江南大学

硕士学位论文

基于EM算法的模型聚类的研究及应用

姓名：岳佳

申请学位级别：硕士

专业：计算机软件与理论

指导教师：王士同

20070601

摘要

在人工智能，模式识别，机器学习等领域中，很多的应用都要用到模型的参数估计。即极大似然估计或极大后验似然估计。EM 算法，又称期望最大算法，就是作为一种参数估计的方法通常用于存在缺失数据的情况下。核心思想就是根据已有的数据来迭代计算似然函数，使之收敛于某个最优值。本文简要介绍了聚类的基础知识，回顾了聚类的典型方法，重点介绍了基于模型的聚类方法。然后，文章深入讨论了 EM 算法并从以下四个方面对 EM 算法进行了深入的研究。

1,实现了基于高斯混合模型的 EM 算法，并针对一个具体的应用实例的数据集，和 Kmeans 方法作了比较，也作为深入研究本文算法的基础。

2,EM 算法收敛的优劣很大程度上取决于其初始参数。运用 EM 算法来实现高斯混合模型的聚类，如何初始化 EM 参数是一个关键的问题。本文在比较其他的初始化方法的基础上，引入用于密度估计的“binning”法来初始化 EM。实验结果表明，应用 binning 法来初始化 EM 的高斯混合模型聚类优于其它传统的初始化方法。

3,半监督聚类是指利用少部分标签的数据辅助大量未标签的数据进行的聚类分析。本文提出了一种基于双重高斯混合模型的 EM 算法，在无监督学习中增加一些有标记的样本，利用已标记的样本得到初始参数，研究了半监督条件下的双重高斯混合模型的 EM 聚类算法。实验结果表明，该算法提升了样本的识别率，具有良好的聚类性能和一定的应用领域。

4,最后，本文研究了基于 Mel 频率倒谱系数和高斯混合模型（GMM）的说话人识别系统。给出了 MFCC 倒谱系数的具体提取过程和算法，并通过实验研究了 EM 的迭代次数和 GMM 模型阶数对识别性能的影响。

关键字：EM 算法；高斯混合模型；双重高斯混合模型；极大似然估计；半监督聚类；初始化；MFCC；说话人识别

Abstract

There are many applications that require the parameter estimation of data model, such as artificial intelligence, pattern-recognition and machine-learning. It is often desired to estimate the maximum-likelihood or maximum-posterior likelihood. EM algorithm, which is named expectation maximum algorithm, is a general-purpose algorithm for maximum likelihood estimation in a wide variety of situations best described as incomplete-data problem. Its core idea is to literately compute the likelihood function until it converges to some optimal value for the given data, This paper introduced the basis of cluster in brief and reviewed the typical cluster methods, and focus on the. In the following of the paper, we further studied the EM algorithm from four aspects:

1, Implemented the EM algorithm and compared with the kmeans algorithm on some application cluster. It also can be used as the basis of the further study.

2, The performance of EM algorithm heavily depends on the initial values of the parameters. When EM algorithm is utilized to realize Gaussian-Mixture-Model based clustering, how to initialize it becomes a pivotal issue. In this paper, the binning method is adopted to initialize EM on the base of comparison of other methods. Our experimental results demonstrate that Gaussian-mixture-model based on clustering using EM with the binning method for initialization outperforms those with other classical initialization methods.

3, Semi-supervised clustering employs a small amount of labeled data to aid clustering analysis. The EM algorithm based on dual Gaussian mixture model has been studied with the added labeled samples as the initial parameters in this paper. Our experimental results demonstrate that the algorithm increase the recognition rate for samples and has good clustering ability and some application fields.

4, Finally, This paper focuses on the speaker recognition system based on Mel-frequency Cepstral coefficients and GMM. It also gives the theory basis and processing arithmetic to compute MFCC in detail; Then influnces on recognition performance of GMM mixure component and the number of EM iteration are discussed by the experiments.

Key words: EM algorithm; Gaussian Mixture Model; dual Gaussian Mixture Model; Maximum Likelihood Estimation; Semi-supervised Cluster; Initialization; MFCC; Speaker recoginition

独 创 性 声 明

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

签名：岳伟　　日期：07年 6月 10日

关于论文使用授权的说明

本学位论文作者完全了解江南大学有关保留、使用学位论文的规定：江南大学有权保留并向国家有关部门或机构送交论文的复印件和磁盘，允许论文被查阅和借阅，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，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文，并且本人电子文档的内容和纸质论文的内容相一致。

保密的学位论文在解密后也遵守此规定。

签名：岳伟　　导师签名：

日期：07 年 6 月 10 日

第一章 绪论

1.1 研究背景和意义

随着人类对自然和社会的认识不断深入，要处理的数据呈现了如下趋势：规模越来越大，相互关系也越来越复杂，分类越来越细，对分类的要求也越来越高。这时仅仅依靠定性分析就不能满足要求。为了解决这个难题，人们开始引入了数学这个得力的工具。在此基础上发展形成了数值分类学（Numerical Taxonomy），对分析对象进行定量的研究[1]。通过定量的研究，分类不再仅仅依靠经验和专业知识，而是通过数据本身的分析来对数据进行分类。数值分类不仅仅可以应用于分类，还可以用到其它领域，于是出现了基于数值分析的聚类分析。

“物以类聚，人以群分”,聚类技术的出发点也是如此。简单来说，聚类就是根据事物之间的相似性把事物聚集成不同类别的一种技术，得到的聚类结果中同类之间相似度较高，而不同类之间的相似度较低。这样，聚类技术就可以把大数据集合中相似度较高的对象聚集在一起，而把相似度较低的对象区分开来。从而使得获得的聚类结果与人们的判断相一致。

通过聚类，人们能够识别密集和稀疏的区域，发现全局的分布模式以及数据属性之间有趣的相互关系。而其本身的研究也是一个蓬勃发展的领域，数据挖掘、统计学、机器学习、空间数据库技术、生物学和市场学的发展推动着聚类分析研究的进展，使它已成为数据挖掘研究中的一个热点。与其他数据挖掘方法不同，在进行聚类分析前用户一般并不知道数据集的特征[2]。因此，从某种角度看，聚类分析是一种无监督的学习过程，是基于观察的学习而不是基于实例的学习。作为数据挖掘中的一个模块，聚类分析可作为一个独立的工具来获取数据分布的情况，观察每个簇的特点，集中对特定的某些簇做进一步分析。如在商务上，聚类分析可以帮助市场分析人员从客户基本库中发现不同的客户群，并且用购买模式来刻画不同的客户群的特征。聚类分析也可以作为数据挖掘中其他算法（如特征和分类等）的预处理步骤，这些算法再在生成的簇上进行处理；此外它还可以完成孤立点挖掘。许多数据挖掘算法试图使孤立点影响最小化，或者排除它们。然而孤立点本身可能是非常有用的，如在欺诈探测中，孤立点有可能预示着欺诈行为。迄今为止，人们提出了大量的聚类算法，算法的选择取决于数据的类型、聚类的目的和应用。如果聚类分析用于描述或探索的工具，可以对同样的数据尝试多种算法，以便发现数据可能隐含的规律与结果[3]。综上，我们可以发现，聚类分析已成为了一个非常活跃的研究课题。

1.2 研究现状

1.2.1 聚类简介

在统计方法中，聚类称聚类分析，它是多元数据分析的三大方法之一（其它两种是回归分析和判别分析）。它主要研究基于几何距离的聚类，如欧式距离、明考斯基距离等[2]。传统的统计聚类分析方法包括系统聚类法、分解法、加入法、动态聚类法、有序样品聚类、重叠聚类和模糊聚类等。这种聚类方法是一种基于全局比较的聚类，它需要考察所有的个体才能决定类的划分;因此它要求所有的数据必须预先给定，不能动态地增加新的数据对象。聚类分析方法不具有线性的计算复杂度，难以适用于数据库非常大的情况。

聚类分析问题可描述为：给定 m 维空间 R^m 中的 n 个向量，把每个向量归属到 S 聚类中的某一个，使得每个向量与其聚类中心的“距离”最小。聚类分析问题的实质是一个全局最优问题。在这里，m 可认为是样本参与聚类的属性个数，n 是样本的个数，S 是由用户预先设定的分类数目。

定义：对于 m 维空间 R^m 中的向量：

$$X_i, X_j, X_i = \{X_{i1}, X_{i2}, ..., X_{im}\}, X_j = \{X_{j1}, X_{j2}, ..., X_{jm}\},$$

向量 X_i, X_j 之间的距离为：

$$d_{ij} = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2} \tag{1.1}$$

聚类与数据挖掘中的分类不同。在分类问题中，我们知道训练例的分类属性值，在那里我们要做的就是将每一条记录分别属于哪一类标记出来；与此相似但又不同的是，聚类分析的输入数据集是一组未标记的对象，也就是说此时输入的对象还没有被进行任何分类，聚类的目的是根据一定的规则，合理地进行分组或聚类，并用显式或隐式的方法描述不同的类别。由于分析可以采用不同的算法，所以对于相同的数据集合可能有不同的划分。在机器学习中，聚类是无监督学习的一个例子，分类是有监督学习的一个例子，两者所采用的方法相差甚远，并且聚类的时间复杂度要比分类大得多。

