计学习中可能出现的过拟和问题，本文采用温度EM（tempered EM，TEM）算法[64]调整混合模型的学习过程。在这种算法中，训练数据集被分为训练集和保留集两部分。在标准EM算法的期望步引入一个控制变量β作为温度，调整公式（3.29）、（3.30）的期望步为：

强制预测：$$P(z_u,z_y|u,x,y)=\frac{[P(u|z_u)P(y|z_y)P(x|z_u,z_y)]^\beta}{\sum_{z_u,z_y}[P(u|z_u)P(y|z_y)P(x|z_u,z_y)]^\beta} \tag{3.44}$$

自由预测：$$P(z_u,z_y|u,x,y)=\frac{P(z_y)[P(z_u|u)P(y|z_y)P(x|z_u,z_y)]^\beta}{\sum_{z_u,z_y}P(z_y)[P(z_u|u)P(y|z_y)P(x|z_u,z_y)]^\beta} \tag{3.45}$$

温度控制变量$\beta=1$时，TEM 与标准 EM 算法相同。通过减小β的值，TEM 算法调整模型参数避免收敛到局部最优解。开始时设定$\beta=1$，运行 EM 算法直到出现过拟和现象，EM 算法早期停止，减小β的值($\beta_{t+1}=\eta\beta_t, 0<\eta<1$)，再次运行 EM 算法直到减小$\beta$不会带来保留数据集上更多的性能提高或者$\beta$小于一个特定值$\beta_{\min}$。最后使用所有训练数据（包括保留数据）和当前的$\beta$值进行训练。在下面的实验中，取$\eta=0.92$，$\beta_{\min}=0.5$，保留数据集和训练集都取全体训练数据。

3.3.4 预测过程

经训练得到的模型参数被用于预测测试集的用户评分。如果将公式（3.16）、（3.17）中的u视为测试用户，则公式右边的各参数中只有$P(z_u|u)$和$P(u|z_u)$是未知量。将训练得到的其他参数作为固定的常数值，并把测试用户加入训练集中，再次运行 EM 算法，即可求得未知参数，从而得到公式（3.16）、（3.17）左边的真实偏好概率。RPGMM 预测真实偏好的期望为：

$$E[x|u,y]=\int_X xP(x|u,y)dx=\sum_{z_u,z_y}P(z_u|u)P(z_y|y)\mu_{z_u,z_y} \tag{3.46}$$

$$E[x,y|u]=\int_X xP(x,y|u)dx=\sum_{z_u,z_y}P(y|z_y)P(z_y)P(z_u|u)\mu_{z_u,z_y} \tag{3.47}$$

根据预测真实偏好求预测评分的方法见 RPGMM 算法的后处理部分。

3.4 实验与讨论

本节介绍对新模型的验证和对比实验。包括纵向对比和横向对比两部分。纵向对比：考察隐变量数量对预测准确度的影响，用户评分习惯与项目公众评价对预测准确度的影响哪个更大，强制预测与自由预测两种方式模型的性能比较。横向对比：新模型与其他几个协同过滤算法的性能比较。

3.4.1 数据集

以下实验中使用的数据集是可公开下载到的 EachMovie 数据集，它包括 61,265 个用户对 1,623 个电影的评分，每个用户平均评价了 45.6 个电影，用户的评分表现为 0.0 到 1.0 的 6 档离散值。在实验中首先将评分转换为自然数评分 $V=\{1,2,...,6\}$，然后从数据集中提取一个子集，包括 1000 个评分数超过 100 的用户，每个用户平均评价了 165.9 个电影。

3.4.2 评价指标

本文用在协同过滤研究领域中被普遍使用的平均绝对误差（mean absolute error，MAE）和均方差（mean square error，MSE）作为算法统计准确度的衡量指标：

$$MAE=\frac{1}{|T|}\sum_{<u,y,v>\in T}\left|v_{u,y}{}'-v_{u,y}\right| \tag{3.48}$$

$$MSE=\frac{1}{|T|}\sum_{<u,y,v>\in T}(v_{u,y}{}'-v_{u,y})^2 \tag{3.49}$$

其中 T 是测试集，$v_{u,y}{}'$ 是预测得到的评分，$v_{u,y}$ 是实际评分。MAE 和 MSE 值越小说明预测结果与实际评分值差别越小，算法的效果也越好。

3.4.3 评价协议

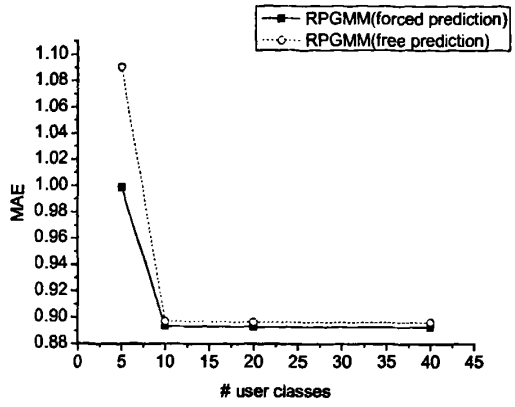
本文采用 AllButOne 和 Given$N$ 两种协议评价预测结果准确性。AllButOne 协议是在每个测试用户的有效评分中随机地选取 1 个作为未知的待预测评分，其他所有评分作为已知评分值。这种协议中算法从每个测试用户获得了最多的已知信息，反映了算法在数据库已经收集到足够多信息，进入平稳状态时预测评分的能力。Given$N$ 是对每个测试用户，随机选取 N 个（例如，5 个，10 个，20 个）作为已知评分，其他所有评分作为待预测评分。这种协议中算法能得到的已知信息相对较少，它能反映在数据库的建立阶段，新用户刚刚进入推荐系统，尚未做出很多评分时算法的性能。

应用这两种协议时，将所有用户同时作为测试用户和训练用户。所有已知的用户评分构成训练数据集，所有待预测评分构成测试数据集。评测时将预测结果与相应的实际评分相比较，考察算法的性能。

3.4.4 纵向对比

纵向对比实验评测模型中各参数对算法性能的影响。在以下的实验中，如果未加特别说明，模型的各参数被设置为：训练集和测试集都是 1000 个用户的评分值，协议为 AllButOne，用户类隐变量 Z_u 数量为 10 个，项目类隐变量 Z_y 数量为 20 个，比例系数 α 取值 0.5，EM 算法迭代次数为 60 次。以下所有实验结果都是独立进行的 10 次实验的平均值。

（1） 隐变量数量对模型预测准确度的影响



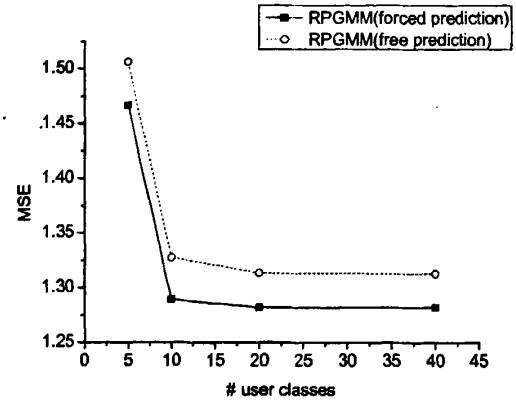
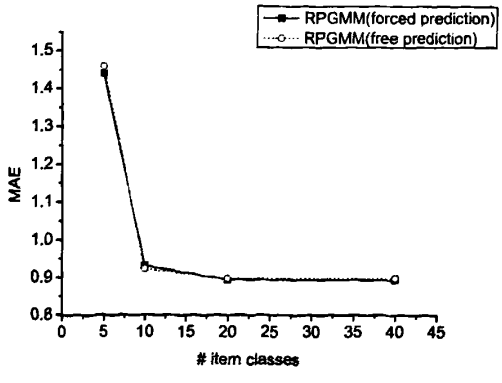


图 3.6 项目类固定为 20，用户类取不同值时 RPGMM 自由预测和强制预测方式的性能



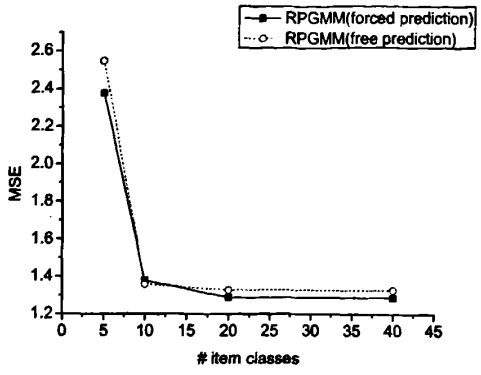


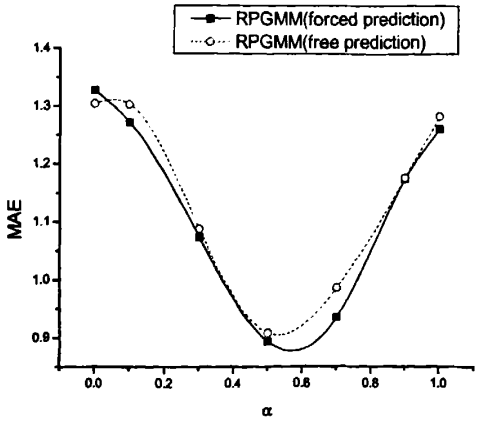
图 3.7　用户类固定为 10，项目类取不同值时 RPGMM 自由预测和强制预测方式的性能

图 3.6 是固定项目类为 20，用户类取 5，10，20 和 40 时 RPGMM 两种预测方式的实验结果。从图中可见，算法预测效果随用户类数量增加而提高。用户类数量为 10 以上时算法基本稳定。

图 3.7 是固定用户类为 10，项目类取 5，10，20 和 40 时 RPGMM 两种预测方式的实验结果。算法性能随项目类数量增加而提高。项目类在 20 以上时算法基本稳定。

（2）　比例参数α对模型预测准确度的影响

图 3.8 显示了比例参数α取不同值时 RPGMM 两种预测方式的性能。在 MAE 和 MSE 两个指标上，RPGMM 的两种预测方式都在$\alpha \in [0.5, 0.7]$时达到最好的预测效果。



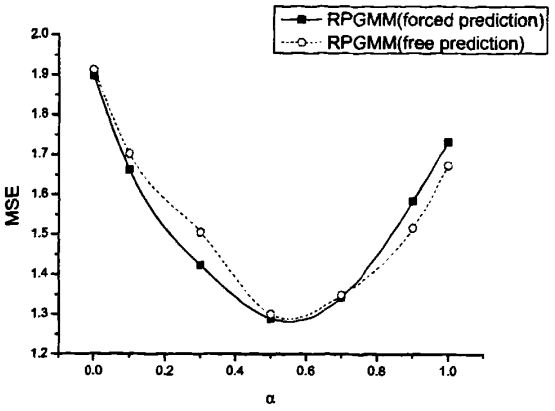
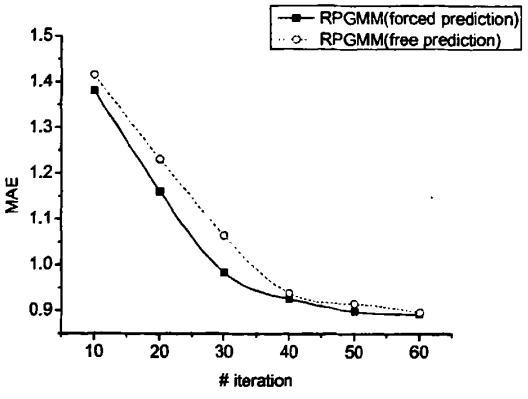


图 3.8 比例参数 α 取不同值时 RPGMM 自由预测和强制预测方式的性能

（3） 迭代次数对模型预测准确度的影响



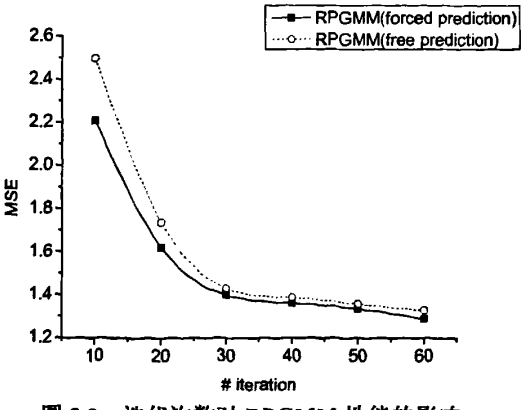


图 3.9 迭代次数对 RPGMM 性能的影响

图 3.9 显示了 EM 算法迭代次数对 RPGMM 两种预测方式性能的影响。从图中可见，RPGMM 性能随迭代次数的增加而提高。迭代次数小于 30 次时，RPGMM 预测准确度较低。迭代次数在 30 次以上时，RPGMM 性能可以接受，但继续增大迭代次数，预测效果没有显著提高。

（4） 已知用户评分信息量大小对模型预测准确度的影响

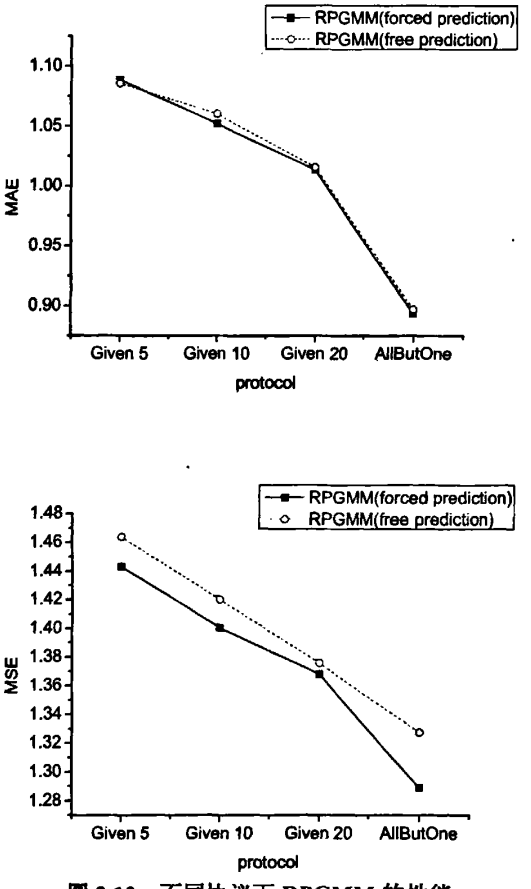


图 3.10　不同协议下 RPGMM 的性能

图 3.10 显示了不同协议下 RPGMM 的性能。从图中可见，Given*N* 协议下算法性能随 *N* 的增加而提高，在 AllButOne 协议下达到最高。这表明已知信息越多，RPGMM 建立的模型越准确，算法预测效果越好。

3.4.5 横向对比

通过横向对比实验比较 RPGMM 与已知的几种协同过滤算法的性能。

表 3.2 各种协同过滤算法的性能比较（AllButOne 协议）

算法 / 指标	BC[18]	BN[18]	PC	VS	PD	GPLSA	RPGMM (free)	RPGMM (forced)
MAE	1.103	1.066	0.952	0.948	1.233	0.902	0.897	**0.893**
MSE	—	—	1.323	1.319	2.070	1.395	1. 328	**1.289**

表 3.2 是在 EachMovie 数据集上对常见的几种协同过滤算法进行实验得到的结果。实验所用协议为 AllButOne，其中 BC（贝叶斯聚类）和 BN（贝叶斯网络）的数据由文献[18]得到，其他几种算法中，PC 是 Pearson 相关系数法，VS 是向量相似度法，PD 是个性诊断，GPLSA 是高斯概率潜在语义分析模型，RPGMM (free)和(forced)分别是本文提出的新模型的自由预测和强制预测模式。GPLSA 模型的隐变量数量被设置为 40。RPGMM 中，表示用户和项目类的隐变量数量被分别设置为 10 和 20，偏好比例系数 α 取值 0.5。

实验结果表明潜在空间模型（GPLSA 和 RPGMM）的性能明显优于传统的基于存储的协同过滤算法（PC，VS 等）以及基于模型的聚类算法（BC，PD 等）。RPGMM 的强制预测方式在 MAE 和 MSE 两个指标上的预测效果达到实验中各种算法中的最优值。

3.4.6 计算复杂度

RPGMM 是一种基于模型的协同过滤方法，它分为离线建模和在线预测和推荐两个阶段。离线建模阶段，对数据的预处理算法复杂度为 $O(N)$。然后用 EM 算法迭代计算模型各参数，计算复杂度为 $O(I \cdot N \cdot k_1 \cdot k_2)$，其中 I 是迭代次数，N 是训练集中 $\langle u,v,y \rangle$ 三元组的个数，k_1，k_2 是用户类和算法类隐变量的个数。RPGMM 建模时，EM 算法迭代次数一般可取为 30 次到 60 次，k_1，k_2 分别取值 10 和 20，假定数据集包括 1000 个用户，平均每人评价了 100 个项目，则 N=100,000。在一台 Pentium4 2.6GHz CPU 的计算机上，一次建模过程需要大约 20 分钟时间。

在线阶段按照公式（3.46）、（3.47）计算评分预测值，模型的各参数在离线建模阶段已被计算得到，因此计算复杂度为 $O(k_1 \cdot k_2)$，即在线预测评分可以在常数时间内完成。