1.2.2 模型的研究现状

对混合模型的研究[3]最早可以追溯到一百年以前，皮尔逊在 1894 年，用具有两个混合元的正态混合模型对一组数据进行了拟和，用矩估计的方法对模型的参数进行了估计，这是有关混合模型最早的研究。在接下来的三十年内，研究人员一直是用矩估计的方法对混合模型参数进行估计，Charlier and Wicksell

在1924年开始讨论二元的正态混合模型，Doetcsh在1928年又把混合元的个数推广到具有多个混合元的情形。但是由于矩估计的方法计算太复杂，所以混合模型的研究进展缓慢。直到计算机的出现，很多复杂的计算问题才得以解决，Tan和Robertson等人在1972年开始用极大似然法对混合模型进行研究，并证明了极大似然比矩估计的方法要好很多[2]。

随着高速运行计算机的出现，人们逐渐把研究的注意力集中到用极大似然估计的方法对混合模型的参数进行估计。Dempster等人在1977年，用EM算法对极大似然估计进行计算使得计算困难迎刃而解。之后，有关混合模型的研究进入了一个新的阶段，研究领域也扩展到了医学，经济学等其他科学。

虽然极大似然方法使得混合模型的研究取得了很大的发展，但是该方法只能在混合元个数已知的假定下，才能对参数进行估计，而对于混合元个数k未知的情况下，则无法给出k的估计，因此研究人员有开始寻找别的方法来解决这一难题。有关混合模型的 贝叶斯估计的最早文章见Diebolt and Robert（1994）[4]。在Richardson and Green（1997）中，作者采用可逆条MCMC方法对未知混合元个数的一元正态混合模型进行了贝叶斯分析，给出了一种估计混合元个数k的贝叶斯方法，在Stephens，M(2000)[5]者构造了一个连续状态空间的马尔可夫链，给出了另外一种可逆条MCMC方法。

混合模型最具代表性的用处是用来对数据进行聚类。假定一组观察值来于k个总体，每个数据被认为是来自于k（k可能未知的）个分布总体中的某一个，混合权ω_i表示观测值来自于第i个总体的频率。这样混合模型提供了一个模拟数据的工具。通过对混合模型的研究，可以对一组来自于不同总体的数据进行分类。此外，混合模型还可以用来对那些不能用标准的参数分布族来拟和的总体进行密度估计或近似。

1.3 需要研究的方向

聚类[7,8,9,10]，是一个富有挑战性的研究领域，其研究工作集中在为大型数据库的有效和实际的聚类分析寻求适当的方法，目前的研究方向包括下列几个方面[6]

（1）算法的可伸缩性:在很多聚类算法中，数据对象小于200个的小数据集合上鲁棒性执行多种数据模型;而对于包含几百万个数据对象的大规模数据库进行聚类时,将会导致有不同的偏差结果。这就需要聚类算法具有高度的可伸缩性，能有效地处理海量数据。

（2）处理不同类型属性的能力：对于设计的很多算法用于聚类数值类型的数据。但在实际应用中可能要求聚类其它类型的数据。

（3）发现任意形状的聚类：许多聚类算法是基于欧几里德距离或者曼哈坦距离，趋向于发现具有相近密度和尺寸的球状簇。但一个簇可能是任意形状的。提出能发现任意形状簇的算法非常重要。

（4）用于决定输入参数的领域知识最小化：在聚类分析中，许多聚类算法要求用户输入一定的参数，如希望簇的数目。聚类结果对于输入参数很敏感，通常参数较难确定，尤其是对于含有高维对象的数据集更是如此。

（5）对于输入记录顺序不敏感：一些聚类算法对于输入数据的顺序是敏感的。如对于同一个数据集合，以不同的顺序提交给同一个算法时，可能产生差别很大的聚类结果。研究和开发对数据输入顺序不敏感的算法具有重要的意义。

（6）高维性：一个数据库可能含有若干维或者属性。很多聚类算法擅长处理低维数据，一般只涉及两到三维。通常最多在三维的情况下能够很好地判断聚类的质量。聚类数据对象在高维空间是非常有挑战性的，尤其是考虑到这样的数据可能高度偏斜，非常稀疏。

（7）处理噪声数据的能力：在现实应用中绝大多数的数据都包含了孤立点，空缺、未知数据或者错误的数据。有些聚类算法对于这样的数据敏感，将会导致质量较低的聚类结果。

（8）基于约束的聚类：在实际应用中有可能需要在各种约束条件下进行聚类。既要找到满足特定的约束，又要具有良好聚类特性的数据分组是一项具有挑战性的任务。

（9）可解释性和可用性：通常用户希望聚类结果是可解释的，可理解的和可用的。因此，应用目标如何影响聚类方法的选择也是一项重要的研究课题。

1.4 本文的研究应用

说话人识别[38][39]指的是根据说话人所发的语音来确定说话人的过程，也就是将声音这种生物特性作为身份认证依据的识别技术。为此，需要从各个说话人的发音中找出说话人之间的个性差异。说话人识别是交叉运用心理学、生理学、语音信号处理、模式识别、统计学习理论和人工智能的综合性研究课题。

说话人识别可以看作是语音识别的一种，它和语音识别一样，都是通过对所接收的语音信号进行处理，提取相应的特征，建立相应的模型，然后据此做出判断。

基于高斯混合模型的说话人识别方法是现代说话人识别技术的重要方法之一。说话人识别是模式识别的一种，而高斯混合模型则是属于统计模式识别的一种方法。人体的发音器官和过程都是很复杂的，直接对人体的发音过程建立模型显然是非常困难的，但可以把声音的产生抽象为一个随机过程(实际上是从

声音中提取出的特征参数在特征空间中的分布是一个随机过程)，这样针对发音过程就可以建立一个概率模型，而高斯混合模型就是这样的一个概率模型。由于它的性能较好、复杂度小、方法简单，所以被认为是当前最好的说话人识别模型之一。

说话人识别这种技术有着广阔的市场应用前景。通过这种技术，可以利用人本身生物特性进行身份鉴别，例如为公安部门进行语音验证，为一般用户提供防盗门开启功能等等。在互联网应用及通信领域，说话人识别技术可以应用于诸如声音拨号、电话银行、电话购物、数据库访问、信息服务、语音 E-mail、安全控制、计算机远程登录等领域。在军事领域，可以用于战场上的侦听，以辨认出敌方的指挥员。在医疗领域，可以用于患者的确认等等。

本文在深入研究了基于高斯混合模型的 EM 算法的基础上，把算法放到了应用领域，即高斯混合模型的说话人识别上。研究了高斯混合模型的阶数和 EM 算法的迭代次数对系统识别性能的影响。

1.5 论文的内容及章节安排

本论文分为六章，各章的主要内容安排如下：

第一章是绪论，介绍了本论文的研究背景和现状。以及研究的应用领域——高斯混合模型的说话人识别。

第二章是聚类的基础知识。这一章主要对聚类的基础知识作了一般性介绍。详细介绍了聚类的典型方法和使用模型来描述数据，重点介绍了我们要使用的混合模型的概念。

第三章是 EM 算法及其初始化问题的研究。这一章重点研究了 EM 算法，混合模型聚类的原理。详细介绍了高斯混合模型的 EM 算法，并通过实验比较了其和 Kmeans 方法的聚类效果。随后，讨论了 EM 算法的初始化问题，提出了用 binning 方法作为一种初始化方法，并通过两个具体的实验验证了结果。

第四章是基于半监督的 EM 算法的研究。介绍了贝叶斯学习理论的基本观点以及最大后验概率。随后，详细介绍了本文的双重高斯混合模型的 EM 算法，并通过实验，在半监督条件下，验证了算法的可行性和可操作性。

第五章是高斯混合模型在说话人识别中的应用。在这一章里，先详细介绍了说话人识别的特征提取技术——MFCC 以及 GMM 模型的训练和识别方法。接着通过具体的实验，研究了高斯混合模型的阶数和 EM 算法的迭代次数对系统识别性能的影响。

第六章是总结与展望。在这一章里，主要是对本文进行了回顾和总结，并对以后的工作方向进行了展望。

第二章 聚类的基础知识

通过第一章的介绍，我们已经对本论文的研究背景，研究的方向有了一个大致的了解。在这一章里，我们将对聚类的基础知识作一般性的介绍。为下文中，对高斯混合模型的 EM 算法的聚类研究工作打下基础。

2.1 聚类分析

2.1.1 类的定义

由于客观事物纷繁芜杂的特性，以及我们在特征提取过程中用来表示样本点性质的特征变量的不同选择，使得样本点的表示很不相同。在不同的问题中关于类的定义也是不同的。要想给类下一个通用严格的定义看来是不可能的，[7]提出以下几种不同的类的定义，不同的定义适用于不同的应用场合。

设G表示一个有k个样本的集合，S_i表示其中的样本，T和V为预设阈值。

定义 1：如果对于任意$S_i,S_j \in G$，都有$D(S_i,S_j) \le T$，则G称为一个类。

定义 2：如果对于每个$S_i \in G$，都有$\frac{1}{k-1}\sum_j D(S_i,S_j) \le T$，那么$G$称为一类。

定义 3：如果对于每个$S_i,S_j \in G$，都有

$$\frac{1}{k*(k-1)}\sum_i\sum_j D(S_i,S_j) \le T \text{ 且 } D(S_i,S_j) \le V,$$

那么G称为一类。

定义 4：对于任意样本S_i，都存在G中一个样本S_j，满足$D(S_i,S_j) \le T$，则G称为一类。

以上几种定义中，定义 1 是要求最高的，凡是满足定义 1 要求的类，肯定满足其他几种定义。凡是满足定义 2 的集合，也必定满足定义 3。

设有类G，类中共有k个样本，我们常常从以下几个角度来刻画类的特征：

（1）均值和中心

$$\bar{S} = \frac{1}{k}\sum_i S_i \tag{2.2}$$

（2）样本协方差矩阵C，表示样本点的散布程度

$$C = \frac{1}{k-1}\sum_i (S_i - \bar{S})(S_i - \bar{S})^T \tag{2.3}$$

（3）类的直径R

$$R=\frac{1}{k-1}\sum_{i}(S_i-\bar{S})^T(S_i-\bar{S}) \tag{2.4}$$

2.1.2 聚类的典型方法

目前存在的并得以应用的聚类算法[3][8][9]很多，但就其主要思想，可将其划分为如下几类：

1．划分方法(partitioning method)：

给定一个n个对象或元组的数据库，一个划分方法将构建数据的k个划分，每个划分表示一个聚簇，并且$k \le n$。也就是说，它将数据集合划分为k个组，同时满足如下要求：(1)每个组至少包含一个对象；(2)每个对象必须属于且只属于一个组。注意在某些模糊划分技术中第二个要求可以放宽。给定要构建的划分的数目k，划分方法首先创建一个初始划分。然后采用一种迭代重定位技术，尝试通过对象在划分之间移动来改进划分的效果。一个好的划分的一般准则是：在同一个簇中的对象之间尽可能的“接近”或相关，而不同簇中的对象之间尽可能“远离”或不同。为了达到全局最优，基于划分的聚类会要穷举所有可能的划分。实际上，绝大多数应用采用以下两种比较流行的启发式划分方法：(1)k—均值算法，在该算法中，每个簇使用该簇中的平均值表示。(2)k—中心点算法，在该算法中，每个簇使用接近聚类中心的一个对象表示。这些启发式聚类方法在中小规模的数据库中发现球状簇很好使用。为了对大规模的数据集进行聚类，以及处理复杂形状的聚类，基于划分的方法需要近一步的扩展。

2．层次的方法(hierarchical method)：

层次的方法对给定数据对象集合进行层次的分解。根据层次的分解如何形成，层次的方法可以分为凝聚的和分裂的。凝聚的方法，也称为自底向上的方法，一开始每个对象作为单独的一组，然后相继地合并相似的对象或簇，直到所有的簇合并为一个(层次的最上层)，或者达到一个终止条件。分裂的方法，也称为自顶向下的方法，一开始将所有的对象置于一个簇中。在迭代的每一步中，一个簇被分裂为更小的簇，直到最终每个对象在单独的一个簇中，或者达到一个终止的条件。

3．基于密度的方法(density-based method)：

绝大多数划分方法基于对象之间的距离进行聚类。这样的方法只能发现球状的簇，而在发现任意形状的簇上遇到了困难。随之提出了基于密度的另一类聚类方法，其主要的思想是：只要临近区域的密度(对象或数据点的数目)超出了某个阀值，就继续聚类。也就是说，对给定类中的每个数据点，在一个给定范围的区域中必须至少包含某个数目的点。这样的方法可以用来过滤“噪声”孤立点数据，发现任意形状的簇。DBSCAN 是一个有代表性的基于密度的方法，

它根据一个密度阀值来控制簇的增长。OPTIC 是另一个基于密度的方法，它为自动的和交互的聚类分析计算一个聚类顺序。

4. 基于网格的方法(grid-based method)：

基于网格的方法把对象空间量化为有限数目的单元，形成了一个网格结构。所有聚类操作都在这个网格结构(即量化了的空间)上进行。这种方法的主要优点是它的处理速度很快，其处理时间独立于处理对象的数目，只与量化空间每一维的单元数目有关。STING 是基于网格方法的一个典型例子。

5. 基于模型的方法(model-based method)：

基于模型的方法为每个簇假定了一个模型，寻找数据对此模型的最佳拟合。一个基于模型的算法可能通过构建反映数据点空间分布的密度函数来定位聚类，它也可能基于标准的统计数字自动决定聚类的数目，考虑“噪声”数据和孤立点，从而产生健壮的聚类方法。在实际应用中，一些聚类算法可能集成了多种聚类方法的思想，所以有时将某个给定的算法划分为属于某类聚类方法是很困难的。近年来，以模型为基础的数据分析方法，得到了人们的关注。它的主要思想是假设数据空间中的每一个数据都是产生于一个统一的模型。例如，当用某种概率密度模型表示数据空间的时候，那么假设数据空间中的每一个数据都服从该概率分布。在确定了产生数据的模型之后，可以通过数学的方法调整模型的各种参数，使得模型能够很好的拟合数据空间，这样就得到了一个可以用来对现有或者未知数据进行分析的模型。模型可以用很多的形式进行表示，其中概率分布是一种典型的表示形式，例如高斯分布或者多项分布。在确定了概率模型之后，需要用数学的方法使模型与数据拟和，这其中最常用的方法是 EM 方法（在下文中会详细介绍），M 步骤帮助确定概率分布中的各种参数。目前，已经研究得出许多的概率模型聚类技术，并且这些技术已经在各个不同的领域产生了很好的效果。

2.1.3 聚类的评价标准

聚类方法的优劣标准本身就是一个值得研究的问题[3]。聚类方法的性能很大程度上取决于聚类的类型。

那么一个好的聚类方法应该满足什么条件呢？现在通用的聚类标准都是从几个方面来衡量，而没有完全使用量化的客观标准。下面给出了六条关于聚类的主要标准：

1. 处理大的数据集合的能力；
2. 处理任意形状，包括有间隙的和嵌套的数据的能力；
3. 算法处理的结果与数据输入的顺序是否相关，也就是说算法是否独立于数据输入顺序；

4. 处理数据噪音的能力；
5. 需不需要预先知道聚类个数，需不需要用户给出领域知识；
6. 算法处理有很多属性的数据的能力，也就是对数据维数是否敏感；

所以，对于一个聚类算法的优劣可以从这几个方面综合衡量。

2.1.4 聚类分析的应用

聚类算法[9][10]被用于许多知识领域，这些领域通常要求找出特定数据中的“自然关联”。自然关联的定义取决于不同的领域和特定的应用，可以具有多种形式。典型的应用例如：商务上，帮助市场分析人员从客户基本资料库中发现不同的客户群，并用购买模式来刻画不同客户群的特征；生物学上，用于推导植物和动物的分类，对基因进行分类，获得对种群固有结构的认识；地理信息方面，在地球观测数据库中相似区域的确定、汽车保险单持有者的分组，及根据房子的类型、价值和地理位置对一个城市中房屋的分组上可以发挥作用；聚类也能用于对文档进行分类.此外，聚类分析可以作为其它数据挖掘算法的预处理步骤，便于这些算法在生成的簇上进行处理。

2.2 使用模型描述数据

2.2.1 模型简介

数据挖掘[9]用于发现大量数据中所蕴含的知识或者是规则，这种发现是建立在数据表示之上的，也就是说在建立了正确的数据表示以后，才能利用各种方法来对数据潜在的信息进行挖掘，所以构造恰当的，并且是适用于当前数据分析的数据表示形式显得尤为重要。通常把能反映整个数据集中某一个部分或者个体的表示称为模式(pattern)。但是这种描述只是局限于数据的局部，只能反映个体数据的情况，不能从全局出发考虑整个数据集合的结构，而在对数据进行分析处理的过程中，不仅仅要从局部出发，更多的时候还需要从全局的角度来进行研究，这个时候就需要建立能够反映数据整体结构的“模型”(model)。所以一般可以将数据的描述分为两种：一种是全局的模型(model)，一种是局部的模式(pattern)。这里把模型定义为对数据集的全局性总结，它对整个测量空间的每一个点做出全局性的描述。例如，如果把数据矩阵的各行看作 p 维向量(也就是 p 维空间中的点)，那么模型可以对这个空间中的每一点(也就是所有对象)做出描述。通过对于这种全局性表示进行分析，可以把一个点分配到一个聚类或者预测出某一个模型变量的值。模型是对一个数据集的高层次、全局性的表示。它能通过一个很大的样本透视总体。模型可以是描述性的——以方便简捷

的方式归纳数据；也可以是推理性的，允许对数据所在的数据总体或者未来数据做出某种判断。一个简单的模型可以表示成这样$Y = aX + c$，其中Y和X是变量，a和c是模型中的参数，通过这个模型可以看出，他重点描绘的并不是某一个数据部分，而是对整个数据空间做出了一个表示，所有符合上面模型的数据，都要落在模型描述的空间中。当然这只是一个比较简单的模型，在实际的应用中，一般使用的模型要更为复杂。但是这个简单的例子已经足够用来了解模型的基本含义。在这个模型表述了变量Y与 X之间的关系，其中的a和c通常称之为参数，在实际的使用中经常用θ来表示一般参数或者一系列参数的向量。在此实例中，$\theta = \{a,c\}$。在给定的模型形式或者结构以后，接下来的任务就是通过估计为模型选择适合的参数值——也就是选择一个适合的评分函数来衡量模型与数据之间的拟合情况，然后通过最小化或者最大化为该函数选择合适的参数。