3.5 本章小结

对协同过滤问题，本章提出了一种新的概率模型真实偏好高斯混合模型（RPGMM）。RPGMM 是一种层次 Bayesian 模型，它对用户和项目分别建模，用户和项目可以依概率属于多个类。它的另一特点是用户最终评分由用户评分习惯，项目公众评价以及用户对项目的"真实偏好"共同决定。这一模型适应性强，可扩展性好，可以在常数时间内计算得到预测和推荐结果。实验表明新模型的性能明显优于其他几种协同过滤模型。

第四章 无监督文本聚类的混合模型

本章研究用于大规模文本数据聚类的有限混合模型方法。有限混合模型聚类的问题之一是特征选择，即在没有类标记的情况下，如何选择适合模型学习的相关特征集。本章定义了一种改进的"特征显著性"方法，将特征对各混合成员的相关性作为隐变量引入混合模型，在估计模型参数的同时完成特征选择。进一步提出了一种有限混合模型进行无监督文本聚类的广义方法，它将模型选择，特征选择以及混合模型的参数估计纳入一个统一的框架中。在这个广义框架下，发展了一种带特征选择的多项式混合模型（MM-FS），并以 BIC 准则作为模型选择方法。

4.1 无监督文本聚类概述

随着数据存储技术和通信网络的发展，人们可以方便地访问和获取大量文本文档，无监督文本聚类已成为文本挖掘的一个重要任务。对大量文本数据的聚类分析涉及机器学习、自然语言处理、数据存储和挖掘等多个领域。对无监督学习（即聚类）算法的研究是文本聚类分析的重要课题。在各种聚类算法中，基于模型的方法从数据中学习一个生成模型，它可以揭示数据内在结构和相互间联系，受到研究人员的广泛关注。本文采用机器学习和统计学方法，考察基于概率模型的聚类算法。

近年来，研究人员提出了多种用于无监督文本聚类的有限混合模型。Liu 等[71]使用 Gaussian 混合和 EM 算法对文档集聚类，并给出了相应的模型选择方法。Nigam 等[38]将一种多项式混合用于文本聚类。文档中词的出现次数依多项式分布，模型由这些多项式分布的混合构成，EM 算法被用于模型的概率推断。Blei 等提出的 LDA 模型[65]由 PLSA 模型[63]发展而来，它考虑了文档主题的概念，用一些潜在主题的混合模型来表示文档，而主题由在词上的分布来体现。Zhong 等[72]对常用于文本建模的多项式混合，von Mises-Fisher 混合，多元 Bernoulli 模型等生成性模型进行了比较实验，结论是 von Mises-Fisher 稍好于多项式混合，而多元 Bernoulli 模型聚类效果最差。

对文本数据进行无监督特征选择的困难在于数据没有指导性的类标记，从而造成一个两难的问题：从多维数据空间中提取有用的特征需要对数据进行好的聚类，而特征选择对聚类效果的好坏又有直接影响。Yang 等[73]对一些常见的适用于文本分类的特征选择方法，如 IG，CHI-Square，MI，DF 等，进行了比较研究。然而，它们中的大部分都需要文档的类标记信息，并不适用于文本的无监督学习。无监督学习的特征选择方法必须不依赖于数据的类标记信息。一般地，特征选择

方法可分为两大类：filter 和 wrapper。前者的特征选择独立于学习过程；后者则将特征选择融合于学习过程中，根据学习算法选取特征子集。Dash 等[74]提出了一种独立于聚类算法的 filter 类特征选择方法，用一种熵度量来考察同一聚类中的数据在各个特征维上的距离，根据熵度量值的大小选择特征子集。对大规模数据库的特征选择，Agrawal 等[29]提出了寻找密集数据的特征子集的 CLIQUE 算法。Dy 等[75]使用 EM 算法进行聚类并以一种 wrapper 方法 FSSEM-k 进行特征选择。Law 等[76]提出了一种"特征显著性"概念，将特征选择以隐变量的形式加入 Guassian 混合模型，同时完成特征选择和参数估计，本章的部分工作即受此方法的启发。

有限混合模型的一个重要问题是模型选择，即如何确定混合成员（即聚类）的个数。混合成员过多会造成模型过度适应，混合成员过少会造成模型不足拟和。研究人员提出了各种模型选择的比较准则，包括 Bayesian 信息准则（BIC）[56]，Akaike 信息准则（AIC）[57]，最小描述长度（MDL）[77]和最小消息长度（MML）[78]等。Kontkanen 等[79]对 AIC 和 BIC 进行了比较实验，结果表明 BIC 准则比 AIC 准则效果更好。Figueiredo 等[32]对 Gaussian 混合模型的模型选择进行了深入研究，并提出了一种改进的 MML 准则，可以在用 EM 算法对 Gaussian 模型的参数估计时动态调整成员数量，减小了对不同模型进行比较的计算量。

对无监督文本学习的有限混合模型方法，研究人员进行了许多工作，但现有的工作还存在两个主要问题：第一，在没有类标记的情况下，如何选择适合模型学习的相关特征集。第二，各种有限混合模型以及模型选择方法的应用，尚缺少有指导意义的一般性方法。针对这两个问题，本文提出了一个基于混合模型对文本数据进行聚类的框架方法，称为"有限混合模型进行无监督文本聚类的广义方法"，简称"广义方法"。它的特点是：第一，面向文档聚类，以 bag-of-words 模型表示文档，同时具有特征选择能力。第二，基于有限混合，是一种生成性概率模型。第三，模型具有很强的普遍性，可应用于多种混合形式，并可依各种准则完成模型选择。在此框架下，本文发展了一种新的面向无监督文本学习的有限混合模型方法，即以一种带特征选择的多项式混合模型对文本建模，并依 Bayesian 信息准则（BIC）进行模型选择。

本章首先介绍广义方法，包括模型定义，参数估计，特征选择和模型选择。然后提出了用于无监督文本聚类的多项式混合和 BIC 模型选择标准。最后通过各文本数据集上的实验来考察广义方法的效果。

4.2 有限混合模型聚类的广义方法

本文研究对文本文档进行无监督聚类的方法。采用在文档分析领域被广泛应用的 Salton 的 bag-of-words 模型来表示文档，忽略词在文档中的位置信息，每个文档由一个固定词汇表上的词的分布向量来表示。定义文档集为 $D=\{\mathbf{w}_1,...,\mathbf{w}_N\}$，

用 $\mathbf{w}=(w_1,...,w_M)$ 表示其中的一个文档，w_i 代表词的出现频率（次数）。

4.2.1 有限混合模型

假定文档数据 D 由 K 个概率模型的混合产生，每个模型的数学形式相同（例如 Gaussian），各有自己的分布，文档向量之间相互独立。一个文档向量 $\mathbf{w}$ 的有限混合模型定义为：

$$p(\mathbf{w};\Theta)=\sum_{k=1}^{K}\lambda_k p(\mathbf{w};\boldsymbol{\theta}_k) \tag{4.1}$$

其中 λ_k 是混合权重，它满足 $\lambda_k>0,k=1,...,K$ 且 $\sum_{k=1}^{K}\lambda_k=1$。$\boldsymbol{\theta}_k$ 是第 k 个混合成员的参数。$\Theta=\{\lambda_1,...,\lambda_K,\boldsymbol{\theta}_1,...,\boldsymbol{\theta}_K\}$ 是混合模型的参数。

为了对文本进行聚类，定义一个指示向量 $\mathbf{c}=(c_1,...,c_K)$，$\mathbf{c}$ 的其中一个分量 c_k 值为 1，其他都是 0，用于指示文档是由第 k 个混合成员生成的。指示向量的集合为 $C=\{\mathbf{c}_1,...,\mathbf{c}_N\}$。文档 $\mathbf{w}$ 是已知的，指示向量 $\mathbf{c}$ 是潜在的，二者构成完全数据 $\mathbf{x}=\{\mathbf{w},\mathbf{c}\},\mathbf{x}\in X=D\cup C$。新的文本数据和聚类标记的联合密度函数为：

$$p(\mathbf{x};\Theta)=p(\mathbf{w},\mathbf{c};\Theta)=\sum_{k=1}^{K}c_k\lambda_k p(\mathbf{w};\boldsymbol{\theta}_k)=\prod_{k=1}^{K}(\lambda_k p(\mathbf{w};\boldsymbol{\theta}_k))^{c_k} \tag{4.2}$$

4.2.2 参数估计

模型的对数似然函数（log-likelihood）为：

$$L=\log p(X;\Theta)=\log p(D,C;\Theta)=\sum_{n=1}^{N}\sum_{k=1}^{K}c_{n,k}\log(\lambda_k p(\mathbf{w}_n;\boldsymbol{\theta}_k)) \tag{4.3}$$

在模型中引入一个 $n\times k$ 参数矩阵 G，表示文档属于某个聚类的可能性，称为聚类概率矩阵。它的元素为：

$$g_{n,k}=\mathrm{E}(c_{n,k}\mid D;\Theta)=P(c_{n,k}=1\mid\mathbf{w}_n;\Theta)=\frac{\lambda_k p(\mathbf{w}_n;\boldsymbol{\theta}_k)}{\sum_{k'=1}^{K}\lambda_{k'}p(\mathbf{w}_n;\boldsymbol{\theta}_{k'})} \tag{4.4}$$

文档依概率矩阵 G“软聚类”，模型的对数似然函数为：

$$\mathcal{L}=\log p(X;\Theta,G)=\sum_{n=1}^{N}\sum_{k=1}^{K}g_{n,k}\log(\lambda_k p(\mathbf{w}_n;\boldsymbol{\theta}_k)) \tag{4.5}$$

对含有隐变量的模型进行参数估计，常用的方法是期望一最大化（Expectation Maximization，EM）算法。EM 算法通过求期望和最大化两步过程的不断迭代，可以收敛到似然函数的局部最优解。EM 算法开始时，设置初始参数为 $\hat{\Theta}^{(0)}$，定义第 t 次迭代得到的参数的期望为 $\hat{\Theta}^{(t)}$。

期望步：根据上一次的参数估计 $\hat{\Theta}^{(t-1)}$ 计算本次的期望 $\hat{\Theta}^{(t)}$。

$$Q(\Theta;\hat{\Theta}^{(t-1)}) = \mathrm{E}(\log p(X;\Theta) \mid D;\hat{\Theta}^{(t-1)}) = \log p(X;\Theta,G) \tag{4.6}$$

$$g_{n,k}^{(t)} = \frac{\hat{\lambda}_k^{(t-1)} p(\mathbf{w}_n;\hat{\theta}_k^{(t-1)})}{\sum_{k'=1}^{K} \hat{\lambda}_{k'}^{(t-1)} p(\mathbf{w}_n;\hat{\theta}_{k'}^{(t-1)})} \tag{4.7}$$

最大化步：计算使最大似然函数 $Q(\cdot)$ 达到最大值的参数值并更新。参数的最大似然估计是：

$$\hat{\Theta}_{ML}^{(t)} = \arg\max_{\Theta}(Q(\Theta;\hat{\Theta}^{(t-1)})) \tag{4.8}$$

最大后验估计是：

$$\hat{\Theta}_{MAP}^{(t)} = \arg\max_{\Theta}(Q(\Theta;\hat{\Theta}^{(t-1)}) + p(\Theta)) \tag{4.9}$$

用 EM 算法进行参数估计的过程，就是不断重复期望步和最大化步，直至达到收敛标准。为加快 EM 算法的收敛速度，本文采用一种改进的 EM 算法，称为 eM[32]和 EM 算法的唯一不同之处是收敛条件变为：

$$\frac{\mathcal{L}^{(t)} - \mathcal{L}^{(t-1)}}{\mathcal{L}^{(t)} - \mathcal{L}^{(0)}} \le 10^{-2} \tag{4.10}$$

其中 $\mathcal{L}^{(t)}$ 表示 t 次迭代的对数相似性，阈值是基于实际效果确定的。

4.2.3 特征选择

本文研究的有限混合模型基于 Naïve Bayes 假定，即如果给定类标记，特征（词）之间相互独立。根据这一假定，引入一种改进的“特征显著性”方法，对数据进行特征选择。

Law 等[76]提出的“特征显著性”是一组实数值参数，用于衡量每个特征（词）对聚类过程的相关程度。这种方法将特征选择作为一个参数估计问题，可以在聚类的同时完成。Law 等用一组隐变量 $\boldsymbol{\varphi} = \{\phi_1,...,\phi_M\}$ 来表示词对聚类的相关性，它的分量定义为：

$$\phi_m = \begin{cases} 1, \text{如果特征} m \text{与聚类相关} \\ 0, \text{反之} \end{cases} \tag{4.11}$$

可以注意到，对文本数据进行聚类时，各主题都有相应出现频率较高的词。例如影视文艺类的文档中常常出现“主演”、“上映”，财经金融类的文档中“股市”、“基金”等词出现频率很高。因此对文献[76]中定义的特征相关性进行改进，将它定义为矩阵形式：$\Phi = \{\phi_{k,m}\}$，并将其作为隐变量。模型的完全数据集为

$X = D \cup C \cup \Phi$。

$$\phi_{k,m} = \begin{cases} 1, \text{如果特征} m \text{与混合成员} k \text{相关} \\ 0, \text{反之} \end{cases} \tag{4.12}$$

假定与某成员相关的特征服从该成员对应的分布，与该成员不相关的特征服从一个共享的分布 $q(w_m;\gamma)$，γ 是参数。(4.2) 式变为：

$$p(\mathbf{x};\Theta) = p(\mathbf{w},\mathbf{c},\Phi;\Theta) = \prod_{k=1}^{K}[\lambda_k \prod_{m=1}^{M}(p(w_m;\theta_k)^{\phi_{k,m}} q(w_m;\gamma)^{1-\phi_{k,m}})]^{c_k} \tag{4.13}$$

其中 $\Theta = \{\lambda_k, \theta_k, \gamma\}$ 是模型的参数。本文把这种具有特征选择能力的有限混合模型记为 FMM-FS（finite mixture model with feature selection）。

模型的对数似然函数为：

$$\begin{aligned} &\mathcal{L} = \log p(X;\Theta,G,S) \\ &= \sum_{n=1}^{N}\sum_{k=1}^{K} g_{n,k}\{\log \lambda_k + \sum_{m=1}^{M}[s_{k,m} \log p(w_{n,m};\theta_k) + (1-s_{k,m}) \log q(w_{n,m};\gamma)]\} \end{aligned} \tag{4.14}$$

其中参数矩阵 $S = \{s_{k,m}\}$ 表示特征和某个混合成员相关的概率，即特征显著性矩阵。

模型中各参数的最大似然估计可以用 EM 算法来求解得到。Q 函数为：

$$Q(\Theta;\hat{\Theta}^{(t-1)}) = \log p(X;\Theta,G,S) \tag{4.15}$$

在 EM 算法的期望步更新参数矩阵 G，根据 Bayes 定理，有：

$$\begin{aligned} g_{n,k}^{(t)} &= \mathrm{E}(c_{n,k} \mid D;\hat{\Theta}^{(t-1)}) = P(c_{n,k}=1 \mid \mathbf{w}_n;\hat{\Theta}^{(t-1)}) \\ &= \frac{\hat{\lambda}_k^{(t-1)} \prod_{m=1}^{M} p(w_{n,m};\hat{\theta}_k^{(t-1)})^{\hat{s}_{k,m}^{(t-1)}} q(w_{n,m};\hat{\gamma}^{(t-1)})^{(1-\hat{s}_{k,m}^{(t-1)})}}{\sum_{k'=1}^{K} \hat{\lambda}_{k'}^{(t-1)} \prod_{m=1}^{M} p(w_{n,m};\hat{\theta}_{k'}^{(t-1)})^{\hat{s}_{k',m}^{(t-1)}} q(w_{n,m};\hat{\gamma}^{(t-1)})^{(1-\hat{s}_{k',m}^{(t-1)})}} \end{aligned} \tag{4.16}$$

在最大化步更新矩阵 S。由 (4.13) 式，有：

$$P(\phi_{k,m}=1, w_{n,m} \mid c_{n,k}=1;\hat{\Theta}^{(t-1)}) \propto p(w_{n,m};\hat{\theta}_k^{(t-1)})^{\hat{s}_{k,m}^{(t-1)}} \tag{4.17}$$

$$P(\phi_{k,m}=0, w_{n,m} \mid c_{n,k}=1;\hat{\Theta}^{(t-1)}) \propto p(w_{n,m};\hat{\theta}_k^{(t-1)})^{1-\hat{s}_{k,m}^{(t-1)}} \tag{4.18}$$