2.2.2 概率模型

对于很大的总体中抽取出的数据，或者可以被看作是从很大总体中抽取出的数据，通过潜在的分布或者是密度函数来描述他们是一种很基本的策略，也就是说在分析数据的时候可以假定这些总体当中的数据是服从于某一个概率分布的。比如经常使用标准正态分布来假设一些现实世界中的一些事情的发生情况等等。概率模型[3]在描述数据方面有着广泛的应用，可以使用概率模型来完成诸如对数据的预测、数据总体的描述的问题。对于很多预测问题经常会遇到这样的情况：对于一个未知的或者待估计的变量(标示为Y)，使用其它变量对其做出预测。数据挖掘中很多的建模问题都属于这一类。另外，还有许多的建模问题是“描述性”的，目标是给出对数据的描述或总结。如果现有数据是完整的(如某一类化合物的全部)，那么就不存在任何推理概念，目标就是简化描述。另一方面，如果现有的数据是一个样本或者带有误差的测量值(因而如果在采集一次数据，那么可能会得到略微不同的值)，那么建模的目的实质上是一种推理——推理出“真实”或者至少是比较好的模型结构。所以，在一般的情况下，可以把要分析的数据假定为由一个潜在的概率函数产生的。也就是说数据空间中的点是服从于某一个概率模型。可以把概率模型分成两类：

1. 参数模型：这种模型采用一种特定的函数形式。对于实数便利经常使用位置参数(平均值)和范围(scale)参数(刻画变化性)来表征这种函数——例如正态分布和二项分布函数。参数模型的优点在于简单明了(易于估计和解释)，但是可能偏差相对较大，因为真实数据可能不遵循假定的函数形式。
2. 非参数模型：在这种模型中分布和密度估计是数据驱动的，事先仅对函数形式作少量的假设。如果把上述两种情况视为极端情况，那么还可以定义一种介

于参数模型和非参数模型之间的中间模型：混合模型(mixture model)。下面的部分将讨论有关混合模型的问题。

2.2.3 混合模型

关于变量 x 的混合密度是这样定义的：$p(x)=\sum_{k=1}^{K}\pi_k p_k(x|\theta_k)$。模型把 X 的整个密度分解为 K 个分量(component)或者类(class)的加权现行组合。每个分量密度 $p_k(x|\theta_k)$ 通常是由一种相对简单的参数模型(参数为 θ_k，这里的 θ_k 可以是一个参数向量)(比如一个正态分布函数)组成的。π_k 代表一个随机抽取的数据点是由第 k 个分量产生的概率，$\sum_k \pi_k = 1$。着 K 的增大，混合模型可以具有非常灵活的函数形式。因为局部的分量函数可用来捕捉局部的数据特征，而混合模型是将所有这些局部的特征综合在一起，形成一个更为细致的分布函数。这样显然要比用一个单一的密度函数去定义全局的特征的效果要好得多。另外 K 的值还控制着模型的复杂程度：因为 K 值越大，得到的模型越灵活，对数据的表示更好，但是同时解释也越复杂、拟合也越困难。以二元正态分布组成的混合模型为例，通过下面图 2.1，说明混合模型的原理。

从上至下：

（a）三个等权值的二元正态分布组合的混合模型产生的数据点。

（b）距离平局值 3σ 处等高线的潜在分量密度。

（c）总的混合密度产生的等高线。

图 a 显示了由三个等权值的二元正态分布组成的混合模型产生的数据点，图 b 表明了这三个分量密度，每一个椭圆表示一个二元正态分布，图 c 是混合密度产生的等高线。由这组图形，可以看出混合模型利用分量捕捉局部特性，然后再局部整合在一起以后能够很好的对数据进行表示。混合模型所蕴含的一般原理具有广泛的用途，这种思想被应用在概率建模的许多领域。例如，使用混合模型能够很好的捕捉层次结构，而且混模型也已经成功地应用在探测数据中的聚类。

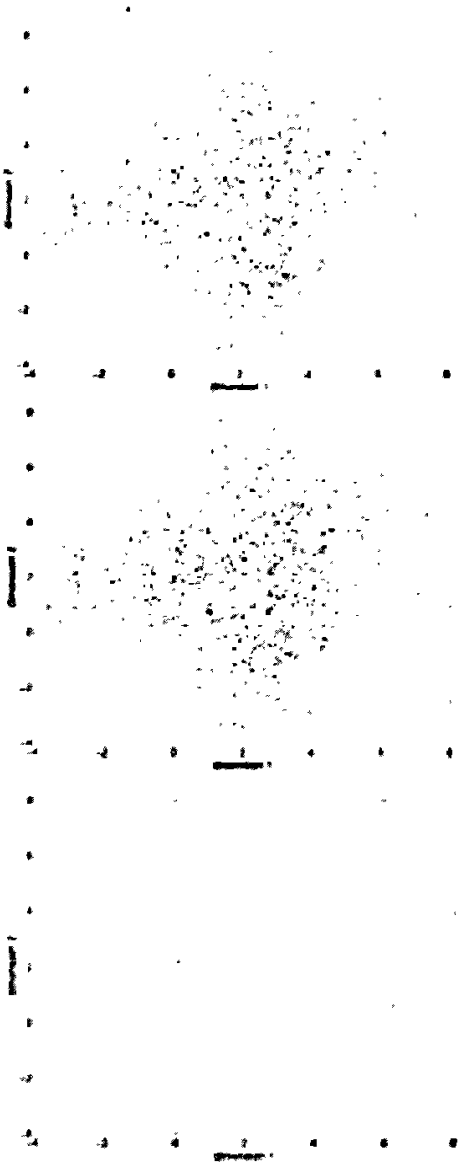

图 2.1 混合模型实例

第三章 EM 算法及其初始化问题的研究

在这一章里，将对 EM 算法作深入的研究。先概述该算法，并讨论了混合模型聚类的原理和基本方法。接着，对高斯混合模型的 EM 算法作了详细的介绍。通过一个实验数据，和 Kmeans 方法作了比较。鉴于 EM 算法对初始值比较敏感，本章讨论了利用不同的初始化方法对 EM 聚类的影响。在比较了传统的方法后，引入用于密度估计的 binning 法，试图用在初始化 EM 的方法上。通过两个实验数据，验证了该方法具有一定的优越性。

3.1 EM 算法

3.1.1 EM 算法的含义

EM 算法是一种迭代算法，是由 Dempster，Laind，Rubin 于 1977 年提出的求参数极大似然估计的一种方法，它可以从非完整数据集中对参数进行最大似然估计，是一种非常简单实用的学习算法[11][12][13]。这种方法可以广泛地应用于处理缺损数据，截尾数据，带有讨厌数据等所谓的不完全数据(incomplete data)。

贝叶斯计算方法大体可以分为两大类。一类是直接运用于后验分布以得到后验均值或后验众数的估计，以及这种估计的渐进方差或其近似。另一类可以总称为数据添加算法，这是近年来发展很快且应用很广的一种算法，它不是直接对复杂的后验分布进行极大化或模拟，而是在观察数据的基础上添加一些“潜在数据”，从而简化计算并完成一系列简单的极大化或模拟。EM 算法就是一种一般的从“不完全数据”中求解模型参数的极大似然估计的方法。所谓“不完全数据”，一般两种情况：一种是由于观察本身的限制或者错误，造成观察数据成为错漏的不完全数据，一种是参数的似然函数直接优化十分困难，而引入额外的参数（隐含的或丢失的）后就比较容易优化，于是定义原始观察数据加上额外数据组成“完全数据”，原始观察数据自然就成为“不完全数据”。

EM 算法的每一步迭代中包括一个 E 步——期望步（Expectation Step）和一个 M 步——极大似然步（Maximum Likelihood Step）。算法的优势在于它在一定意义下可靠地收敛到局部极大，也就是说在一般条件下每次迭代都增加似然函数值，当似然函数值是有界的时候，迭代序列收敛到一个稳定值的上确界。EM 算法的缺点是当缺失数据比例较大时候，它的收敛比率比较缓慢。

3.1.2 EM 算法的原理

基本原理[14]可以表述如下：我们可以观察到的数据是Y，完全数据$X=(Y,Z)$，Z是缺失数据，θ是模型参数。θ关于Y的后验分布$p(\theta|Y)$很复杂，难以进行各种不同的统计计算。假如缺失数据Z已知，则可能得到一个关于θ的简单的添加后验分布$p(\theta|y,z)$，利用$p(\theta|y,z)$的简单性我们可以进行各种统计计算。然后，回过头来，我们又可以对Z的假定作检查和改进。如此进行，我们就将一个复杂的极大化抽样问题转化为一系列简单的极大化或抽样问题。

EM 算法本质上与多元空间中的局部爬山形式很相似，E和M步骤隐含(而且自动的)确定每一步的方向和距离。因此，与爬山算法一样，EM 算法对初始条件敏感，所以选取不同的初始条件可以得到不同的局部最大值。正因为如此，本文后面将对不同的初始化方法对 EM 聚类结果的影响进行研究。EM 算法可能相当慢的收敛到最终的参数值，所以可以把它与传统的优化技术一起用力加速收敛。虽然如此，标准的 EM 算法因为具有广阔的适用范围和可以相当轻松地移植到各种不同的问题而被广为应用。

3.2 混合模型聚类的原理

对于数据集$X=\{X_1,\cdots,X_n\}$，用混合模型聚类[15]就是将数据划分到它最可能的簇中去（或成分中去），拟合出一个的混合分布。这里隐藏了一个重要的事实就是要找到数据集X的最可能的标签向量集$Z=\{Z_1,\cdots,Z_n\}$，X_i的标签向量为$Z_i=(z_{i1},\cdots,z_{iG})^T$，如果$X_i$的所在簇完全确定了（比如$X_i$在第 k 个簇中），那么$z_{ik}=\begin{cases}1 & X_i\text{在第k个簇中}\\ 0 & \text{其他}\end{cases}$。$Z_i$通常认为是相互独立的随机变量，且有$p(z_{ik}=1|\theta)=\pi_k$，这样$Z_i$服从一多项分布，记为$Z_i\sim M_G(\pi_1\cdots\pi_G)$，$(i=1,\cdots,n,k=1,\cdots,G)$。要得到$X$的最有可能的标签向量集，就要合理的从数据中估计模型的参数θ（这里实际是估计参数集θ中的各个参数，后面出现这种提法可类似理解），使得$\forall X_i\in X$都能找到它最有可能的标签。

从上面的聚类思想中可以看出，标签向量集Z可以看成是数据集中的隐含变量集（缺损变量集），如果混合模型各混合成分的形式是明确的，而且我们从数据中学习知道θ的当前极大似然是$\hat{\theta}$，那么我们就可以给数据集的每个观察

数据 X_i 一个类概率 t_{ik}，计算表示式为：

$$t_{ik}(X|\hat{\theta}) = pr(z_{ik}=1|X_i;\hat{\theta}) = \frac{\hat{\pi}_k h(X_i|\hat{\lambda}_k)}{\sum_{j=1}^{G}\hat{\pi}_k h(X_i|\hat{\lambda}_j)}(1\le i\le n),(1\le k\le G) \tag{3.1}$$

这样 X_i 的标签变量 Z_i 就完全由这些类概率决定，其中最大的一个标志它所在的簇。从这里我们可以看出模型聚类的实质就是寻找到每个观察数据的最大后验类概率。

3.3 高斯混合模型的 EM 算法的实现

3.3.1 极大似然方法

由于极大似然[16]的渐进最优性质，它已成为参数估计的一种常用方法。极大似然估计是以使观测值出现的概率最大作为准则的。

设 x 为连续随机变量，其分布密度函数为 $p(x|\theta)$，$\theta=\{\theta_1,.....,\theta_M\}$，即该分布密度函数是由参数 θ 决定的。假设从分布密度为 $p(x|\theta)$ 的总体中独立抽取 n 个观测值，记 $X=\{x_1,.....,x_n\}$，则：

$$p(X|\theta)=\prod_{i=1}^{n}p(x_i|\theta)=L(\theta|X) \tag{3.2}$$

函数 $L(\theta|X)$ 称为似然函数。当 X 固定是，$L(\theta|X)$ 是 θ 的函数，极大似然参数估计的实质就是求出使得 $L(\theta|X)$ 达到极大值的 θ，即：

$$\theta^{*}=\arg\max L(\theta|X) \tag{3.3}$$

为了计算的方便，我们通常用求 $\log(L(\theta|X)$ 的最大值来代替 $L(\theta|X)$。

假设 $p(x|\theta)$ 是一维高斯分布，$\theta=(\mu,\sigma^2)$，我们可以令 $\log(L(\theta|X)=0$，直接求出 μ 和 σ^2，得到均值和方差的标准方程：

$$\hat{\mu}=\frac{1}{n}\sum_{i=1}^{n}x_i \qquad \sigma^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu})^2 \tag{3.4}$$

3.3.2 高斯混合模型

假设有一系列观测值由某混合分布 P 产生，该分布又是由 G 个成分构成，每一个成分都代表一个不同的类别（cluster），假设观测样本 $X=\{x_1,.....,x_n\}$，每个向量 x_i 都是 p 维的矢量。$f_k(x_i|\theta_k)$ 表示 x_i 是第 k 类的密度函数，θ_k 是相应的参数。最大化混合似然函数：

$$L_M(\theta_1,.....,\theta_G;\gamma_1,.....,\gamma_n \mid x) = \prod_{i=1}^{n}\sum_{k=1}^{G}\pi_k f_k(x_i \mid \theta_k) \tag{3.5}$$

其中π_k是某一观察值属于第 k 类的概率。$(\pi_k \geq 0; \sum_{k=1}^{G}\pi_k = 1)$

如果$f_k(x_i \mid \theta_k)$是多元正态分布，即高斯分布，则此混合聚类的模型即为我们的高斯混合模型（GMM），G 个成分即为 G 个独立同方差的高斯分布[21][22]。参数θ_k由均值μ_k和协方差矩阵Σ_k组成。密度函数$f_k(x_i \mid \theta_k)$具有以下形式[33]：

$$f_k(x_i \mid \mu_k, \Sigma_k) = \frac{\exp\left\{-\frac{1}{2}(x_i - \mu_k)^T \Sigma_k^{-1}(x_i - \mu_k)\right\}}{(2\pi)^{p/2}|\Sigma_k|^{1/2}} \tag{3.6}$$

所以该分布P可由G个高斯密度函数的加权平均所表示的概率密度函数描述如下：

$$P(x \mid \theta) = \sum_{k=1}^{G}\pi_k f_k(x_i \mid \mu_k, \Sigma_k) \tag{3.7}$$

π_k是混合模型中基模型高斯密度函数的权重。

可以用下图来表示高斯混合模型：

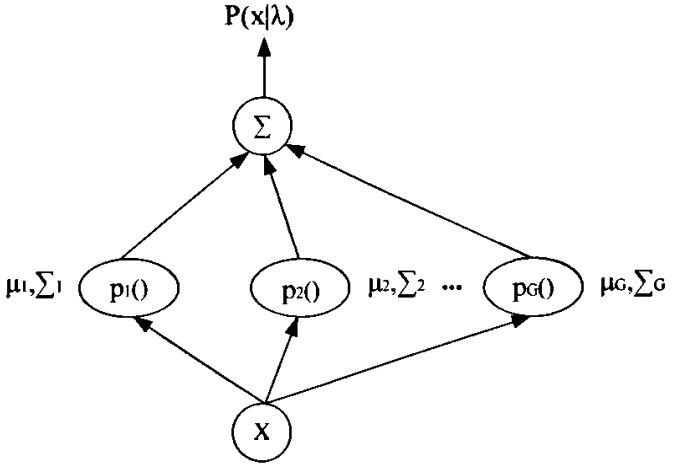

图 3.1 高斯混合模型示意图

由此可见，高斯混合模型的各个分量$p_k(x_k)$可由均值向量μ_k和协方差举证Σ_k来描述，故上述 GMM 模型可由参数集$\lambda = \{\pi_k, \mu_k, \Sigma_k (k = 1,2,\cdots,G)\}$来表示。

聚类是以均值μ_k为中心的椭圆体分布。其它的几何学性质（方向，体积，形状）由协方差矩阵Σ_k决定。Banfield 和 Raftery[12]提出了分解Σ_k的特征值的模型聚类的框架：$\Sigma_k = \lambda_k D_k A_k D_k^T$。

3.3.3 聚类的 EM 算法

假设存在一个完整数据集$Y=(X,Z)$，$X=\{x_1,.....,x_n\}$是不完整的数据集，Z_i是引入的隐含变量。$Z_i\in\{1,2,.....,M\}$，M是给定的有限整数。于是$Y=\{(x_1,z_1),.....,(x_n,z_n)\}$则完整数据的似然函数为：

$$L(\theta|X,Z)=p(X,Z|\theta)=\prod_{i=1}^{n}p(x_i,z_i|\theta)\qquad Z=\{z_1,.....,z_n\}\tag{3.8}$$

该似然函数的期望值：

$$E(L(\theta|X,Z))=\int_Z p(X,Z|\theta)f(Z)d_Z\tag{3.9}$$

采用 EM 算法的基本思想[23][24]是对于上述的完整数据集Y，假设这些数据独立同分布于我们已知的某一个模型，如 GMM，而我们知道该模型的参数，因此可以根据该模型推出属于每个成分的各数据点的概率，然后修改每个成分的值，重复该过程直到收敛到结束条件。

假设初始参数为θ^0，多步迭代运算，每次运算都产生新的参数θ；EM 运算的每次迭代都由 Expectation 和 Maximization 两步组成：

E-step：引入辅助函数$Q(\theta,\theta^{(i-1)})$，其定义是：

$$Q(\theta,\theta^{(i-1)})=E(\log L(\theta|X,Z))=\int_Z \log L(\theta|X,Z)f(Z|,\theta^{(i-1)})d_Z\tag{3.10}$$

显然，辅助函数$Q(\theta,\theta^{(i-1)})$的值就是$\log(L(\theta|X,Z))$的期望值，并且是θ的函数，$\theta^{(i-1)}$是上一步迭代运算求得的参数值。

M-step:求解θ^*，使得$Q(\theta^*,\theta^{(i-1)})$得到极大值，即：

$$\theta^*=\theta^{(i)}=\arg\max Q(\theta,\theta^{(i-1)})\tag{3.11}$$

可以看出，随机向量Z的分布是由X和$\theta^{(i-1)}$决定的，若θ_i^*表示第i次迭代的最大似然函数值，Q_{i-1}^*表示第$i-1$次迭代的最大似然函数值，可知证明，EM 算法能够保证$\theta_i^*\geq Q_{(i-1)}^*$，并且算法是收敛的。