由全概率公式，有：

$$\begin{aligned} &P(\phi_{k,m}=1, c_{n,k}=1 \mid \mathbf{w}_n;\hat{\Theta}^{(t-1)}) \\ &= \frac{P(\phi_{k,m}=1, w_{n,m} \mid c_{n,k}=1;\hat{\Theta}^{(t-1)}) \cdot P(c_{n,k}=1 \mid \mathbf{w}_n;\hat{\Theta}^{(t-1)})}{\sum_{\phi=0}^{1} P(\phi_{k,m}=\phi, w_{n,m} \mid c_{n,k}=1;\hat{\Theta}^{(t-1)})} \\ &= \frac{p(w_{n,m};\hat{\theta}_k^{(t-1)})^{\hat{s}_{k,m}^{(t-1)}} \cdot g_{n,k}^{(t)}}{p(w_{n,m};\hat{\theta}_k^{(t-1)})^{\hat{s}_{k,m}^{(t-1)}} + q(w_{n,m};\hat{\gamma}^{(t-1)})^{1-\hat{s}_{k,m}^{(t-1)}}} \end{aligned} \tag{4.19}$$

$$P(\phi_{k,m}=0,c_{n,k}=1\mid \mathbf{w}_n;\hat{\Theta}^{(t-1)})=\frac{p(w_{n,m};\hat{\boldsymbol{\theta}}_k^{(t-1)})^{(1-\tilde{s}_{k,m}^{(t-1)})}\cdot g_{n,k}^{(t)}}{p(w_{n,m};\hat{\boldsymbol{\theta}}_k^{(t-1)})^{\tilde{s}_{k,m}^{(t-1)}}+q(w_{n,m};\hat{\boldsymbol{\gamma}}^{(t-1)})^{1-\tilde{s}_{k,m}^{(t-1)}}} \quad (4.20)$$

$$\begin{aligned} s_{k,m}^{(t)}=P(\phi_{k,m}=1)&=\frac{\sum_{n=1}^{N}P(\phi_{k,m}=1,c_{n,k}=1\mid \mathbf{w}_n;\hat{\Theta}^{(t-1)})}{\sum_{n=1}^{N}\sum_{\phi=0}^{1}P(\phi_{k,m}=\phi,c_{n,k}=1\mid \mathbf{w}_n;\hat{\Theta}^{(t-1)})} \\ &=\frac{1}{N}\sum_{n=1}^{N}\frac{p(w_{n,m};\hat{\boldsymbol{\theta}}_k^{(t-1)})^{\tilde{s}_{k,m}^{(t-1)}}\cdot g_{n,k}^{(t)}}{p(w_{n,m};\hat{\boldsymbol{\theta}}_k^{(t-1)})^{\tilde{s}_{k,m}^{(t-1)}}+q(w_{n,m};\hat{\boldsymbol{\gamma}}^{(t-1)})^{1-\tilde{s}_{k,m}^{(t-1)}}} \end{aligned} \quad (4.21)$$

参数 λ_k，$\boldsymbol{\theta}_k$ 和 $\boldsymbol{\gamma}$ 根据 $p(\cdot;\boldsymbol{\theta}_k)$ 和 $q(\cdot;\boldsymbol{\gamma})$ 的分布进行相应的更新。在 4.3 节给出了多项式混合的实例来加以说明。

4.2.4 基于准则的模型选择

用 EM 算法估计模型的参数，需要预先确定混合成员的个数，即聚类的数量，这就是模型选择问题。现有的大部分模型选择方法是依据某种准则对不同混合成员数量的模型进行比较，选择最优的模型。此外还有一种统计学方法，根据 Markov Chain Monte Carlo（MCMC）采样，同时完成模型选择和模型的参数估计。由于 MCMC 采样方法计算量巨大[32]，因此对大规模数据聚类的模型选择问题，一般采用基于准则的方法。

模型的成员数量根据某种准则 $\mathrm{Cr}(\Theta(K),K)$ 确定，其中 $\Theta(K)$ 是成员个数为 K 时模型的参数。模型选择的过程就是根据不同的准则，选择使准则最大或者最小的 K 值。已提出的准则包括 BIC，AIC，MDL，MML 等。

4.2.5 一种广义方法

本文提出一种用有限混合模型对文本聚类的广义方法或框架，它由两步组成：首先，采用 *v*-fold 方式，在数据集的若干小规模子集上进行模型选择和特征选择；然后，在全部数据集上对模型进行参数估计。

方法的第一步是模型选择和特征选择，算法描述如下：

算法 4.1　有限混合的模型选择和特征选择算法

输入：文档集 D，词汇表集 $\mathbf{w}$，*v*-fold 交叉检验次数 v，混合成员数量的最大值 $K_{\max}$ 和最小值 $K_{\min}$，特征选择阈值 h；

输出：最佳混合成员数 $\hat{K}$，选择的特征子集 $\mathbf{f}$；

步骤：1，将数据集随机划分为大小相同的 v 个子集 $D^t(t=1,\ldots,v)$，对每个子集运行步骤 2 至 8，并用 D' 替代前面公式中的 D；

2，令 $K=K_{max}$；

3，初始化模型参数 $\Theta(K)$：随机设置各混合模型的参数值，设置特征显著性矩阵的所有元素为 0.5；

4，设置 $\Theta=\Theta(K)$，运行 EM 算法直至收敛；

5，计算模型选择准则值 $\mathrm{Cr}(\Theta(K),K)$，并记录估计得到的特征显著性矩阵 S 为 $S^{(K)}$；

6，如果 $K=K_{min}$，转到步骤 8）；

7，令 $K=K-1$，转到步骤 3）；

8，设置 $\hat{K}^t=\arg(\max/\min)_K(\mathrm{Cr}(\Theta(K),K))$，并记录特征显著性值高于 h 的特征子集 $\mathbf{f}^t$；

9，令 $\hat{K}=\left\lfloor\frac{1}{v}\sum_{t=1}^{v}\hat{K}^t\right\rfloor$，$\mathbf{f}=\bigcup_{t=1}^{v}\mathbf{f}^t$，返回 $\hat{K}$ 和 $\mathbf{f}$。

算法结束

第二步，将上一步得到的特征集作为词汇表，混合成员的数量也由上一步确定，在全部文档集上对模型进行参数估计。特征选择完成后，混合模型也可不加入特征显著性的隐变量，它的形式是（4.2）式。

与其他基于有限混合模型的聚类方法相比，本文提出的广义方法具有以下特点：首先，它是面向文本聚类问题的一种综合性方法，不仅用有限混合对文本建模和通过估计模型参数完成聚类，还包括模型选择和特征选择，并将它们纳入一个统一的框架中。由于采用了特征显著性的概念，将特征选择转化为混合模型中的参数估计问题，成为广义模型统一框架中的一部分，而不需另外的特征选择步骤。其次，模型选择和特征选择过程需要多次迭代，计算量比较大。采用 v-fold 的方式，可以有效减少单次计算量，代价是增大了模型不足拟和的风险。第三，方法对有限混合的具体形式未做限定，各种有限混合都可以被引入这个广义框架，同样也可以应用各种模型选择准则，因此称为广义方法，具有很好的灵活性和适应性。

4.3 广义模型的一种应用

本文提出的广义方法有很好的灵活性，Gaussian，多项式，Dirichlet，von Mises-Fisher 等各种有限混合均可引入其中。框架还包含了基于准则的一般性模型选择方法，同样可以应用 BIC，AIC，MDL，MML 等各种模型选择准则。在此框架下，本节对一种多项式混合模型以及 BIC 模型选择方法进行了具体研究。

4.3.1 多项式混合模型

Nigam 等[38]给出了一种用多项式混合（multinomial mixture）对文本建模的方法。以这种混合模型为例来具体说明本文提出的广义方法。

这一模型的文档生成过程是：首先，依概率 $\boldsymbol{\lambda}=(\lambda_1,...,\lambda_K)$ 生成文档 $\vec{w}$ 的主题；然后，在词汇表的所有词上依一个多项式分布生成文档向量，参数是 $\boldsymbol{\theta}_k=(\theta_{k,1},...,\theta_{k,M}), k=1,...,K$。混合模型的表达式为：

$$p(\mathbf{w};\Theta)=\sum_{k=1}^{K}\lambda_k p(\mathbf{w},\boldsymbol{\theta}_k)=\sum_{k=1}^{K}\lambda_k\frac{(\sum_{m=1}^{M}w_m)!}{\prod_{m=1}^{M}w_m!}\prod_{m=1}^{M}\theta_{k,m}^{w_m} \qquad (4.22)$$

其中 $\Theta=\{\lambda_1,...,\lambda_K,\boldsymbol{\theta}_1,...,\boldsymbol{\theta}_K\}$ 是混合模型的参数集。$\theta_{k,m}$ 是第 k 个混合成员的参数向量的第 m 个分量，$\sum_{m=1}^{M}\theta_{k,m}=1, k=1,...,K$。$\lambda_k$ 是混合权重，$\sum_{k=1}^{K}\lambda_k=1$，$\lambda_k>0, k=1,...,K$。

依照前面的广义模型，在多项式混合中加入指示向量 $\mathbf{c}=(c_1,...,c_K)$。$\mathbf{w}$ 和 $\mathbf{c}$ 构成完全数据集 $\mathbf{x}$，它的多项式混合表达式是：

$$p(\mathbf{x};\Theta)=p(\mathbf{w},\mathbf{c};\Theta)=\prod_{k=1}^{K}(\lambda_k\frac{(\sum_{m=1}^{M}w_m)!}{\prod_{m=1}^{M}w_m!}\prod_{m=1}^{M}\theta_{k,m}^{w_m})^{c_k} \qquad (4.23)$$

4.3.2 参数估计

用多项式混合对文档建模，模型的对数似然函数（软聚类意义下，忽略常数）为：

$$\mathcal{L}=\sum_{n=1}^{N}\sum_{k=1}^{K}g_{n,k}(\log\lambda_k+\sum_{m=1}^{M}w_{n,m}\log\theta_{k,m}) \qquad (4.24)$$

为了估计模型的参数，求似然函数的最大后验（MAP）估计，引入对应于参

数$\{\lambda_k\}$和$\{\theta_{k,m}\}$的 Dirichlet 先验，定义二者的 Dirichlet 超参数为d_λ和d_θ。这种通过加入 Dirichlet 先验对模型进行 MAP 估计的方法也称 Laplace 平滑。

用 EM 算法迭代更新模型的参数，期望步是：

$$g_{n,k}^{(t)} = P(c_{n,k}=1|\mathbf{w}_n;\hat{\Theta}^{(t-1)}) = \frac{\hat{\lambda}_k^{(t-1)}\prod_{m=1}^{M}(\hat{\theta}_k^{(t-1)})^{w_{n,m}}}{\sum_{k'=1}^{K}\hat{\lambda}_{k'}^{(t-1)}\prod_{m=1}^{M}(\hat{\theta}_{k'}^{(t-1)})^{w_{n,m}}} \tag{4.25}$$

最大化步更新参数$\{\lambda_k\}$和$\{\theta_{k,m}\}$：

$$\lambda_k^{(t)} \propto d_\lambda - 1 + \sum_{n=1}^{N} g_{n,k}^{(t)} \tag{4.26}$$

$$\theta_{k,m}^{(t)} \propto d_\theta - 1 + \sum_{n=1}^{N} w_{n,m} g_{n,k}^{(t)} \tag{4.27}$$

Rigouste 等[80]在 Reuters 数据集上的实验表明d_λ的值对聚类几乎没有影响，而d_θ在 0.1 和 2 之间取值聚类效果较好。取$d_\lambda = 1$，$d_\theta = 2$。由参数的约束条件$\sum_{k=1}^{K}\lambda_k = 1$和$\sum_{m=1}^{M}\theta_{k,m} = 1, k = 1,...,K$，有：

$$\lambda_k^{(t)} = \frac{d_\lambda - 1 + \sum_{n=1}^{N} g_{n,k}^{(t)}}{\sum_{k'=1}^{K}(d_\lambda - 1 + \sum_{n=1}^{N} g_{n,k'}^{(t)})} = \frac{\sum_{n=1}^{N} g_{n,k}^{(t)}}{N} \tag{4.28}$$

$$\theta_{k,m}^{(t)} = \frac{d_\theta - 1 + \sum_{n=1}^{N} w_{n,m} g_{n,k}^{(t)}}{\sum_{m=1}^{M}(d_\theta - 1) + \sum_{m'=1}^{M}\sum_{n=1}^{N} w_{n,m'} g_{n,k}^{(t)})} = \frac{1 + \sum_{n=1}^{N} w_{n,m} g_{n,k}^{(t)}}{M + \sum_{m'=1}^{M}\sum_{n=1}^{N} w_{n,m'} g_{n,k}^{(t)})} \tag{4.29}$$

4.3.3 特征选择

在多项式混合中引入表示特征与成员相关性的隐变量矩阵Φ以及与成员不相关的特征服从的共享分布$q(\cdot;\gamma)$，称这种模型为 MM-FS（multinomial mixture with feature selection），它的形式为：

$$p(\mathbf{x};\Theta) = p(\mathbf{w},\mathbf{c},\Phi;\Theta) = \prod_{k=1}^{K}[\lambda_k \frac{(\sum_{m=1}^{M} w_m)!}{\prod_{m=1}^{M} w_m!}\prod_{m=1}^{M}(\theta_{k,m}^{w_m})^{\phi_{k,m}}(\gamma_m^{w_m})^{1-\phi_{k,m}}]^{c_k} \tag{4.30}$$

对数似然函数为：

$$\mathcal{L}=\sum_{n=1}^{N}\sum_{k=1}^{K}g_{n,k}\{\log\lambda_k+\sum_{m=1}^{M}w_{n,m}[s_{k,m}\log\theta_{k,m}+(1-s_{k,m})\log\gamma_m]\} \tag{4.31}$$

其中 G，S 分别是前面定义的聚类概率矩阵和特征显著性矩阵。

应用 Laplace 平滑方法，在模型中加入参数 $\{\lambda_k\}$，$\{\theta_{k,m}\}$ 和 $\{\gamma_m\}$ 的 Dirichlet 先验，超参数分别为 $d_\lambda=1$，$d_\theta=2$ 和 $d_\gamma=2$。

在期望步更新 G 的元素：

$$\begin{aligned}g_{n,k}^{(t)}&=P(c_{n,k}=1\mid\mathbf{w}_n;\hat{\Theta}^{(t-1)})\\&=\frac{\hat{\lambda}_k^{(t-1)}\prod_{m=1}^{M}[(\hat{\theta}_{k,m}^{(t-1)})^{\tilde{s}_{k,m}^{(t-1)}}(\hat{\gamma}_m^{(t-1)})^{(1-\tilde{s}_{k,m}^{(t-1)})}]^{w_{n,m}}}{\sum_{k'=1}^{K}\hat{\lambda}_{k'}^{(t-1)}\prod_{m=1}^{M}[(\hat{\theta}_{k',m}^{(t-1)})^{\tilde{s}_{k',m}^{(t-1)}}(\hat{\gamma}_m^{(t-1)})^{(1-\tilde{s}_{k',m}^{(t-1)})}]^{w_{n,m}}}\end{aligned} \tag{4.32}$$

在最大化步根据（4.21）式更新参数矩阵 S 的元素：

$$s_{k,m}^{(t)}=\frac{1}{N}\sum_{n=1}^{N}\frac{(\hat{\theta}_{k,m}^{(t-1)})^{w_{n,m}\tilde{s}_{k,m}^{(t-1)}}\cdot g_{n,k}^{(t)}}{(\hat{\theta}_{k,m}^{(t-1)})^{w_{n,m}\tilde{s}_{k,m}^{(t-1)}}+(\hat{\gamma}_k^{(t-1)})^{w_{n,m}(1-\tilde{s}_{k,m}^{(t-1)})}} \tag{4.33}$$

同时更新模型的其他参数：

$$\lambda_k^{(t)}=\frac{\sum_{n=1}^{N}g_{n,k}^{(t)}}{N} \tag{4.34}$$

$$\theta_{k,m}^{(t)}=\frac{1+\sum_{n=1}^{N}w_{n,m}s_{k,m}^{(t)}g_{n,k}^{(t)}}{M+\sum_{m'=1}^{M}\sum_{n=1}^{N}w_{n,m'}s_{k,m'}^{(t)}g_{n,k}^{(t)}} \tag{4.35}$$

$$\gamma_m^{(t)}=\frac{1+\sum_{n=1}^{N}\sum_{k=1}^{K}w_{n,m}(1-s_{k,m}^{(t)})g_{n,k}^{(t)}}{M+\sum_{m'=1}^{M}\sum_{n=1}^{N}\sum_{k=1}^{K}w_{n,m'}(1-s_{k,m'}^{(t)})g_{n,k}^{(t)}} \tag{4.36}$$

4.3.4 BIC 模型选择准则

Bayesian 信息准则（BIC）[56]是一种在统计学习领域广泛应用的模型选择准则，用于衡量模型表现数据的能力。有限混合模型的成员个数为 K 时，BIC 的形式为：

$$BIC(K)=\log p(X\mid\hat{\Theta}(K))-\frac{1}{2}V(K)\log N \tag{4.37}$$

其中 $\hat{\Theta}(K)$ 是模型参数 $\Theta(K)$ 的最大似然估计，$V(K)$ 是模型中需要估计的独立参数的数量。BIC 的值越大，表明模型反映数据的能力越强。

将 BIC 准则引入广义框架，对多项式混合进行模型选择。模型选择和特征选择过程见算法 4.1。

4.4 实验与讨论

通过实验考察本文提出的聚类方法的性能。具体考察以下问题：

第一，本文提出了一种包含特征选择功能的多项式混合模型，它的实际聚类效果如何？对此，在多个数据集上，将本文提出的模型与其他常见聚类方法进行性能比较。

第二，广义方法包括模型选择和特征选择，这二者是否能达到可接受的效果？以多项式混合模型和 BIC 模型选择准则为例，考察模型选择和特征选择的性能。

4.4.1 文档数据集

实验中用到了 4 种标准文档数据集：Classic-3[1]，Hitech-6[2]，Reuters-21578[3]和 Newsgroup-20[4]。

Classic-3 数据集有 3,891 个文档，6,729 个词和 3 个类。它由三个数据集的文档组成，1,398 个 CRANFIELD（航空系统论文）文档，1,033 个 MEDLINE（医学期刊）文档和 1,460 个 CISI（信息检索论文）文档。

Hitech-6 数据集有 1,530 个文档，10,919 个词和 6 个类。这个数据集的文档来自 San Jose Mercury 报的文章，它的 6 个类别是计算机，电子，健康，医药，研究和技术。

Reuters-21578 数据集有 21,578 个文档，文档内容来自 Reuters 的新闻文章。由于各类中的文档极不均衡，取它的一个子集，由包含文档数最多的 10 个类的文档组成，记为 Reuters-10。Reuters-10 数据集有 5,771 个文档，15,996 个词和 10 个类。

Newsgroups-20 数据集有 20,000 个文档，文档内容来自新闻组的讨论文章。对文档数据集进行预处理，去除空文档和在全部文档中出现次数少于 3 次的词，得到的数据集有 19,949 个文档，43,586 个词和 20 个类。

4.4.2 评价指标

采用在文本聚类领域常用的 F-measure 和 NMI 两种指标来评价聚类方法的效

[1] ftp://ftp.cs.cornell.edu/pub/smart/

[2] http://trec.nist.gov

[3] http://www.daviddlewis.com/resources/testcollections/reuters21578/

[4] http://kdd.ics.uci.edu/databases/20newsgroups/20newsgroups.html

果。

F-measure[81]是一种综合了准确率（precision）和召回率（recall）的聚类评价指标。每个聚类中的文档被视为一个查询的结果，准确率和召回率分别定义为：

$$precision_{i,j} = \frac{N_{i,j}}{N_j} \tag{4.38}$$

$$recall_{i,j} = \frac{N_{i,j}}{N_i} \tag{4.39}$$

其中 $N_{i,j}$ 是实际类标记为 i 并被划分到聚类 j 中的文档数，N_i 是类标记为 i 的全部文档数，N_j 是聚类 j 中的全部文档数，N 是全部文档数。

聚类 j 和类标记 i 的 F-measure 定义为：

$$F_{i,j} = \frac{2precision_{i,j} \cdot recall_{i,j}}{precision_{i,j} + recall_{i,j}} \tag{4.40}$$

对全部聚类结果进行加权求和，可得聚类的总体 F-measure：

$$F = \sum_i \frac{n_i}{n} \max_j F_{i,j} \tag{4.41}$$

F-measure 的取值范围为(0,1)，值越大表示聚类效果越好。

归一化互信息（normalized mutual information，NMI）[82]也是一种常见的聚类评价指标。假设 S 是表示某文档所属聚类的随机变量，T 是表示该文档实际类标记的随机变量，则聚类结果与类标记的互信息为：

$$I(S,T) = \sum_{s \in S, t \in T} p(s,t) \log \frac{p(s,t)}{p(s)p(t)} \tag{4.42}$$

根据 S 和 T 的熵 $H(S)$ 和 $H(T)$，对互信息进行归一化，得到 NMI：

$$NMI = \frac{I(S,T)}{\sqrt{H(S)H(T)}} = \frac{\sum_{i,j} N_{i,j} \log(\frac{N \cdot N_{i,j}}{N_i N_j})}{\sqrt{(\sum_i N_i \log \frac{N_i}{N})(\sum_j N_j \log \frac{N_j}{N})}} \tag{4.43}$$

NMI 的取值范围为(0,1)，值越大表示聚类结果与实际类标记越相符合。

4.4.3 实验结果

以 4.3 节提出的有特征选择功能的多项式混合模型（MM-FS）以及 BIC 模型选择准则为例，通过实验考察广义方法的聚类性能。以下的实验结果，如未作特殊说明，都是采用 v-fold（v=20）方法，特征选择阈值为 0.5，重复 10 次的平均值。

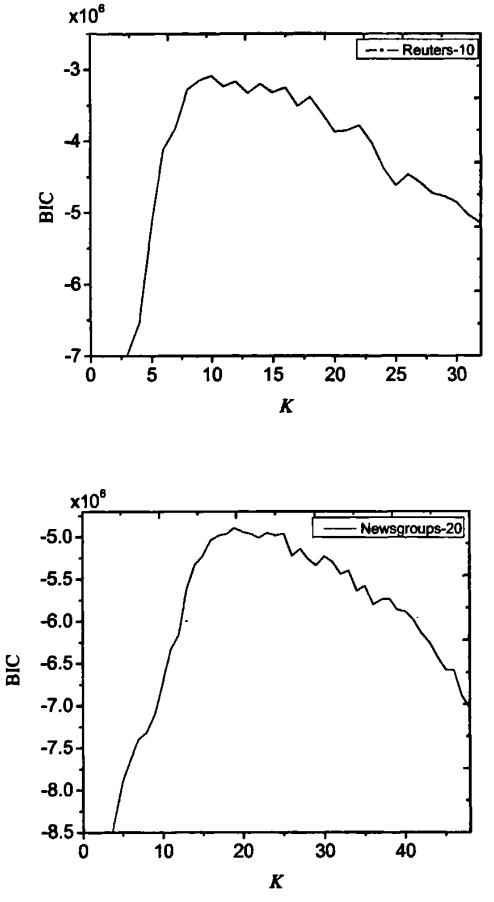


图 4.1 MM-FS 在 Reuters-10（上）和 Newsgroups-20（下）上的 BIC 曲线

表 4.1 MM-FS 在各数据集上的模型选择和特征选择效果

	Classic-3	Hitech-6	Reuters-10	Newsgroups-20
实际类数	3	6	10	20
计算得到的混合数	3	6	10	19
全部词汇数	41,681	10,080	15,996	43,586
选择得到的特征数	2,387	2,096	1,885	3,950

首先考察广义方法的模型选择和特征选择效果。图 4.1 是 MM-FC 方法在 Reuters-10 和 Newsgroups-20 数据集上进行模型选择的 BIC 曲线。最大成员数分别初始化为 32 和 48，图中所示 BIC 值是采用 v-fold （v=20）方法各子集上 BIC 值的平均值。表 4.1 是 MM-FS 在各数据集上模型选择和特征选择结果。从图 4.1 和

表 4.1 中可见，MM-FS 配合 BIC 模型选择准则可以较准确地确定文本数据集聚类的数目，同时通过特征选择极大地降低了原特征空间的维度。对模型的下一步学习过程而言，准确的模型成员数量，较小规模的有效特征集可以提高参数估计的准确性，降低其计算复杂度。

表 4.2 MM-FS 在 Classic-3 数据集上特征选择的效果

特征选择阈值(h)	选择的特征数	F-measure	NMI
0	41,681	0.92	0.83
0.1	35,844	0.93	0.84
0.3	16,072	0.96	0.88
0.5	2,387	0.94	0.86
0.7	1,264	0.81	0.79

表 4.3 MM-FS 在 Classic-3 数据集上选择的最佳特征

聚类	最佳的 10 个特征
1	cells, patients, hormone, blood, cancer, children, renal, deaths, abortions, rats
2	aero, heat, engine, plane, wing, supersonic, laminar, shock, transfer, power
3	retrieval, system, libraries, research, library, science, scientific, computer, information, literature

在 Classic-3 数据集上，MM-FS 的特征选择效果如表 4.2 和表 4.3 所示。表 4.2 是在设定不同特征选择阈值 h 的情况下得到的有效特征数，和以这些特征为词汇表对全部文档集进行聚类的聚类效果（评价指标为 F-measure 和 NMI）。其中 h=0 表示不进行特征选择，算法直接在原有词汇表上进行的情况。一般来说，特征选择阈值越小，选择得到的特征越多，模型也更加复杂。聚类算法通过选择一部分最"有用"的特征，适当降低模型复杂性，以减少计算量。由表 4.2 可见，当特征选择阈值为[0.5, 0.7]之间时，算法可以极大降低特征维数（选择的特征集规模为原特征集的 5%左右），同时达到较好的聚类效果。表 4.3 是 MM-FS 得到的数据集在各聚类上的前 10 个特征的列表。由表中可见，选择到的特征基本反映了构成数据集的三类文档各自相关的内容。

最后，在 4 个文本数据集上对几种常见的聚类方法进行了比较实验，实验结果如表 4.4 所示。选用的算法分别是 *k*means（随机设置初始中心点），多项式混合（MM）[38]，提出的带特征选择的多项式混合（MM-FS）以及基于相似性矩阵特征向量的谱聚类（spectral）[83]方法。MM-FS 与 MM 的主要区别是前者将特征选

择作为模型参数估计的一部分，后者不包括特征选择。需要说明的是，为得到最佳的聚类效果，在 MM-FS 方法中直接指定模型的混合成员数为数据集的实际类别数。从实验结果可见，在 F-measure 和 NMI 两项指标上均达到或接近达到 4 种聚类方法的最优值。可以认为，MM-FS 对文档数据集的建模与其他几种方法相比更为准确。特别是与没有特征选择能力的多项式混合相比，聚类性能有明显的提高。这可能是由于通过特征选择，降低了不相关的特征对聚类的影响，从而相应提高了建模的准确性。

表 4.4 各种聚类方法的聚类效果

数据集	F-measure				NMI			
	*k*means	MM	MM-FS	spectral	*k*means	MM	MM-FS	spectral
Classic-3	.66±.08	.92±.01	.96±.01	**.97±.02**	.40±.06	.83±.03	.88±.01	**.93±.02**
Hitech-6	.40±.07	.41±.02	**.44±.01**	.43±.02	.20±.05	.23±.04	**.26±.01**	.24±.01
Reuters-10	.43±.04	.67±.01	**.72±.02**	.68±.03	.33±.06	.51±.02	**.54±.01**	.52±.01
Newsgroups-20	.11±.03	.45±.03	**.48±.02**	**.48±.03**	.08±.04	.55±.03	.57±.02	**.58±.02**

4.4.4 相关工作

本文提出的广义方法很大程度上受 Law 等[76]的启发：与文献[76]的区别在于：第一，不仅限于 Gaussian 混合，多项式混合等更适于文本建模的混合模型。第二，特征显著性的定义不同，Law 的特征显著性定义为一组向量，表示特征对全部聚类的相关性；本文的特征显著性是矩阵的形式，表示特征对各聚类的相关性。第三，Law 的模型选择方法基于 MML 准则；本文的广义方法不仅限于 MML 准则，NEC[84]，BIC 等各种模型选择方法都可应用于广义方法中。

与本文类似的工作还有文献[85]，同样是结合了特征选择的混合模型聚类方法。本文工作与文献[85]的最大区别是：本文中的特征显著性的定义是特征与某个聚类相关的概率，并将特征相关性作为混合模型中的隐变量；G 的特征显著性是二进制值，即将特征相关性直接作为混合模型的参数而不是隐变量。此外，文献[85]提出的混合模型参数估计过程中，EM 算法的极大似然函数是负（极小化）BIC 开销，这一点也广义模型有区别。

4.4.5 算法复杂度

对 MM-FS 模型，如果不考虑模型选择过程，只进行特征选择和聚类，对混合

模型进行参数估计时 E 步和 M 步运行一次的时间复杂度都是 $O(NMK)$。假设 EM 算法的迭代次数是 I，则混合模型聚类一次的时间为 $O(2INMK)$。假设在特征选择后，用于聚类的特征数为 F，则聚类的时间复杂度为 $O(2INFK)$。

如果预先在 v 个规模为 ρN 的子集上进行模型选择，时间为 $O(2\rho vINMK)$。总的时间复杂度为 $O(2INK(F+\rho vM))$。

由于广义方法中的有限混合模型是一种生成性模型，在文档来源比较稳定的情况下，只要进行一次模型选择和特征选择，之后的聚类过程可以在确定的聚类数目和较少的特征集上进行。如果采用适合的启发式方法，模型选择的时间还可以进一步减少。

4.4.6 方法的局限性

广义方法的主要局限之处在于采用了传统的模型选择方法，混合成员数目的确定需要进行多次尝试。将来的工作将考虑采用 Figueiredo 等[32]的做法，将 MML 模型选择准则作为混合模型似然函数的一部分。广义方法的另一个局限是只适用于基于词的文档模型，如多项式混合，Gaussian 混合，Von Mises 混合等；基于文档的模型，如 Dirichlet 混合，LDA 模型以及 PLSA 模型，由于它们的概率模型中文档无法表示为在各个词上的概率分布，因此不适用于本文提出的广义方法。

4.5 本章小结

本文提出了用有限混合模型进行无监督文本聚类的一种广义方法，这种方法将模型选择，特征选择以及混合模型的参数估计纳入一个统一的框架中。特别地，将指示文档特征的向量作为隐变量加入有限混合模型中，特征选择问题就转化成对作为模型参数的特征显著性矩阵的最大似然估计，可通过 EM 算法求解得到。在这个广义方法框架下，发展了一种带特征选择的多项式混合模型（MM-FS），并以 BIC 准则作为模型选择方法。

第五章 基于密度的半监督学习模型

在大量数据挖掘的实际问题中，人的手工标记由于成本昂贵而难于获取，因此如何利用有限的监督信息构造正确的学习器成为当前机器学习研究领域的一个热点问题。在很多情况下，数据集包括大量无标记的数据，并辅以少量先验知识作为监督信息。研究人员提出了半监督学习方法处理这一类问题。半监督学习包括半监督聚类和半监督分类，本章对这两类半监督学习问题分别加以研究。

5.1 半监督学习问题概述

半监督学习是机器学习中一个相对较新的子领域，近年来受到了广泛的关注和研究。在许多实际问题中，有标记的数据一般需要专家的人工分类得到，通常较为困难、昂贵且需要大量时间。而大量无标记的数据相对较容易收集，但在有监督学习中，无法作为学习的依据。为解决这一问题，研究人员提出了各种半监督学习方法。通过同时使用有标记和无标记信息，半监督学习方法可以得到比有监督方法更好的学习器。由于半监督学习需要更少的人工劳动，可以得到更好的学习精度，因此它的理论和实践都得到了重要发展。

半监督学习包括半监督分类和半监督聚类，它们的定义和概念已在本文第一章给出。下面分别介绍这两类问题的研究现状，并提出本文的研究思路。

5.1.1 半监督分类

半监督分类中监督信息一般是少量的有标记样本。如何利用大量无标记样本的结构信息，研究人员由不同角度出发提出了各种方法。现有的半监督分类方法包括混合模型和 EM 算法[37]，自训练（self-training）[39,86]，协同训练（co-training）[40]，基于图的方法[42]，以及 TSVM[41]等。

基于图的半监督分类方法[42,87,88]成为近期的研究热点之一。这一类方法一般是根据数据集建立一个图，图本身通常反映了整个数据集的流型特征。数据集中的有标记样本和无标记样本作为节点，图中的边反映了样本之间的相似性。然后，基于图的方法在图上构造一个分类器并对之加以平滑。

近期有关半监督分类的工作中，人们提出了许多基于图的方法，其中很大一部分基于随机行走（random walks）方法。随机行走也称布朗运动，是一种正态分布的独立增量随机过程。Szummer 和 Jaakkola[87]在加权数据图上提出了 Markov 随机行走方法，这种方法假设如果两个样本是相似的，那么很容易从其中之一“行走”到另一个。Zhu 等[88]将 Markov 随机行走加以扩展，将图学习问题定义为 Gaussian

随机场以及相应的调和函数。这样，某个样本的标记可以根据它的邻居上的随机行走来计算。最近，一些研究者提出了用于半监督分类的基于图的核方法（graph based kernel methods，GBKM）。Sindhwani 等[89]对半监督学习引入了流型学习方法，定义了一种图形核函数，将基于图的学习方法的转导（transiductive）性扩展为直导（inductive）性。这些 GBKM 的主要优点是它们与传统的核机器方法密切相关，同时利用有标记和无标记样本来构造图。

尽管 GBKM 在很多半监督学习问题中有成功的应用，图形核方法仍有进一步改进的可能。大部分现有的 GBKM 都使用径向基函数（radial basis function，RBF），数据集的密度分布信息并没有被充分利用。然而，和数据集相关的"定制"核被证明可以得到比标准 RBF 核更好的预测精度。相关工作包括 Fischer 等的连接性核（connectivity kernels）[90]和 Sajama 等的基于密度距离的聚类方法[91]。

本章第二节提出一种新型的非参数的基于密度的图形核算法用于半监督分类，新方法称为基于密度的 Laplacian 核（density-based Laplacian kernels，DBLK）。在 DBLK 方法中，图中的各个边根据本文定义的"边密度"被重新加权。然后采用"点云"（point cloud）Laplacian 核方法[89]，在图上建立直导 SVM 分类器。DBLK 方法可以完全利用数据集的结构信息来提升分类器的性能，计算得到的分类器可以对整个特征空间上的数据进行分类，从而突破了大部分 GBKM 的转导性质。

5.1.2 半监督聚类

半监督聚类中常见的先验知识表现为反映样本间相似关系的约束条件。参考 Wagstaff 等对约束条件的定义[92]，两个样本属于同一类即为 must-link，不属于同一类的则为 cannot-link。约束条件用于半监督聚类主要有两大类方法：基于距离的方法和基于约束的方法。前者根据约束构造某种距离度量并以此为基础运行各种聚类算法[92,93]；后者是将约束作为聚类目标的一部分直接作用于聚类算法[44,45,94]。