在高斯混合模型（GMM）里，假设完整数据为$y_i=(x_i,z_i)$，x_i为可观测变量，z_i为隐含变量，$z_i=(z_{i1}.....z_{iG})$

$$Z_{ik}=\begin{cases}1 & if\ \ x_i\ \ belongs\ \ to\ \ group\ \ k\\ 0 & otherwise\end{cases}\tag{3.12}$$

设z_i是独立同分布于G类，其概率分别为$\pi_1,.....,\pi_G$，并且由x_i给出的z_i的密度为：$\prod_{k=1}^{G}f_k(x_i|\ \theta_k)^{z_{ik}}$

完整数据的 log 似然函数为：

$$L(\theta_k,\pi_k,z_{ik}|\ x)=\sum_{i=1}^{n}\sum_{k=1}^{G}z_{ik}[\log\pi_k f_k(x_i|\ \theta_k)] \tag{3.13}$$

聚类的 EM 算法是在 E-step 和 M-step 之间迭代[17][18]。在 E-step 里，由可观测变量 x 和当前的参数估计，计算出完整数据 log 似然的条件期望值 z_{ik}，M-step 中，根据 E-step 的值，计算使得 log 似然函数值最大的参数估计一权重，均值，协方差矩阵。

下面给出了基于高斯混合模型的 EM 算法的流程图：

表 3.1：基于高斯混合模型的聚类的 EM 算法

初始化 z_{ik} (可以用一个离散的分类来表示(0-1))

Repeat

M-step:由 z_{ik} 计算最大似然参数估计值

$$n_k \leftarrow \sum_{i=1}^{n} z_{ik}$$

$$\pi_k \leftarrow n_k / n$$

$$\mu_k \leftarrow (\sum_{i=1}^{n} z_{ik} x_i) / n_k$$

$$\Sigma_k \leftarrow ((\sum_{i=1}^{n} z_{ik} (x_i - \mu_k)(x_i - \mu_k)') / n_k \text{ (取决于模型)}$$

(再计算似然函数式)

E-step:由M-step的参数计算 z_{ik}

$$z_{ik} \leftarrow (\pi_k f_k(x_i|\ \mu_k, \sum_k)) / \sum_{j=1}^{G} \pi_j f_j(x_i|\ u_j, \sum_j)$$

until 满意的收敛标准

3.3.4 实验

为了验证上述的高斯混合模型的 EM 算法，我们利用一个数据集：diabetes，用传统的 kmeans 方法和其进行比较。该数据集是包含 145 个三维的数据点。在后面的实验中还会详述。分别用三组图：(1,2)维，(1,3) 维，(2,3) 维，来表示聚类的结果。如下图：

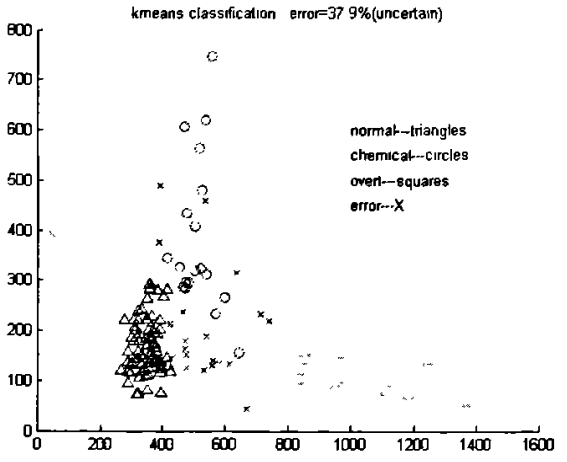

图 3.2：Kmeans 聚类

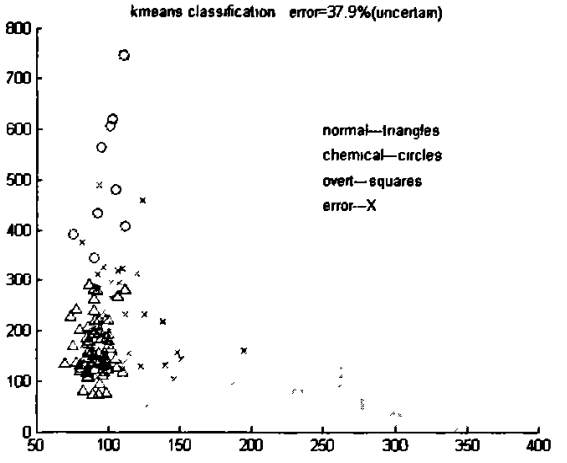

图 3.3 ： Kmeans 聚类

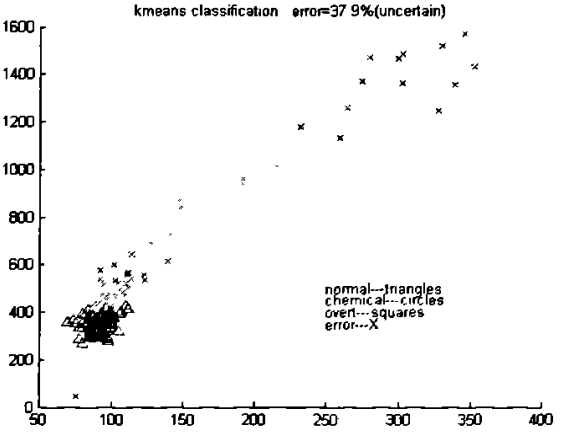

图 3.4 ： Kmeans 聚类

对于同样的该组数据，我们采用本文的高斯混合模型的 EM 方法聚类。样本识别的错误率有了明显的下降，并且多次的实验结果发现，该方法比 Kmeans 方法稳定，多次的实验结果的差别不是特别大。

具体的实验结果如下图所示：

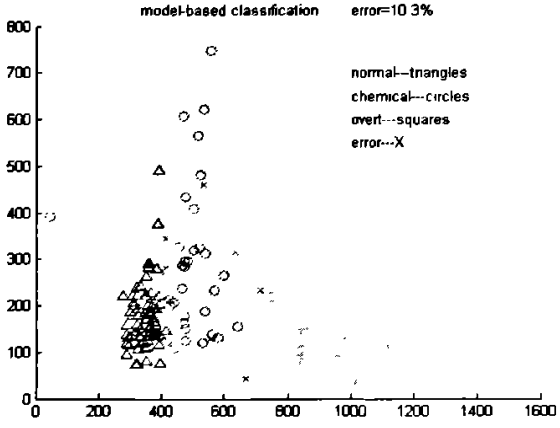

图 3.5：EM 方法聚类

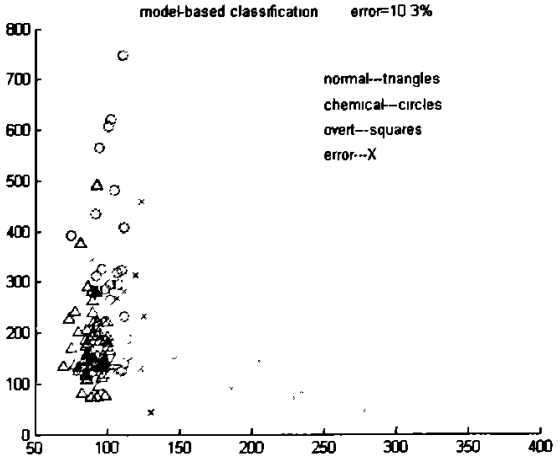

图 3.6：EM 方法聚类

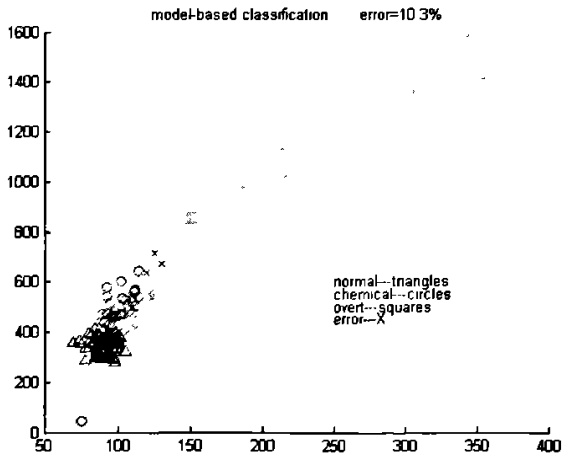

图 3.7：EM 方法聚类

3.4 初始化问题的讨论

EM 算法的迭代由设定模型的初始参数开始，把推导得到的新的模型参数用作下一次迭代过程中模型参数的初始值，迭代直到满足收敛条件后结束。通过上文的讨论和实验，我们发现，在高斯混合模型的 EM 算法中，初始值的获取对算法的聚类结果有较大的影响。换句话说，就是算法对使用何种初始化方法得到的参数比较敏感。所谓 EM 算法的初始化[19][20][21][25]就是使用某一算法，来获得 EM 迭代的初始值。在高斯混合模型聚类中，我们要得到的 EM 初始值就是参数：权重π_0，均值μ_0，协方差Σ_0，即$\psi_0=(\pi_0,\mu_0,\Sigma_0)$

3.4.1 随机初始化

在 EM 作为一种参数估计的方法提出以后，多数情况下我们都是使用随机初始化方法。即：在我们的数据集里随机的抽取 n 个点作为聚类中心，这里 n 等于聚类（类别）数。然后，对数据集里的每个数据点（除了这 n 个点以外），计算其与这n个点的每个点的欧氏距离，根据距离最短原则，把每个点放入n类中的某一类，并给同一个类里的数据点赋予相同的类标。这样，我们就可以得到一个初始的类标。于是，我们对于每一类的数据，可以计算出其权重，均值，协方差。即ψ_0。随机初始化的方法优点是过程简单，操作方便。但是因为是随机抽取的聚类中心，所以结果偏差较大。但在要求不严格的情况下，它也只是一个粗糙的分类，EM 才是整个聚类的主体。所以，随机初始化 EM 也得到了较为广泛的应用。