已有的半监督聚类方法很少将数据集的空间结构信息加以利用。基于距离的方法仅根据约束信息调整样本间的距离，而大部分情况下可用的约束条件数量较少，因而数据集的空间分布信息无法得到有效利用。一个自然的考虑是将数据集结构信息以某种形式引入基于约束的聚类方法中，以超越有限的约束条件来得到更好的聚类效果。相关工作包括 Wang 等提出的密度敏感的半监督谱聚类算法 DS-SSC[93]，构造样本间基于密度的距离矩阵，并在此基础上进行有约束条件的谱聚类。类似的工作还包括 Chapelle 等提出的半监督分类算法低密度分离（LDS）[95]，该方法通过寻找聚类之间的低密度区域来帮助半监督学习。

本章试图将数据集结构信息直接转化为约束条件，对约束进行扩展。首先将全部样本视为一个无向图，并提出一种基于密度的距离来度量图上两点间距离，然后根据样本点之间的距离关系和已知约束产生新的约束条件，作为半监督聚类算

法的依据。这种做法的好处是：将数据集结构信息引入聚类，在约束条件数量较少，不足以反映数据集分布特点时，可望达到更好的聚类效果；结构信息以约束条件的形式参加学习过程，不影响原有的聚类算法，适用于现有的各种半监督聚类算法。相关内容的详细介绍在本章第四节。

5.2 半监督分类中基于密度的 Laplacian 核方法

5.2.1 半监督分类问题的图形定义

大部分半监督分类问题中，数据集是 d 维向量集 $X=\{\mathbf{x}_1,...,\mathbf{x}_l,\mathbf{x}_{l+1},...,\mathbf{x}_n\}$，其中 $L=\{\mathbf{x}_1,...,\mathbf{x}_l\}$ 是有标记数据集，$U=\{\mathbf{x}_{l+1},...,\mathbf{x}_n\}$ 为无标记数据集。本文只考虑二值分类问题，因此标记 $Y=\{y_1,...,y_l\}$ 为一个二进制集，$y_i\in\{-1,+1\}$，$1\le i\le l$。采用一对多策略[96]，二值分类问题可以方便地扩展为多类问题。

由数据集可以建立一个图 $G=(V,E)$。其中节点是数据点，每个连接数据点的边都被赋予一个权重值，通常是节点的成对相似度。许多图形构建方法假定图是完全连接的，即每个节点和其他节点之间都存在边。连接节点 i 和节点 j 的边权重通常设置为 $w_{ij}'=\exp(-(\|\mathbf{x}_i-\mathbf{x}_j\|_2)^2/(2\sigma^2))$，其中参数 σ 是用于 Parzen 窗口密度估计的 Gaussian 分布宽度。在本章第四节的半监督聚类问题中，采用这种形式的边权重定义。但对半监督分类问题，本文根据 k 近邻（k nearest neighbor，kNN）的方法确定节点之间的边，图中的边在每个节点和与距离它最近的 k 个节点间设置，边权重设置为它们的欧氏距离：$w_{ij}'=\|\mathbf{x}_i-\mathbf{x}_j\|_2$。最后，令 $w_{ij}'=w_{ji}'$ 将边权重矩阵 $\mathbf{W}'=\{w_{ij}'\}$ 对称化。

5.2.2 kNN 密度距离

为了探寻数据的内在流型结构，很多方法在图上定义某种合适的距离或度量，使得属于同一流型的数据点相似度更高，位于不同流型中的数据点相似度较低。本文提出具有这样功能的一种结构简单的密度距离形式，称为 kNN 密度距离。正如它的名称所示，这种距离形式根据某个数据点的 k 个最近邻居节点定义它的密度。对一个数据点 $\mathbf{x}_i$，用 $\Gamma^i=\{\gamma_1^i,...,\gamma_k^i\}$ 表示它的 k 个最近邻居，将 $\mathbf{x}_i$ 的密度函数定义为：

$$d(\mathbf{x}_i)=g(\frac{1}{k}\sum_{l=1}^{k}\|\mathbf{x}_i-\gamma_l^i\|_2) \tag{5.1}$$

其中 $g(\cdot)$ 是一个单调递减的函数，定义为 $g(t)=\exp(-\alpha t)$，α（$\alpha>0$）是一个超参数，为方便起见，定义为 $\alpha=1$。

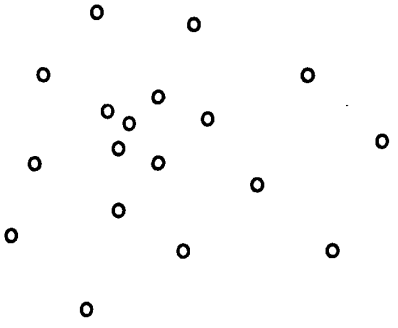


图 5.1　一个数据图

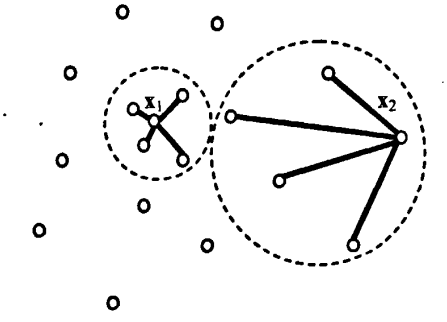


图 5.2　kNN（k=4），$\mathbf{x}_1$ 和 $\mathbf{x}_2$ 的点密度，$d(\mathbf{x}_1)>d(\mathbf{x}_2)$

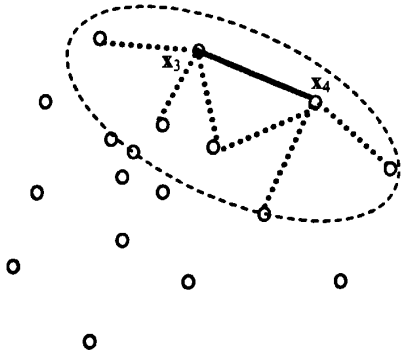


图 5.3　图的 kNN（k=4）边密度

图 5.1 是一个数据集的图形表示，图 5.2 显示了图上两个数据点 $\mathbf{x}_1$ 和 $\mathbf{x}_2$ 的 4 最近邻居节点和它们分别占据的区域。显然点密度 $d(\mathbf{x}_1)>d(\mathbf{x}_2)$。根据公式（5.1）定义的点密度，定义边$(i,j)$的 kNN 边密度为：

$$Den(i,j)=g(\|\mathbf{x}_i-\mathbf{x}_j\|_2+\frac{\beta}{k}(\sum_{l=1}^{k}\|\mathbf{x}_i-\boldsymbol{\gamma}_l^i\|_2+\sum_{l=1}^{k}\|\mathbf{x}_j-\boldsymbol{\gamma}_l^j\|_2)) \tag{5.2}$$

其中 β 是一个自由参数，控制基于密度调整的程度。kNN 边密度如图 5.3 所示。

直观上，边密度的提出是为了"拉长"处于低密度区域中的边，"缩短"处于高密度区域中的边，使得位于同一流型中的数据点间距离比它们之间的原始欧氏距离更"近"。注意公式（5.2）中的边密度只在图中的邻居节点上计算，然后图的权重矩阵 $\mathbf{W}=\{w_{ij}\}$ 被设置为 $w_{ij}=Den(i,j)$。

5.2.3 Laplacian 核方法

给定一个图 G 和相应的权重矩阵 $\mathbf{W}$，它的图形 Laplacian 矩阵 $\mathbf{L}$ 定义为：

$$\mathbf{L}=\mathbf{D}-\mathbf{W} \tag{5.3}$$

其中 $\mathbf{D}=\mathrm{diag}(d_i: i=1,\ldots,n)$，$d_i=\sum_{j=1}^{n}w_{ij}$。$\mathbf{L}$ 的归一化形式为：

$$\tilde{\mathbf{L}}=\mathbf{D}^{-1/2}\mathbf{L}\mathbf{D}^{-1/2} \tag{5.4}$$

上式称为归一化图 Laplacian 矩阵，它的伪逆 $\tilde{\mathbf{L}}^{+}$ 是一个半正定矩阵[97]，在一些半监督学习方法中被用来作为分类器的核。本文采用 Sindhwani 等提出的 Laplacian 核方法[89]，以下简要介绍这种方法。

记 $\mathcal{H}$ 为关于函数集 $X\to\mathbb{R}$ 的一个 Hilbert 空间，空间上的核函数定义为：

$$k(\mathbf{x},\mathbf{x}')=\langle k(\mathbf{x},\cdot),k(\mathbf{x},\cdot)\rangle_{\mathcal{H}} \tag{5.5}$$

给定数据点集 $X=\{\mathbf{x}_1,\ldots,\mathbf{x}_n\}$，Sindhwani 等[89]引入一个映射 $S:\mathcal{H}\to\mathbb{R}^n$，且 $Sf=(f(\mathbf{x}_1),\ldots,f(\mathbf{x}_n))$，根据数据集调整 Hilbert 空间。它在 $\mathbb{R}^n$ 空间上的半模（semi-norm）根据一个对称半正定矩阵 $\mathbf{M}$ 给出：

$$\|Sf\|_2=(Sf)^T\mathbf{M}(Sf) \tag{5.6}$$

根据一个核 $\tilde{k}(\mathbf{x},\mathbf{x}')$，可得到一个再生核 Hilbert 空间（reproducing kernel Hilbert space，RKHS）$\tilde{\mathcal{H}}$。$\tilde{k}(\mathbf{x},\mathbf{x}')$即再生核，定义为：

$$\tilde{k}(\mathbf{x},\mathbf{x}')=k(\mathbf{x},\mathbf{x}')-k_{\mathbf{x}}^{\mathrm{T}}(\mathbf{I}+\mathbf{M}\mathbf{K})^{-1}\mathbf{M}k_{\mathbf{x}'} \tag{5.7}$$

其中 $\mathbf{K}$ 是所有数据点上 $k(\mathbf{x},\mathbf{x}')$ 的核矩阵，$k_{\mathbf{x}}$ 表示向量 $(k(\mathbf{x}_1,\mathbf{x}),\ldots,k(\mathbf{x}_n,\mathbf{x}))^T$，且 $\mathbf{I}$ 是单位矩阵。令 $\mathbf{M}=\gamma\tilde{\mathbf{L}}$，记 $\tilde{\mathbf{L}}$ 是归一化图 Laplacian 矩阵，γ 是控制核变形（deformation）程度的参数。再生核矩阵记为 $\tilde{\mathbf{K}}$，它可用来作为 SVM 中的定制核矩阵。

这种 Laplacian 核方法最突出的特点是它是一种直导方法。与大部分 GBKM 不同，这种方法产生的分类器占据全部特征空间，可以对所有已知或未知样本进行

分类。而且，计算得到的 Laplacian 核矩阵与数据集相关，分类器可以由此利用图的结构信息。

本章提出的基于密度的 Laplacian 核（density based Laplacian kernels，DBLK）分类方法的步骤是：1，将数据集表示为一个图；2，根据 *k*NN 边密度计算图上各邻居节点的相似性权重；3，在此基础上构造一个 Laplacian 核分类器。

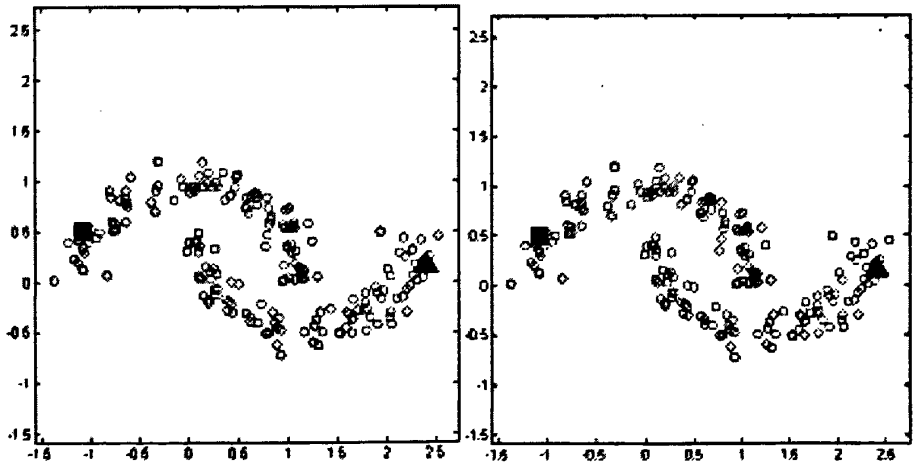
5.3 半监督分类实验与讨论

通过实验考察提出的半监督分类方法的性能。首先在人工数据集上对比 DBLK 和 TSVM 两种半监督分类方法，然后在一些真实数据集上比较 DBLK 和其他方法的分类效果。

5.3.1 Two-moons 数据集实验

Two-moons 数据集是对许多半监督学习算法性能评价都采用的一个基准数据集，它的图形分布如图 5.7(a)所示。对两个类分别取一个有标记数据点作为监督信息，其他数据点都是无标记的。

在同一数据集上分别运行 TSVM 和第二节提出的 DBLK 方法。这两种方法都是基于 SVM 的，它们区别是 TSVM 采用 RBF 核（参数设置$\sigma=0.5$，$C=10$），DBLK 采用与数据集相关的 Laplacian 核。



（a）TSVM　　（b）DBLK

图 5.4　在 two-moons 数据集上 TSVM 和 DBLK 学习得到的分类器

实验结果如图 5.4 所示。图中有标记样本由方块（正例）和三角（反例）表示，无标记样本由空心圆圈表示。分类器将图划分为阴影区域和无阴影区域，分别表

示两个类。由图中可见 DBLK 与 TSVM 相比可以得到更准确的分类器。当有标记样本点很少时，TSVM 在很多时候难以发现数据集的内在流型特征。而在 DBLK 方法中，基于密度的 Laplacian 核可以提供关于图结构的信息，使 SVM 可以在有标记样本很少的情况下取得很好的分类性能。

5.3.2 真实数据集实验

实验中使用的真实数据集包括 COIL20、USPS-TEST、UCI-ADULT 和 UCI-MUSHROOM。其中 COIL20 和 USPS-TEST 是多类图像数据集，另外两个都是 UCI 机器学习标准集中的 2 类数据集。实验中的各数据集是这四个数据集经随机选取得到的子集，它们的详细信息如表 5.1 所示。

表 5.1 真实数据集属性

数据集	类数	特征数	样本数	有标记样本数
COIL20	20	1,024	1,440	40
USPS-TEST	10	256	2,007	50
UCI-ADULT	2	6	4,000	40
UCI-MUSHROOM	2	22	4,000	40

实验中用于比较的算法包括 SVM、TSVM、Laplacian SVM（LapSVM）和本文提出的 DBLK 方法。DBLK 和 LapSVM 的主要区别是 DBLK 中图权重矩阵的建立是采用基于密度的 *k*NN 距离权重，而 LapSVM 根据节点间的欧氏距离设置边权重。所有四种方法都是基于 SVM 的二值分类方法。对于多类学习问题，可采用一对多（one against all）协议将其转化为二值分类问题。

对每个数据集，进行 5 等分交叉检验：数据被分为 5 等份，其中一份作为训练集，其他作为测试集。

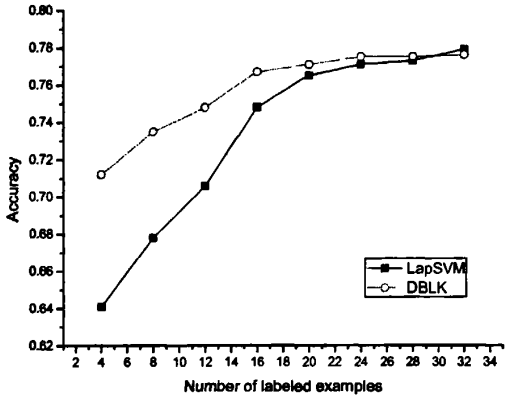
表 5.2 无标记数据集上的平均分类精度

数据集	SVM	TSVM	LapSVM	DBLK
COIL20	0.642	0.703	0.924	**0.949**
USPS-TEST	0.685	0.716	0.820	**0.861**
UCI-ADULT	0.522	0.525	0.714	**0.767**
UCI-MUSH ROOM	0.537	0.579	0.803	**0.880**

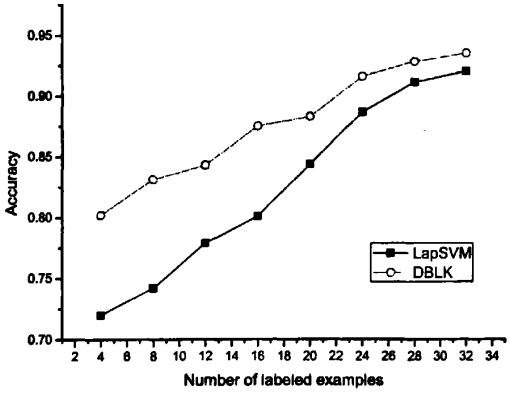
分类结果如图 5.2。基于 Laplacian 核的两种方法 LapSVM 和 DBLK 与 SVM 和 TSVM 相比达到了更好的分类精度。原因是 LapSVM 和 DBLK 中采用的数据相关

的核与 SVM 和 TSVM 的标准 RBF 核相比可以产生更准确的分类器。表中显示 DBLK 的性能稍微优于 LapSVM。

为进一步考察 LapSVM 和 DBLK 方法的区别，在 UCI-ADULT 和 UCI-MUSHROOM 数据集上进行分类实验。在每个数据集中随机选取 1,000 个样本，有标记数据设置为 4 到 32 个。实验结果如图 5.5 所示。



（a）UCI-ADULT



（b）UCI-MUSHROOM

图 5.5 两个 UCI 数据集上 LapSVM 和 DBLK 的分类结果

由图 5.5 可见，在有标记样本足够多时，两种方法的预测性能非常相近。然而，当有标记样本数较少时，DBLK 的分类性能明显强于 LapSVM。正如前面提到的，两种分类方法的最主要区别是，本文提出的 DBLK 方法使用了一种新型基于密度的距离来建立图的权重矩阵，而 LapSVM 没有考虑数据在图上的的密度分布特点。因此，可以认为这是有标记样本数量较少时 DBLK 可以取得较好性能的主要原因。

5.4 半监督聚类中基于密度的约束扩展

5.4.1 半监督聚类问题的图形定义

一般形式的半监督聚类问题中，已知的数据集包含 n 个样本点 $\mathbf{x}_i, 1 \le i \le n$。由数据集可建立图 $G=(V,E)$，图中的节点集是样本点 $V=\{\mathbf{x}_i\}_{i=1}^n$，边集为 $E \subset V \times V$。根据欧氏距离 $d(i,j)=\|\mathbf{x}_i-\mathbf{x}_j\|_2$，对任意两点间的边 $(i,j) \in E, i \ne j$，定义其权重为 $w_{i,j}=\exp(-d(i,j)^2/(2\sigma^2))$，其中参数 σ 是用于 Parzen 窗口密度估计的 Gaussian 分布宽度，且 $w_{i,i}=0$。

5.4.2 基于路径的距离

半监督学习和无监督学习中常依据“聚类假设”，即假定同一类中或同属于同一高密度区域中的两个样本点间应存在较小的距离。根据“聚类假设”，在由数据集生成的图中，将样本点间的相似性定义为某种基于密度的距离，与欧氏距离相比能够更好地反映数据集的结构，作为分类和聚类的指导。以图 5.6 中的数据图形为例，在基于密度的距离度量下，样本点 $\mathbf{x}_1$ 和 $\mathbf{x}_2$ 的距离小于 $\mathbf{x}_2$ 和 $\mathbf{x}_3$ 的距离。文献[90,95]等提出了各种基于密度的距离度量，用于半监督和无监督学习。

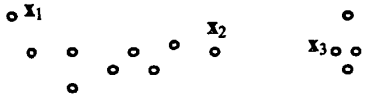


图 5.6 基于密度的距离度量下， $\mathbf{x}_1$ 和 $\mathbf{x}_2$ 的距离小于 $\mathbf{x}_2$ 和 $\mathbf{x}_3$ 的距离

采用 Fischer 等[90]的“连接核”方法，定义一种密度敏感的基于路径的距离。令 p 为长度为 $|p|$ 的一条路径，路径上各边表示为 $(p_l, p_{l+1}), 1 \le k < |p|$。定义路径 p 两端点间基于路径 p 的相似度为路径上各边权重的最小值：

$$w^p_{p_1,p_{|p|}} = \min_{l<|p|} \left\{ \exp(-d(p_l,p_{l+1})^2/(2\sigma^2)) \right\} \tag{5.8}$$

假设 $P_{i,j}$ 为连接点 i 和 j 所有路径的集合，定义边(i,j)基于路径的距离权重为连接点 i 和 j 的所有路径的基于路径的相似度的最大值：

$$w_{i,j}=\max_{p\in P_{i,j}}\left(w_{i,j}^{p}\right)=\max_{p\in P_{i,j}}\min_{l<|p|}\left\{\exp(-d(p_l,p_{l+1})^2/(2\sigma^2))\right\} \tag{5.9}$$

对样本相似度这种形式的定义不依赖于路径长度，因此可能会对不同聚类中间的噪音点敏感，造成错误聚类。Chapelle 等[95]提出了一种"低密度分离"方法以解决此问题，将公式（5.9）放松为：

$$w_{i,j}\approx\exp\left[-\left(\min_{p\in P_{i,j}}\operatorname{smax}^{\rho}(p)\right)^2/(2\sigma^2)\right] \tag{5.10}$$

$$\operatorname{smax}^{\rho}(p)=\frac{1}{\rho}\ln\left(1+\sum_{l=1}^{|p|-1}(\exp(\rho d(p_l,p_{l+1}))-1)\right) \tag{5.11}$$

其中参数 ρ 的作用是将公式（5.9）定义的基于路径的距离以路径长度进行调整，这样可以消除噪音点的影响。ρ 可以取 $(0,+\infty)$ 的任何值，在后面实验中设置 $\rho=2$。

5.4.3 基于密度的约束扩展

半监督聚类的监督信息通常由样本点间的约束关系表示。约束关系包括 must-link 和 cannot-link，表示两个样本点属于或不属于同一类。一般地，已知的约束条件越多，可以越清晰地反映数据集的分布信息，半监督聚类算法的聚类效果越好。

前面定义了样本间基于密度的相似性，在此基础上本文提出一种对现有约束进行扩展的方法，称为基于密度的约束扩展（density-based constraints expansion，DCE）。约束扩展的提出主要基于这样的假设：关系是 must-link 的样本点处于同一高密度区域中，关系是 cannot-link 的样本点处于不同的高密度区域。约束扩展的方法是：根据已知的某约束中两个样本点的关系，计算它们与其他样本点的基于密度的图形相似度，寻找最相似的点，即基于密度的距离最接近的点，指定这些点之间的 must-link 和 cannot-link 关系，添加新的约束条件，扩展已有的约束集。扩展后的约束集将包括更多的约束条件，而新增的约束由基于密度的距离计算得到，可以将数据集空间分布信息引入聚类方法，可望获得更好的聚类效果。

下面给出算法描述：

算法 5.1　基于密度的约束扩展算法（DCE）

输入：数据集 $\mathbf{X}=\{\mathbf{x}_i\}_{i=1}^{n}$，must-link 约束集 $\mathbf{M}=\{(\mathbf{x}_i,\mathbf{x}_j)\}$，cannot-link 约束集 $\mathbf{C}=\{(\mathbf{x}_i,\mathbf{x}_j)\}$

输出：扩展的 must-link 约束集 $\mathbf{M}^+$，扩展的 cannot-link 约束集 $\mathbf{C}^+$

步骤：1，根据样本点构造图形 G，根据公式（5.10）、（5.11），计算样本点间基于密度的图形相似度矩阵 $\mathbf{W}=\{w_{i,j}\}$；

2，初始化 $\mathbf{M}^+=\mathbf{M}$，$\mathbf{C}^+=\mathbf{C}$；

3，对 $\mathbf{M}^+$和 $\mathbf{C}^+$中的所有点进行约束传递：对任一点 $\mathbf{x}$，如果$(\mathbf{x},\mathbf{x}_i)\in\mathbf{M}^+$且$(\mathbf{x}_i,\mathbf{x}_j)\in\mathbf{M}^+$，则扩展 $\mathbf{M}^+$为$\mathbf{M}^+\cup\{(\mathbf{x},\mathbf{x}_j)\}$；如果$(\mathbf{x},\mathbf{x}_i)\in\mathbf{M}^+$且$(\mathbf{x}_i,\mathbf{x}_j)\in\mathbf{C}^+$，则扩展 $\mathbf{C}^+$为$\mathbf{C}^+\cup\{(\mathbf{x},\mathbf{x}_j)\}$。记所有与 $\mathbf{x}_i$关系为 must-link 的点集为$\mathbf{M}_i^+$，所有与 $\mathbf{x}_i$关系为 cannot-link 的点集为$\mathbf{C}_i^+$；

4，对 $\mathbf{M}^+$中的任一约束$(\mathbf{x}_i,\mathbf{x}_j)$，计算 $\mathbf{x}_i$和 $\mathbf{x}_j$的 k 最近邻居节点集，分别记为$\mathbf{X}^i=\{\mathbf{x}_h^i\}$和$\mathbf{X}^j=\{\mathbf{x}_h^j\}$，$1\le h\le k$；

5，对$\mathbf{X}^i$中的任一点$\mathbf{x}_h^i$，假设它在数据集 $\mathbf{X}$ 中的下标为 d，即$\mathbf{x}_h^i=\mathbf{x}_d$，如果$w_{d,j}\le w_{i,j}$，且$\mathbf{M}_d^+\cap\mathbf{C}_j^+=\phi$，$\mathbf{C}_d^+\cap\mathbf{M}_j^+=\phi$，则扩展 $\mathbf{M}^+$为$\mathbf{M}^+\cup\{(\mathbf{x}_d,\mathbf{x}_j)\}$。用同样的方法扩展$\mathbf{X}^j$中的满足同样条件的点；

6，重复步骤 3，4，5 直至 $\mathbf{M}^+$和 $\mathbf{C}^+$规模不再增加。返回 $\mathbf{M}^+$和 $\mathbf{C}^+$。

算法结束

注意算法每次只扩展样本点间的 must-link 关系，cannot-link 关系由算法步骤 3 的约束传递自动扩展得到。

经扩展的约束集可以用于现有的各种半监督聚类算法。下面以 Klein 等的约束完全连接（CCL，constraint complete link）[44]和 Basu 等的成对约束 K 均值（PCKMEANS，pairwise constraint kmeans）方法[45]为例介绍约束扩展方法的应用。

在 CCL 算法基础上增加基于密度的约束扩展，称为 DCE-CCL。由于约束扩展算法中已包括 must-link 和 cannot-link 约束在整个数据集上的传递关系，因此去掉 CCL 原有的约束传递部分。DCE-CCL 算法描述为：

算法 5.2　基于密度的约束扩展用于约束完全连接聚类（DCE-CCL）

输入：数据集 $\mathbf{X}=\{\mathbf{x}_i\}$，must-link 约束集 $\mathbf{M}=\{(\mathbf{x}_i,\mathbf{x}_j)\}$，cannot-link 约束集 $\mathbf{C}=\{(\mathbf{x}_i,\mathbf{x}_j)\}$

输出：聚类 *Clusters*

步骤：1，运行 DCE 算法，得到扩展的约束集 $\mathbf{M}^+$和 $\mathbf{C}^+$，以及基于密度的图形相似度矩阵 $\mathbf{W}=\{w_{i,j}\}$；

2，对每个样本点分配一个类标记：$Clusters=\{c_i\}_{i=1}^n$；

3，初始化连接集 *Linkage* 为空集，初始化 $\delta(c_i,c_j)=\dfrac{1}{w_{i,j}}$；

4，当 $|Clusters|>1$，运行 5，6；

5，选择最近的聚类 $(c_1,c_2)=\underset{c_1,c_2\in Clusters}{\arg\min}\ \delta(c_1,c_2)$，将 (c_1,c_2) 加入连接集 *Linkage*，将 *Clusters* 中的 c_1 和 c_2 融合为 c_{new}；

6，对 $c_i\in Clusters$，重新计算聚类间距离 $\delta(c_i,c_{\text{new}})=\max\{\delta(c_i,c_1),\delta(c_i,c_2)\}$。

算法结束

DCE 算法扩展的约束集有可能存在误差，即将本属于同一类的样本点约束关系设置为 cannot-link，或将不属于同一类的样本点间约束设置为 must-link。由于 CCL 聚类方法的基本思想是将已有的约束关系在整个样本空间上进行传递，错误的约束关系可能会随着约束传递而扩散，影响聚类的效果，这个问题可采用容错性更好的半监督聚类算法解决。在 PCKMEANS 算法基础上增加基于密度的约束扩展，称为 DCE-PCKMEANS。算法描述为：

算法 5.3　基于密度的约束扩展用于成对约束 K 均值聚类（DCE-PCKMEANS）

输入：数据集 $\mathbf{X}=\{\mathbf{x}_i\}$，must-link 约束集 $\mathbf{M}=\{(\mathbf{x}_i,\mathbf{x}_j)\}$，cannot-link 约束集 $\mathbf{C}=\{(\mathbf{x}_i,\mathbf{x}_j)\}$，聚类数 k，约束的违反权重 w

输出：聚类 $Clusters=\{\mathbf{X}_h\}_{h=1}^k$

步骤：1，运行 DCE 算法，得到扩展的约束集 $\mathbf{M}^+$和 $\mathbf{C}^+$，以及基于密度的图形相似度矩阵 $\mathbf{W}=\{w_{i,j}\}$；

2，初始化各聚类中心点：根据 $\mathbf{M}^+$和 $\mathbf{C}^+$创建 λ 个邻居集 $\{N_p\}_{p=1}^{\lambda}$。

如果 $\lambda\geq k$，取 $\{N_p\}$ 中最大的 k 个邻居集的中心作为初始聚类中心 $\{\boldsymbol{\mu}_h^{(0)}\}_{h=1}^k$；

否则$\lambda < k$，初始化$\{\boldsymbol{\mu}_h^{(0)}\}_{h=1}^{\lambda}$为$\{N_p\}_{p=1}^{\lambda}$的中心，

如果存在与所有邻居集关系是 cannot-link 的点 $\mathbf{x}$，初始化$\boldsymbol{\mu}_{\lambda+1}^{(0)}$为 $\mathbf{x}$，随机生成其余聚类中心；

3，迭代 4，5，6 步直至收敛；

4，将各样本点指定类标记 h^*，使得

$$h^* = \arg\min_h \left\{ \frac{1}{2} \left\| \mathbf{x} - \boldsymbol{\mu}_h^{(t)} \right\|^2 + w \sum_{(\mathbf{x}_i, \mathbf{x}_j) \in \mathbf{M}^+} [h \neq l_j] + w \sum_{(\mathbf{x}_i, \mathbf{x}_j) \in \mathbf{C}^+} [h \neq l_j] \right\}$$

其中 l_j 中是样本点 j 的当前类标记，$[\cdot]$是指示函数，$[true]=1$，$[false]=0$；

5，计算聚类中心点：$\{\boldsymbol{\mu}_h^{(t+1)}\}_{h=1}^{k} \leftarrow \{\frac{1}{\left|\mathbf{X}_h^{(t+1)}\right|} \sum_{\mathbf{x} \in \mathbf{X}_h^{(t+1)}} \mathbf{x}\}_{h=1}^{k}$

6，$t \leftarrow t+1$

算法结束

DCE-PCKMEANS 算法与 DCE-CCL 的最大区别是它允许算法以一定的权重 w 违反约束，设置合适的违约权重可以在一定程度上消除约束扩展的误差，因此与 DCE-CCL 相比有更好的容错性。

5.5 半监督聚类实验与讨论

通过实验比较第三节提出的两种算法在多个数据集上的聚类效果。以下的实验结果都是在数据集上重复 100 次的平均值，已知约束随机生成，且 must-link 和 cannot-link 约束等比例产生。在 DCE-PCKMEANS 算法中，根据经验将约束违反权重设置为数据集图形样本点最小距离的 1/10：$w = \frac{1}{10} \min_{(i,j) \in E} d(i,j)$。

5.5.1 评价指标

实验中所有数据集的分类结果已知，采用 Wagstaff 等[92]提出的 CRI 指标：

$$CRI = \frac{(\text{正确成对决策数} - \text{已知约束数})}{(\text{全部成对决策数} - \text{已知约束数})} \tag{5.12}$$

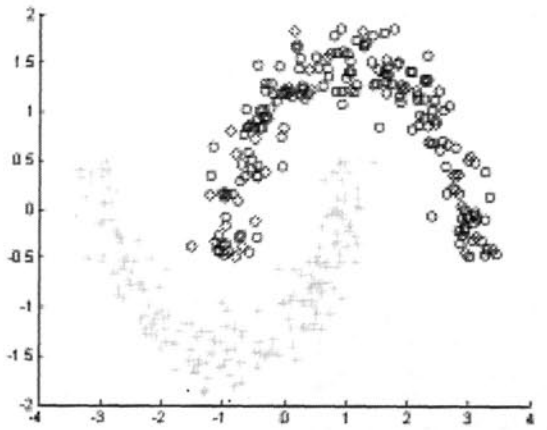
其中成对决策是算法对所有样本两两之间是否属于同一类判断结果，全部成对决策数等于 $n(n-1)/2$。注意评价指标中不计已知约束数，因为作为监督信息的约束不能反映聚类算法的效果。

此外，为测量约束扩展的实际效果，定义 DCE 扩展得到的约束准确率：

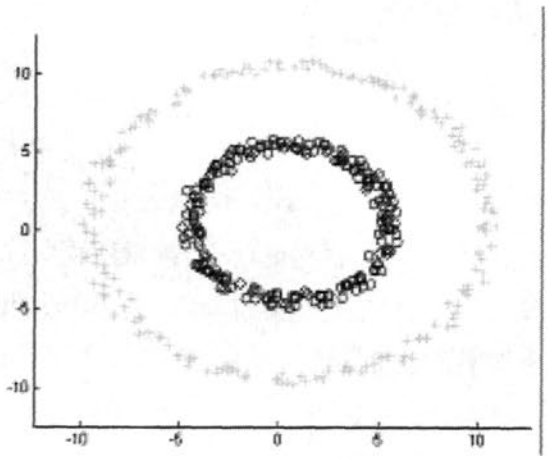
$$CP=\frac{\text{正确扩展约束数}}{\text{全部扩展约束数}} \tag{5.13}$$