3.4.2 层次聚类初始化

这里我们使用的是凝集（自底向上）层次聚类[34]。即:首先将n个样本分成n类,每一类正好含有一个样本。然后将样本凝集成 n－1 类，n－2 类，直到所有的样本都凝集成我们所需要的k类为止。我们把每一次凝集的过程称为一个阶段。凝集的原则是找最相似的类，然后把它们合并。可以用距离来反映相似性，也可以根据最大似然标准来凝集（合并）类。在模型聚类的方法里，我们多采用最大似然标准作为准则。层次聚类的一个显著特点就是时间效率低，在初始数据比较庞大的情况下，执行效率不高，并且需要相当于原始空间近乎两倍的存储空间。所以，使用层次聚类的初始化受到了一定程度的限制。

3.4.3 Kmeans 初始化

Kmeans 聚类即 K 均值聚类，属于聚类分析方法中一种基本的且应用最广泛的划分算法。目标就是要找到k个均值向量，这里的k就是我们的聚类数目。基于给定的聚类目标函数（或者说是聚类效果判别准则），算法采用迭代更新的方法，每一次迭代过程都是向目标函数值减小的方向进行，最终的聚类结果使目标函数值取得极小值，达到较优的聚类效果。根据聚类结果的表达方式又可以分为硬 K-means(HCM)算法、模糊 K-means 算法（FCM）和概率 K-means 算法（PCM）。算法首先随机选取k个点作为初始聚类中心，然后计算各个数据对象到各聚类中心的距离，把数据对象归到离它最近的那个聚类中心所在的类；对调整后的新类计算新的聚类中心，如果相邻两次的聚类中心没有任何变化，说明数据对象调整结束，聚类准则函数 Jc 已经收敛。本算法的一个特点是在每次迭代中都要考察每个样本的分类是否正确，若不正确，就要调整。在全部数据调整完后，再修改聚类中心，进入下一次迭代。如果在一次迭代算法中，所有的数据对象被正确分类，则不会有调整，聚类中心也不会有任何变化，这标志着 Jc 已经收敛，至此算法结束。Kmeans 算法已得到了广泛的应用，作为对原始数据的一个粗糙的划分，它操作方便，执行效率也相对较高。但是聚类的结果不稳定，常常有较大的偏差。

3.4.4 Binning 初始化

Binning 法中文意思是装箱法，可以想象成把数据空间在各维上划分成一个个的箱子，再把数据点投射到对应的箱子里去。在统计学里，它是用于密度估计的一种方法。在这里，初始化 EM 的任务就是找到最优聚类中心。我们可以把这个问题视为密度估计的问题。根据原始数据集，最好的聚类中心可能就是概率密度函数最稠密的部分。于是，我们可以通过 binning 法来寻找概率密度函

数最稠密的部分。

所谓 binning 法，就是我们根据一定的 bin 宽，将每一维上的整个数据空间分成若干个 bin，然后把每个数据的每一维放到对应的 bin 里面，再计算每一个 bin 里所含的数据点的个数。含的点数多的则为概率密度相对大的区域。也就是聚类中心最可能存在的区域。

在得到了每一个 bin 里的数据点的个数后，我们可以作进一步的优化：

1，给每一个数据点一个向量标记，表示其各维所在的 bin 位置。

2，计算出 bin 里数据点的均值。把所有小于均值的 bin 去除考虑范围。

3，按照 bin 的数目，大致估计聚类中心的位置。

4，按照相似度标准和已估计的聚类中心，针对每一个数据，比较其相似度。

5，根据相似度重新排列数据点，相似度高的为同一类。

6，给相同的类的点以相同的类标。

我们的相似度标准是按照每一个向量标记，用相同位置且值也相同的分量的个数除以总的维数。Binning 法的一个关键点就是找到每一维上最优或者近似最优的 bin 宽。文献[22]给出了高斯分布的一元数据的最优 bin 宽。应用[22]里的推导，我们假设：每一维上的数据都近似一个高斯分布的概率密度函数。对于和高斯分布相差很大的每一维的数据，[22]里给出的 bin 宽可能只能只是一个次优值。

3.5 初始化的实验

我们采用了医疗诊断数据集（diabetes）和植物开花数据集（Iris data）来比较以上四种初始化 EM 的方法。

数据下载地址：http://www.ics.uci.edu/%7Emlearn/MLRepository.html

◆ 实验一

基于模型的聚类方法有了很广泛的实际应用。包括字符识别，组织细胞分割，纺织品瑕疵的鉴别，医疗数据分析以及大量数据的分类等等。在这里，我们通过对糖尿病诊断的数据实例，分析说明基于模型的 EM 算法。利用极大似然标准，多次迭代，直到收敛。我们用以上四种方法给 EM 赋初始值。通过比较，分析不同的初始化 EM 的方法对聚类效果的影响。该糖尿病诊断的数据集包含 145 个三维的数据点。每一维分别代 glucose,insulin,sspg 这样三个分量。该数据集又是由三类数据点混合而成，分别代表正常细胞，化学的糖尿病细胞，明显的糖尿病细胞。我们把这三维的数据点用三个二维的平面图形来表示，(1，2)维，(1，3)维和(2，3)维。选择(1，3)维作为代表进行分析。

原始的点图表示如下：

normal---triangle(三角); chemical---circle(圆圈); overt---square(方块)

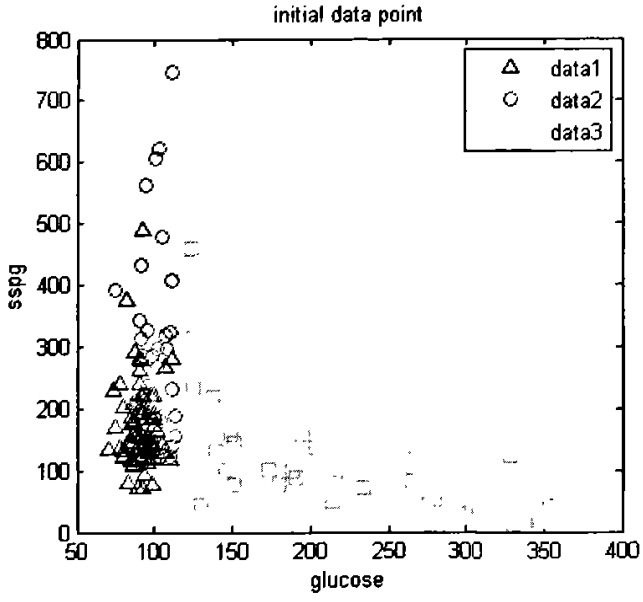

图 3.8：原始的三类数据点

使用四种不同的初始化方法的 EM 聚类结果如下图表示，‘叉’表示错误的分类。

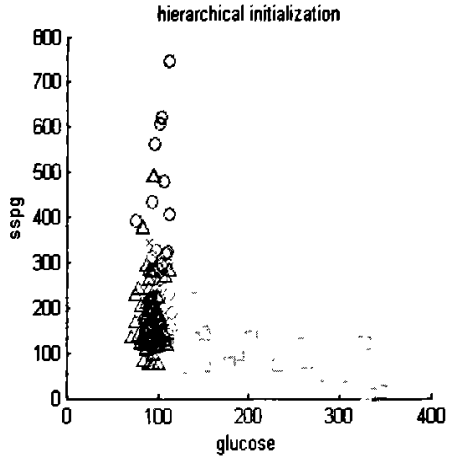

图 3.9：层次聚类初始化 EM

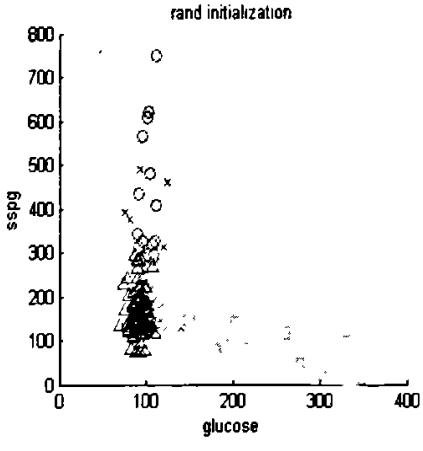

图 3.10：随机初始化 EM

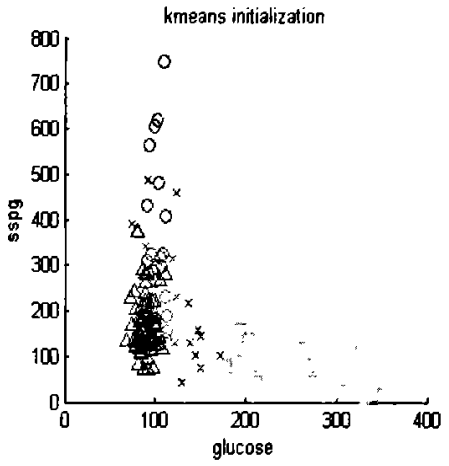

图 3.11：kmeans 初始化 EM

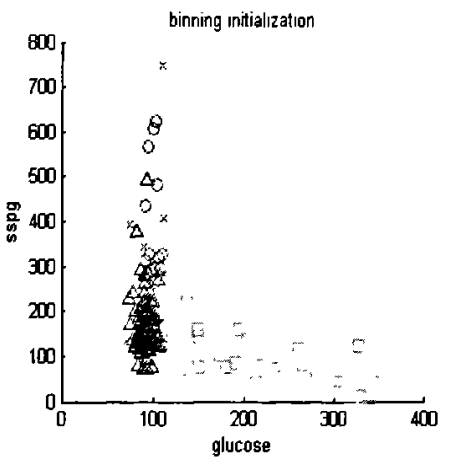

图 3.12：binning 初始化 EM

对四种初始化方法的 EM 聚类的错误率以及每一类数据正确的聚类点数，我们给出下表来直观地反映聚类结果：

表 3.2：初始化方法对 diabetes 聚类效果，错误率的比较

初始化方法	Normal	Chemical	Overt	错误率（百分比）
随机	73	25	29	12.4138
层次聚类	70	30	33	8.2759
Kmeans	75	21	21	19.3103
Binning	73	26	33	8.9655