5.5.2 人工数据集

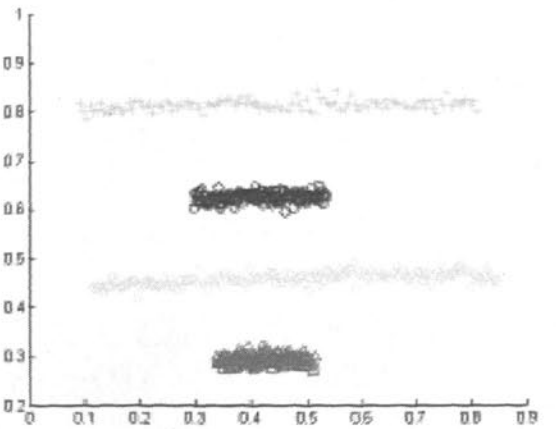
实验采用半监督学习问题中常见的三个人工数据集 two-moons， two-circles 和 four-lines，它们都符合“聚类假设”，分布如图 5.7 所示。



（a）two-moons



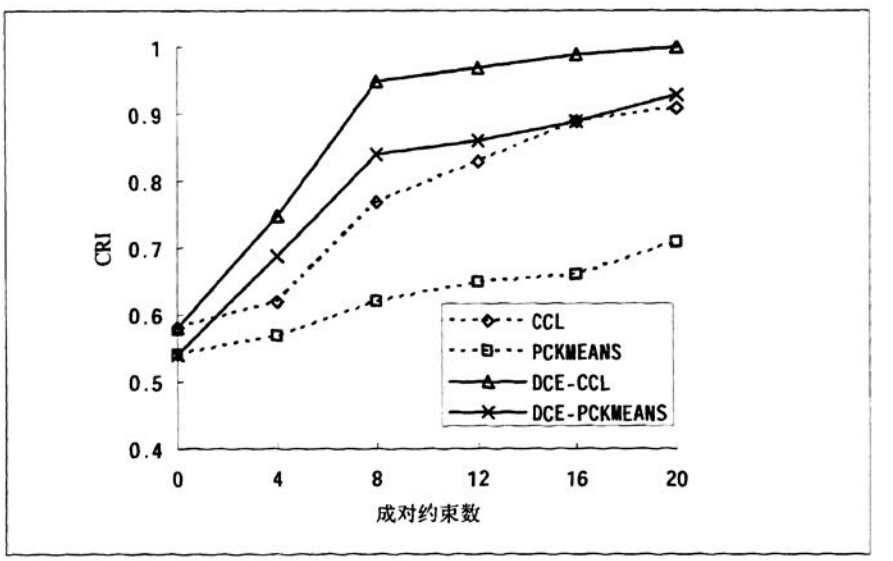
（b）two-circles



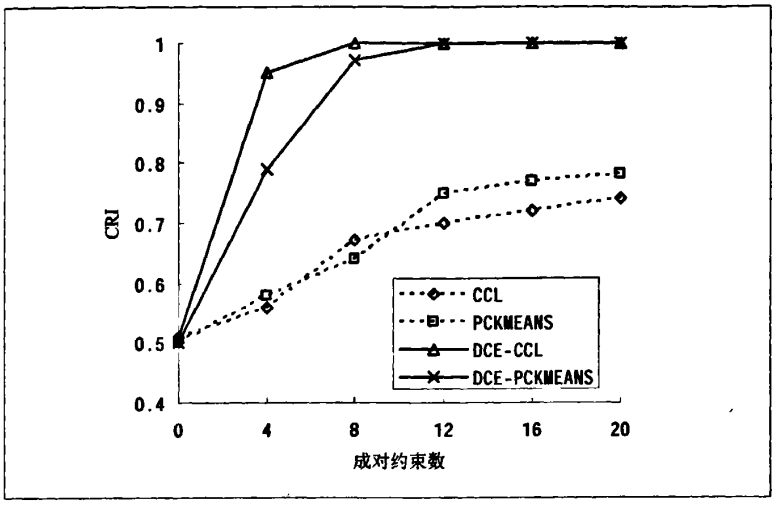
（c）four-lines

图 5.7　人工数据集分布图：two-moons，two-circles 和 four-lines

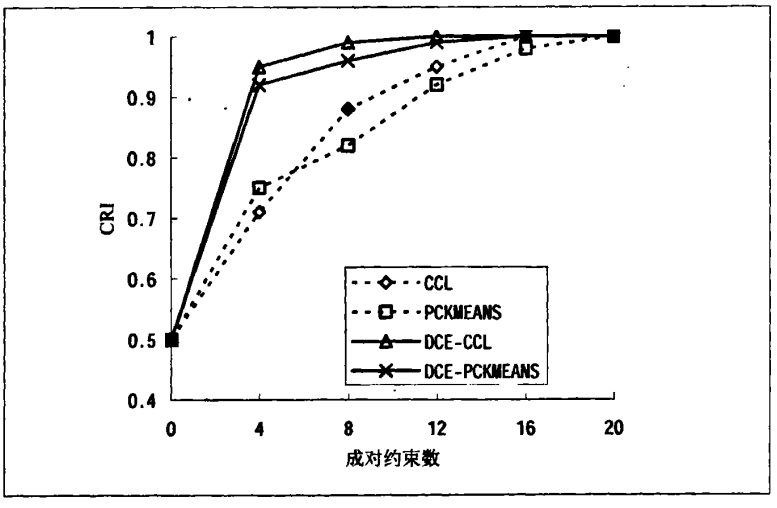
图 5.8 是 4 种算法在 3 种人工数据集上的实验结果。由图中可见，应用约束扩展的半监督聚类算法 DCE-CCL 和 DCE-PCKMEANS 与未经约束扩展的相应算法相比，性能有明显提高。特别地，在初始约束条件较少时，约束扩展的半监督聚类算法可迅速得到较好的聚类准确率，原因是扩展的约束集为聚类算法提供了更多可利用的监督信息。在初始约束数量达到一定水平后，约束扩展算法的聚类性能与未经约束扩展的算法相近，这是因为此时约束所提供的监督信息已达到聚类算法所需的程度，增加更多的约束不能继续提升算法的聚类性能。在人工数据集上，DCE-CCL 算法的聚类性能略优于 DCE-PCKMEANS，原因是两种算法的聚类机制不同，前者基于层次聚类，后者基于 KMEANS 算法。



（a）two-moons



（b）two-circles



（c）four-lines

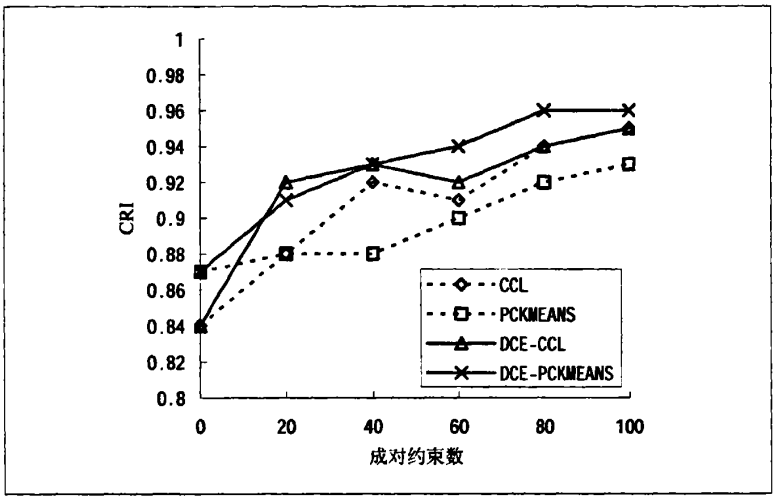
图 5.8　算法在各数据集上的聚类性能：two-moons，two-circles 和 four-lines

5.5.3 真实数据集

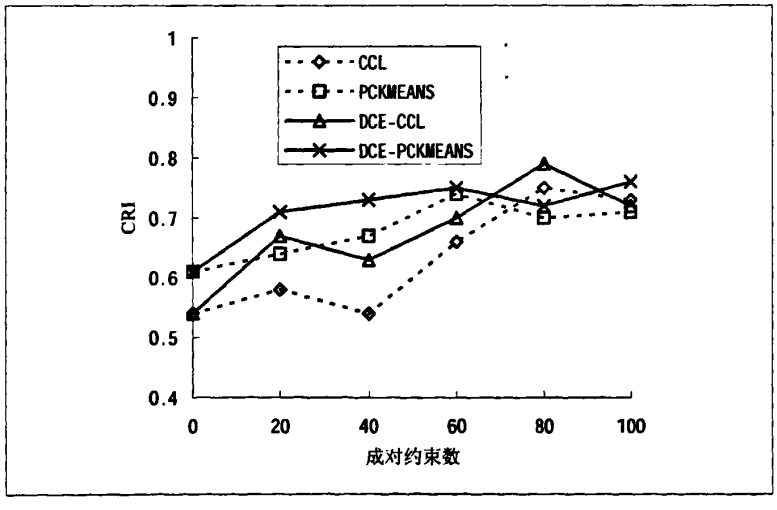
在真实数据集上进行实验，选取 UCI 机器学习标准数据集中的 iris，ionosphere 和 soybean。数据集的描述见表 5.3。

表 5.3 真实数据集描述

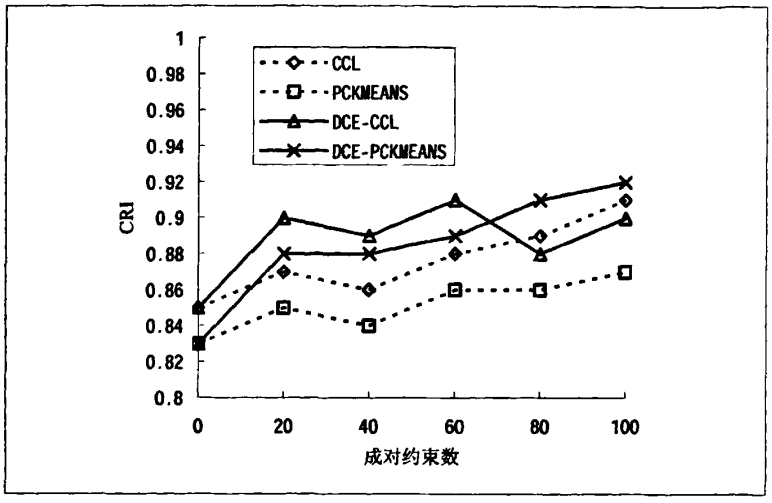
数据集	样本数	数据维度	类数
iris	150	4	3
ionosphere	351	34	2
soybean	307	35	19



（a）iris



（b）ionosphere



(c) soybean

图 5.9　算法在各数据集上的聚类性能：iris，ionosphere 和 soybean

约束扩展的聚类算法在真实数据集上的实验结果如图 5.9 所示，与人工数据集上的表现相似，同样可以在较少的约束条件下达到较好的聚类准确率。在真实数据集上，DCE-PCKMEANS 算法的整体性能略优于 DCE-CCL，原因可能是如前面提出的，前者比后者"容错性"更强。

在 3 种真实数据集上，DCE 算法的约束扩展效果如表 5.4 所示。在不同的数据集和初始约束数情形下，扩展得到的约束准确率均达到了较高水平。

表 5.4　真实数据集上 DCE 算法的约束扩展效果

初始约束数	iris		ionosphere		soybean	
	扩展后约束总数	约束扩展准确率	扩展后约束总数	约束扩展准确率	扩展后约束总数	约束扩展准确率
20	76.4	0.97	59.7	0.89	68.7	0.92
40	135.6	0.96	86.2	0.86	122.4	0.92
60	269.3	0.96	154.0	0.91	200.5	0.94
80	305.5	0.95	218.3	0.85	289.0	0.89
100	334.4	0.97	284.6	0.92	375.1	0.94

5.6 本章小结

本章研究了半监督分类和半监督聚类方法。主要的研究思路是在图方法框架

下，采用基于密度的距离形式表示节点间的相似性关系。这样做的目的是挖掘数据集内在的流型特征，充分利用大量无标记数据中所含的数据集结构信息，以获得更好的学习效果。

对半监督分类问题，提出一种基于密度的 Laplacian 核方法 DBLK，根据文中定义的一种 *k*NN 密度为图中的边设定权重，然后采用一种与数据集相关的 Laplacian 核构造学习器。学习得到的分类器可以在整个特征空间上工作，突破了大多数基于图的半监督学习方法的转导性限制。

对半监督聚类问题，提出一种基于密度的约束扩展方法 DCE。将数据集以图的形式表达，定义了一种基于密度的图形相似度。根据样本点间的距离和相似度关系，对已知约束集进行扩展，扩展后的约束集可用于各种半监督聚类算法。并以约束完全连接聚类和成对约束 K 均值方法为例，说明了约束扩展方法的应用。

第六章 总结与展望

随着计算机和网络技术的迅速发展，基于网络的各种应用和服务的不断涌现，Internet 已成为人们获取各类信息的重要来源和互相沟通交流的重要工具。网上浩如烟海、良莠不齐的各类信息资源，这给人们对寻找和利用有价值的信息带来了重重障碍，为解决这一问题，各种信息过滤和信息检索系统和工具应运而生。机器学习方法在各种信息过滤和检索中起着重要作用，是众多信息系统中的核心技术。本文的主要研究内容是与信息过滤和信息检索问题相关的机器学习方法，重点是针对聚类、分类等机器学习问题，提出适当的模型和算法予以分析解决，并通过实证的方式加以检验。

6.1 论文的主要工作和创新性

论文首先介绍了信息过滤和信息检索的概念和发展情况，概括说明了协同过滤、机器学习中的聚类和分类等研究主题。然后介绍了研究工作的理论基础，主要包括有限混合模型和 EM 算法、基于图的机器学习方法以及支持向量机等基于统计的机器学习方法。在此基础上，论文在以下几个方面进行研究，并取得了具有一定创新性的成果：

- 分析了协同过滤问题中传统方法的不足之处，提出了一种基于概率模型的协同过滤方法，称为真实偏好高斯混合模型。在这种模型中，用户－项目的最终评价由三个因素决定：用户对项目主题和内容的真实偏好，用户的评分习惯，以及项目的公众评价。通过实验证实了新模型的预测性能明显优于几个传统协同过滤算法。
- 针对文本聚类问题，提出了用有限混合模型进行无监督文本聚类的一种广义方法。它将模型选择，特征选择以及混合模型的参数估计纳入一个统一的框架。在提出的框架下，发展了一种带特征选择的多项式混合模型（MM-FS），并以 BIC 准则作为模型选择方法，作为广义方法的实例做了详细的说明。在 4 个大规模文本数据集上的实验结果表明这一方法在模型选择，特征选择和聚类结果三个方面都取得了很好的效果。
- 提出了一种半监督分类方法，称为基于密度的 Laplacian 核方法。定义了一种基于密度的 *k*NN 距离来反映数据点间的距离，并以一种 Laplacian 核方法来构造整个数据集上的超分类面。这种方法与其他基于 SVM 的半监督分类方法相比，突出的特点是在有标记样本数量极少时也可以达到很好的分类精度。
- 对半监督聚类问题，现有方法很少利用数据集空间结构信息，有限的约束条

件限制了聚类算法的性能。对此，提出了一种基于密度的约束扩展方法。这一方法根据样本点间基于密度的距离和相似度关系，对已知约束集进行扩展，扩展后的约束集包含了数据集内在的结构信息。以两种常见的半监督聚类算法为例，通过实验说明了约束扩展方法的实际效果。

6.2 今后研究工作展望

今后的研究工作将包括对机器学习方法本身的进一步研究以及将机器学习方法应用于信息过滤和信息检索中的具体问题的研究两部分，即理论研究和应用研究。具体地，研究工作可以集中在以下方向：

- 对基于模型的协同过滤方法的进一步研究。理论研究包括降低模型预测的在线计算量，将基于内容的先验信息加入到模型中。应用研究包括将本体论概念引入模型，设计分布式、并行式协同过滤系统，以及在电子商务等其他领域的应用等。
- 对无监督文本聚类广义方法的进一步研究。理论研究包括：第一，将模型选择与混合模型的参数估计过程相结合以减少模型选择所需的计算量。第二，将聚类问题划分为多个弱学习器，采用 Boosting 等方法将多个弱学习器的能力综合起来。应用研究方面，可以考察广义框架中其他形式有限混合模型的聚类效果。
- 对半监督学习问题，已提出的基于密度的聚类和分类方法在小规模数据集上表现出很好的学习性能。但对实际问题中常见的大规模数据集特别是文本数据集，还有很多具体问题需要解决。主要的理论和应用研究包括：第一，在学习器中引入特征选择方法降低问题的复杂性。第二，考虑采用分治法等方法降低数据集的规模。第三，设计可扩展性更好的基于密度的半监督学习方法。

参考文献

[1] 中国互联网络信息中心,《第 18 次中国互联网络发展状况统计报告》, 2006

[2] Belkin, N. J., Croft, B. B., Information filtering and information retrieval: two sides of the same coin?, Communications of the ACM, 1992, 35: 29-38

[3] Luhn, H. P., A business intelligence system, IBM Journal of Research and Development, 1958, 2(4): 314-319

[4] Denning, P. J., Electronic junk, Communications of the ACM, 1982, 25(3): 163-165

[5] Malone, T. W., Grant, K. R., Turbak, F. A., et al, Intelligent information sharing systems, Communications of the ACM, 1987, 30(5): 390-402

[6] Yan, T., Garcia-Molina, H., SIFT – A tool for wide-area information dissemination, In: Proceedings.1995 USENIX Technical Conference, 1995, 177-186

[7] Shardanand, U., Maes, P., Social information filtering: algorithms for automating "word of mouth", In: Proceedings of ACM CHI'95 Conference on Human Factors in Computing Systems, 1995, 1: 210-217

[8] Schapire, R. E., Singer, Y., Singhal, A., Boosting and Rocchio applied to text filtering, In: Proceedings of SIGIR-98, 21st ACM International Conference on Research and Development in Information Retrieval, 1998, 215-223

[9] Resnick, P., Iacovou, N., Suchak, M., et al, GroupLens: An open architecture for collaborative filtering of netnews, In: Proceedings of ACM 1994 Confernece on Computer Supported Cooperative Work, 1994, 175-186

[10] Lieberman, H., Van Dyke, N. W., Vivacqua, A. S., Let's browse: A collaborative web browsing agent, In: Proceedings of the 1999 International Conference on Intelligent User Interfaces (IUI'99), 1999, 65-68

[11] Chan, P., A non-invasive learning approach to building web user profiles, Workshop on web usage analysis and user profiling, Fifth International Conference on Knowledge Discovery and Data Mining, 1999

[12] Bollacker, K., Lawrence, S., Giles, C. L., Citeseer: An autonomous web agent for automatic retrieval and identification of interesting publications, In: Proceedings of the 2nd International Conference on Autonomous Agents, 1998, 116-123

[13] Mooers, C. N., Zatocoding applied to mechanical organization of knowledge.

American Documentation, 1951, 2, 20-32

[14] Salton, G., The SMART retrieval system - Experiments in automatic document processing, Prentice-Hall, 1971

[15] Sparck-Jones, K., A statistical interpretation of term specificity and its application in retrieval, Journal of Documentation, 1972, 28(1):11-21

[16] Goldberg, D., Nichols, D., Oki, B. M., et al, Using collaborative filtering to weave an information tapestry, Communications of the ACM, 1992, 35(12): 61～70

[17] Sarwar, B., Karypis, G., Konstan, J., et al, Item-based collaborative filtering recommendation algorithm, In: Proceedings of the 10th International World Wide Web Conference, 2001, 285-295

[18] Breese, J. S., Heckerman, D., Kardie, C., Empirical analysis of predictive algorithms for collaborative filtering, In: Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence, 1998, 43-52

[19] Pennock, D. M., Horvitz, E., Lawrence, S., et al, Collaborative filtering by personality diagnosis: A hybrid memory- and model-based approach, In: Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence, 2000, 473-480

[20] Hofmann, T., Puzicha, J., Latent class models for collaborative filtering, In: Proceedings of the International Joint Conference in Artificial Intelligence, 1999, 688-693

[21] Bellman, R., Adaptive control processes: A guided tour, Princeton University Press, 1961

[22] Kohavi, R., John, G. H., Wrappers for feature subset selection, Artificial Intelligence, 1997, 97(1-2):273-324

[23] Pearson, K., On lines and planes of closest fit to systems of points in space, Philosophical Magazine, 1901, 2: 559-572

[24] Fisher, R. A., The use of multiple measurements in taxonomic problems, Annals of Eugenics, 1936, 7, 179-188

[25] Herault, J., Jutten, C., Space or time adaptive signal processing by neural network models, In: Neural Networks for Computing, Proceedings of AIP Conference, 1986, 206-211

[26] Sneath, P. H. A., Sokal, R. R., Numerical Taxonomy, Freeman, 1973

[27] King, B., Step-wise clustering procedures, Journal of the American Statistical Association, 1967, 69: 86–101

[28] McQueen, J., Some methods for classification and analysis of multivariate

observations, In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1967, 281–297

[29] Agrawal, R., Gehrke, J., Gunopulos, D., et al, Automatic subspace clustering of high dimensional data for data mining applications, In: Proceedings of the 1998 ACMSIGMOD International Conference on the Management of Data, 1998, 94–105

[30] Hinneburg, A., Keim, D., An efficient approach to clustering in large multimedia databases with noise, In: Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining, 1998

[31] Ester, M., Kriegel, H. P., Sander, J., et al, A density-based algorithm for discovering clusters in large spatial databases with noise, In: Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, 1996

[32] Figueiredo, M., Jain, A. K., Unsupervised learning of finite mixture models, IEEE Trans on Pattern Analysis and Machine Intelligence, 2002, 24: 381-396

[33] Quinlan, J.R., Induction of decision trees, Machine Learning, 1986, (1), 81-106

[34] Freund, Y., Schapire, R.E., Experiments with a new boosting algorithm, In: Proceedings of the Thirteenth International Conference on Machine Learning, 1996, 148-156

[35] Minsky, M., Papert, S., Perceptrons, MIT Press, Cambridge, MA, 1969, 1988

[36] Vapnik, V. N., Support vector method for function approximation, regression estimation and signal processing, Neural Information Processing Systems, 1996, 9: 281-287

[37] Dempster, A. P., Laird, N. M., Rubin, D. B., Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society, 1977, 39: 1-38

[38] Nigam, K., McCallum, A. K., Thrun, S., et al, Text classification from labeled and unlabeled documents using EM, Machine Learning, 2000, 39(2/3): 103–134

[39] Yarowsky, D., Unsupervised word sense disambiguation rivaling supervised methods, In: Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics, 1995, 189–196

[40] Blum, A., Mitchell, T., Combining labeled and unlabeled data with co-training. In: Annual Conference on Computational Learning Theory, COLT'98, 1998, 92-100

[41] Joachims, T., Transductive inference for text classification using support vector machines. In: Proceedings of the 16th International Conference on Machine

Learning, 1999, 200–209

[42] Blum, A., Chawla, S., Learning from labeled and unlabeled data using graph mincuts. In: Proceedings of the 18th International Conference on Machine Learning, 2001, 19-26

[43] Cohn, D., Caruana, R., McCallum, A., Semi-supervised clustering with user feedback, Technical Report TR2003-1892, Cornell University, 2003

[44] Klein, D., Kamvar, S. D., Manning, C., From instance-level constraints to space-level constraints: Making the most of prior knowledge in data clustering, In: Proceedings of the 19th International Conference on Machine Learning, 2002, 307-314

[45] Basu, S., Banerjee, A., Mooney, R., Active semi-supervision for pairwise constrained clustering, In: Proceedings of the 2004 SIAM International Conference on Data Mining, 2004, 333-344

[46] Demiriz, A., Bennett, K., Embrechts, M., Semi-supervised clustering using genetic algorithms. In: Intelligent Engineering Systems Through Artificial Neural Networks, 1999, 9: 809–814

[47] Jain, A. K., Dubin, R., Mao, J., Statistical pattern recognition: A review, IEEE Trans on Pattern Analysis and Machine Intelligence, 2000, 22(1): 4-38

[48] Jain, A. K., Dubes Algorithms for clustering data, Englewood Cliffs, New Jersey, Prentice Hall, 1988

[49] McLachlan, G., Basford, K., Mixture models: Inference and application to clustering , New York, Marcel Dekker, 1988

[50] McLachlan, G., Peel, D., Finite mixture models, New York, John Wiley & Sons, 2000

[51] Xu, L., Jordan, M., On convergence properties of the EM algorithm for Gaussian mixtures, Neural Computation, 1996, 8: 129-151

[52] Chrétien, S., Hero III, A., Kullback proximal algorithms for maximum likelihood estimation, IEEE Trans on Information Theory, 2000, 46: 1800-1810

[53] Campbell, J., Fraley, C., Murtagh, F., et al, Linear flaw detection in woven textiles using model-based clustering, Pattern Recognition Letters, 1997, 18: 1539-1548

[54] Dasgupta, A., Raftery, A., Detecting features in spatial point patterns with cluster via model-based clustering, Journal of American Statistical Association, 1998, 93: 294-302

[55] Fraley, C., Raftery, A., How many clusters? Which clustering method? Answers via model-based cluster analysis, Technical Report 329, Dept.

Statistics, Univ. Washington, Seattle, WA, 1998

[56] Schwarz, G., Estimating the dimension of a model, Annals of Statistics, 1978, 461-464

[57] Akaike, H., A new look at the statistical model identification, IEEE Transactions on Automatic Control, 1974, AC-19(6):716-723

[58] Roberts, S., Husmeier, D., Rezek, I., et al, Bayesian approaches to Gaussian mixture modeling, IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20(11): 1133-1142

[59] Hastie, T., Tibshirani, R., Discriminate analysis by Gaussian mixtures, Journal of Royal Statistical Society (B), 1996, 58: 155-176

[60] Hofmann, T., Buhmann, J., Pairwise data clustering by deterministic annealing, IEEE Transactions on Pattern Analysis and Machine Intelligence, 1997, 19(1): 1-14

[61] Meinicke, P., Ritter, H., Resolution-based complexity control for Gaussian mixture models, Neural Computation, 2001, 13(2): 453-475

[62] Blei, D. M., Jordan, M. I., Ng, A. Y., Hierarchical Bayesian models for applications in information retrieval, Bayesian Statistics, 2003, 7: 25-43

[63] Hofmann, T., Probabilistic latent semantic analysis, In: Proceedings of Uncertainty in Artificial Intelligence, UAI'99, 1999

[64] Hofmann, T., Probabilistic latent semantic indexing, In: Proceedings of the 22nd Annual International SIGIR Conference, 1999

[65] Blei, D. M., Ng, A. Y., Jordan, M. I., Latent Dirichlet allocation, Journal of Machine Learning Research, 2003, 3:993-1022

[66] Vapnik, V. N., The nature of statistical learning theory, Berlin, Springer-Verlag, 1995

[67] Platt, J., Using analytic QP and sparseness to speed training of support vector machines, Advances in Neural Information Processing System, 1999, 11: 169-184

[68] 邓爱林, 朱扬勇, 施伯乐, 基于项目评分预测的协同过滤推荐算法, 软件学报, 2003, 14: 1621-1628

[69] Hofmann, T., Collaborative filtering via Gaussian probabilistic latent semantic analysis, In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2003, 259-266

[70] Si, L., Jin, R., Flexible mixture model for collaborative filtering, In: Proceedings of the 20th International Conference on Machine Learning, 2003,

704-711

[71] Liu, X., Gong, Y., Xu, W., et al, Document clustering with cluster refinement and model selection capabilities, In Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2002, 11-15

[72] Zhong, S., Ghosh, J., A unified framework for model based clustering and its applications to clustering time sequences, Technical Report, ECE Dept., University of Texas at Austin, 2002

[73] Yang, Y., Pedersen, J. O., A Comparative Study on Feature Selection in Text Categorization, In: Proceedings of ICML-97, 14th International Conference on Machine Learning, 1997, 412—420

[74] Dash, M., Choi, K., Scheuermann, P., et al, Feature selection for clustering - A filter solution, In: Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM 2002), 2002, 115-122

[75] Dy, J. G., Brodley, C. E., Feature subset selection and order identification for unsupervised learning, In: Proceedings of the 17th International Conference on Machine Learning, 2000, 247-254

[76] Law, M. H. C., Figueiredo, M. A. T., Jain, A. K., Simultaneous feature selection and clustering using mixture models, IEEE Transactions on Pattern Analysis and Machine Intelligence, 26(9): 1154-1166, 2004

[77] Rissanen, J., Stochastic complexity and modeling, Annals of Statistics, 1986, 14(3), 1080-1100

[78] Wallace, C. S., Freeman, P. R., Estimation and inference by compact coding, Journal of Royal Statistics Society (B), 1987, 49, 223-265

[79] Kontkanen, P., Myllymaki, P., Tirri, H., Comparing Bayesian model class selection criteria by discrete finite mixtures, In: Proceedings of the ISIS'96 Conference, 1996, 364-374

[80] Rigouste, L., Cappe, O., Yvon, F., Evaluation of a probabilistic method for unsupervised text clustering, In: Applied Stochastic Models and Data Analysis (ASMDA 2005), 2005

[81] Van Rijsbergen, C. J., Information Retrieval , Butterworths, London,1979

[82] Strehl, A., Ghosh, J., Cluster ensembles – A knowledge reuse framework for combining partitions, Journal of Machine Learning Research, 2002, 3:583–617

[83] Ng, A. Y., Jordan, M. I., Weiss, Y., On spectral clustering: Analysis and an algorithm, Advances in Neural Information Processing Systems, 2002, 14: 849-856

[84] Biernacki, C., Celeux, G., Govaert, G., An improvement of the NEC criterion for assessing the number of clusters in a mixture model, Pattern Recognition Letter, 1999, 20: 267-272

[85] Graham, M. W., Miller, D. J., Unsupervised learning of parsimonious mixtures on large feature spaces, Penn State EE Dept. Technical Report, 2004.

[86] Rosenberg, C., Hebert, M., Schneiderman, H., Semi-supervised selftraining of object detection models, In: The 7th IEEE Workshop on Applications of Computer Vision, 2005

[87] Szummer, M., Jaakkola, T., Partially labeled classification with Markov random walks, Advances in Neural Information Processing Systems, 2001

[88] Zhu, X., Ghahramani, Z., Lafferty, J., Semi-supervised learning using Gaussian fields and harmonic functions, In: Proceedings of the 20th International Conference on Machine Learning, 2003, 20: 912-919

[89] Sindhwani, V., Niyogi, P., Belkin, M., Beyond the point cloud: From transductive to semi-supervised learning, In: Proceedings of the 22nd International Conference on Machine Learning, 2005, 824-831

[90] Fischer, B., Roth, V., Buhmann, J. M., Clustering with the connectivity kernel, Advances in Neural Information Processing Systems, 2004

[91] Sajama, S., Orlitsky, A., Estimating and computing density based distance metrics. In: Proceedings of the 22nd International Conference on Machine Learning, 2005, 768-775

[92] Wagstaff, K., Cardie, C., Rogers, S., et al., Constrained k-means clustering with background knowledge , In: Proceedings of the 18th International Conference on Machine Learning, 2001, 577-584

[93] Wang, L., Bo, L., Jiao, L., A modified k-means clustering with a density-sensitive distance metric, RSKT 2006, LNAI 4062, 2006, 544–551

[94] Bilenko, M., Basu, S., Mooney, R. J., Integrating constraints and metric learning in semi-supervised clustering, In: Proceedings of the 23rd International Conference on Machine Learning, 2004, 81-88

[95] Chapelle, O., Zien, A., Semi-supervised classification by low density separation, AISTATS, 2005

[96] Schölkopf, B., Smola, A. J., Learning with kernels, MIT Press, Cambridge, 2002

[97] Herbster, M., Pontil, M., Wainer, L., Online learning over graphs. In: Proceedings of the 22nd International Conference on Machine Learning, 2005, 305-312

发表论文和参加科研情况说明

1. 发表的论文

[1] 张亮，李敏强，面向协同过滤的真实偏好高斯混合模型，系统工程学报，已录用

[2] 张亮，李敏强，有限混合模型进行无监督文本聚类的一种广义方法，模式识别与人工智能，已录用

[3] 张亮，李敏强，半监督聚类中基于密度的约束扩展方法，计算机工程，已录用

[4] Zhang Liang, Li Min-qiang, Real Preference Gaussian Mixture Model, Journal of Computational Information Systems, 2005, 1(4):663-670

[5] Zhang Liang, Li Min-qiang, A New Mixture Model for Text Clustering, to appear in: The 2007 Workshop on High Performance Data Mining and Application

[6] Zhang Liang, Li Min-qiang, Density-based Laplacian Kernels for Semi -supervised Learning, to appear in: The 2007 Workshop on High Performance Data Mining and Application

2. 参加的科研项目

[1] 国家自然科学基金项目(70171002)

[2] 国家自然科学基金项目(70571057)

[3] 高等学校博士学科点专项科研基金项目(20020056047)

致 谢

论文得以完成，必须要感谢许多人。

感谢我的导师李敏强教授。在我的三年博士生生涯中，李教授在我的学习、工作、生活各方面都给予了悉心指导和亲切的关怀，使我如沐春风、若饮醇醪。三年来，李教授无私地带给我的，是一流的科研环境、广博的学术视野、温馨的实验室氛围。李教授严谨的治学态度、勤奋的工作作风将对我今后的工作产生极大的帮助，必将终生受益。在此对李教授致以诚挚的谢意和深深的祝福。

感谢实验室的各位老师：陈富赞、林丹、南国芳；各位同学：王宏、田津、杨钤雯、李航、刘鎏、王志春；还有我的朋友们：宁禄乔、张正、陈亦平、邹田春。大家的关心和帮助伴我渡过了愉快而难忘的博士生活，谢谢大家。

感谢我的女友宁娴，如奇迹般出现在我的生活中，带给我心灵的宁静和欢乐。

最后感谢多年来父母对我的无私支持和默默奉献。我人生的每一步，都凝聚着他们的心血和期望。

基于机器学习的信息过滤和信息检索的模型和算法研究

作者： 张亮
学位授予单位： 天津大学