◆ 实验二

Iris 数据集包含 150 个样本数据，分别取自三种不同的莺尾属植物 setosa. versicolor 和 virginica 的花朵样本，其中每个数据含有四个属性，即萼片长度(sepal length)萼片宽度(sepal width)、花瓣长度(petal length)和花瓣宽度(petal width)，单位为厘米。

所以，数据集分布为分为 3 类的 150 个 4 维数据向量。我们依然用 EM 算法对该数据集进行聚类。通过四种不同的初始化 EM 的算法进行比较。实验结果如下：

表 3.3：初始化方法对 Iris data 聚类效果，错误率的比较

初始化方法	Class-1	Class-2	Class-3	错误率（百分比）
随机	3	7	4	9.3333
层次聚类	1	6	3	6.6666
Kmeans	0	5	0	3.3333
Binning	1	5	2	5.3333

实验分析：

我们对两组实际数据集,运用 EM 算法对其进行聚类。使用了四种方法来初始化 EM。实验中我们得到了算法对每一类数据可正确聚类的数目和了总体的聚类的错误率。随机中心初始化操作方便，但是出于随机性，实际效果有较大的偏差，这里取的是多次情况下最好的结果。层次聚类初始化可以聚类到我们指定的类别数，但是迭代的次数较多。在数据规模较大的情况下，效率很低，即耗时和需要较大的存储空间。Kmeans 初始化也很常见，但是结果不够稳定，每次实验结果不完全一致。Binning 法相对操作方便，也达到了较好的聚类效果。

3.6 本章小结

本章我们对 EM 算法的基本思想，原理进行了介绍以后，深入研究了的高斯混合模型下的 EM 算法。先是利用已知的数据集，对二维空间里的三维数据运用传统的 Kmeans 方法和这里的 EM 方法进行聚类。由实验结果，我们可以看出 GMM 下的 EM 方法，有较好的聚类效果。接着，我们对于 EM 算法进行了较深入的研究。由于该算法对初始数值的设置十分敏感，于是，采用了几种不同的初始化方法对 EM 算法的实际聚类效果进行了研究。在比较了传统的随机初始化方法，层次聚类方法，kmeans 方法后，我们发现，用于密度估计的 binning 方法在作为一种初始化 EM 算法的方法，得到了较好的聚类效果，在操作上，也比较方便可行。

第四章 半监督的EM算法的研究

本章对半监督条件下的EM算法[26][27]进行了研究。介绍了贝叶斯学习理论的基本观点，尤其是最大后验概率。接着，详细介绍了本文了双重高斯混合模型的EM算法。在无监督学习中增加一些有标记的样本，利用已标记的样本得到初始参数，研究了半监督条件下的双重高斯混合模型的EM聚类算法。

4.1 半监督学习的聚类

半监督学习(Semi-supervised Learning)[26][27][28][29]是模式识别和机器学习中的重要研究领域。近几年随着机器学习理论在数据分析和数据挖掘的实际问题，例如网页检索和文本分类，基于生物特征的身份识别，图像检索和视频检索，医学数据处理等问题中的广泛应用，半监督学习在理论和实际应用研究中都获得了长足的发展。半监督学习研究主要关注当训练数据的部分信息缺失的情况下，如何获得具有良好性能和推广能力的学习机器，这里的信息缺失涵盖数据的类别标签缺失或者存在噪声，数据的部分特征维缺失等多种情况。半监督学习问题从样本的角度而言是利用少量标注样本和大量未标注样本进行机器学习，从概率学习角度可理解为研究如何利用训练样本的输入边缘概率$P(x)$和条件输出概率$P(y|x)$的联系设计具有良好性能的分类器。

4.2 双重高斯混合模型的基础知识

在讨论双重高斯混合模型的EM算法前，需要先了解一些相关基础知识。即：贝叶斯学习理论和最大后验概率。

4.2.1 贝叶斯学习理论的基本观点

贝叶斯学习理论[30][35][36]利用先验知识和样本数据来获得未知样本的分布性质，而概率是先验信息和样本数据信息在贝叶斯学习理论中的表现形式。贝叶斯估计研究如何获得对未知变量的分布和参数估计。

贝叶斯分析方法的特点是使用概率去表示所有形式的不确定性，用概率规则来实现学习和推理。贝叶斯学习的结果表示为随机变量的概率分布，它可以理解为我们对不同可能性的信任程度。贝叶斯学派的起点是贝叶斯的两项工作：贝叶斯定理和贝叶斯假设。贝叶斯定理将事件的先验概率与后验概率联系起来。假定随机向量x,θ的联合分布密度是$p(x,\theta)$，它们的边际密度分别为$p(x)$，

$p(\theta)$。一般情况下设x是观测向量，θ是未知参数向量，通过观测向量获得未知参数向量的估计，贝叶斯定理记作：

$$p(\theta|x)=\frac{\pi(\theta)*p(x|\theta)}{p(x)}=\frac{\pi(\theta)*p(x|\theta)}{\int\pi(\theta)*p(x|\theta)d\theta}\quad(\pi(\theta)\text{是}\theta\text{的先验分布})\quad(4.1)$$

从上式我们可以看出，对未知参数向量的估计综合了它的先验信息和样本信息，而传统的参数估计方法只从样本数据获取信息如最大似然估计。贝叶斯方法对未知参数向量估计的一般过程为：

1. 将未知参数看成是随机向量。这是贝叶斯方法与传统的参数估计方法的最大区别。
2. 根据以往对参数θ的知识，确定先验分布$\pi(\theta)$，它是贝叶斯方法容易引起争议的一步，因此受到经典统计界的攻击。
3. 计算后验分布密度，做出对未知参数的推断。

在第二步，如果没有任何以往的知识来帮助确定$\pi(\theta)$，贝叶斯提出可以采用均匀分布作为其分布，即参数在它的变化范围内，取到各个值的机会是相同的，称这个假定为贝叶斯假设。贝叶斯假设在直觉上易于被人们所接受，然而它在处理无信息先验分布，尤其是未知参数无界的情况却遇到了困难。经验贝叶斯估计 EM（Empirical Bayesian Estimator）把经典的方法和贝叶斯方法结合在一起，用经典的方法获得样本的边际密度$p(x)$，然后通过下式来确定先验分布$\pi(\theta)$：

$$p(x)=\int_{-\infty}^{+\infty}\pi(\theta)*p(x|\theta)d\theta \quad (4.2)$$

贝叶斯定理的计算学习机制是将先验分布中的期望值与样本均值按各自的精度进行加权平均，精度越高者其权值越大。在先验分布为共轭分布的前提下，可以将后验信息作为新一轮计算的先验，用贝叶斯定理与进一步得到的样本信息进行综合。多次重复这个过程后，样本信息的影响越来越显著。由于贝叶斯方法可以综合先验信息和后验信息，既可避免只使用先验信息可能带来的主观偏见和缺乏样本信息时的大量盲目搜索与计算，也可避免只使用后验信息带来的噪音影响。因此，适用于具有概率统计特征的数据采掘和知识发现问题，尤其是样本难以取得或代价昂贵的领域。合理准确地确定先验，是贝叶斯方法进行有效学习的关键问题。目前先验分布的确定依据只是一些准则，没有可操作的完整理论。在许多情况下先验分布的合理性和准确性难以评价。对于这些问题还需要进一步深入研究。

4.2.2 最大后验概率

期望最大化方法 EM 方法[37]迭代地计算最大似然估计（Maximum Likelihood Estimation MLE）和最大后验概率[31]（Maximum APosterior MAP）。它处理不完全数据分为以下几个步骤：（1）含有不完全数据的样本的缺项用该项的最大似然估计代替；（2）把第一步中的缺项值作为先验信息，计算每一缺项的最大后验概率，并根据最大后验概率计算它的理想值。（3）用理想值替换（1）中的缺项。（4）重复（1—3），直到两次相继估计的差在某一固定阀值内。EM 算法的一个缺点是易陷入局部最优。

朴素贝叶斯分类模型（Naive Bayes 或 Simple Bayesian ）假定特征向量的各分量间相对于决策变量是相对独立的，也就是说各分量独立地作用于决策变量，如图 4.1。尽管这一假定一定程度上限制了朴素贝叶斯模型的适用范围，然而在实际应用中，不仅指数级的降低了贝叶斯网络构建的复杂性，而且在违背这种假定条件下的许多领域，朴素贝叶斯也表现出相当的健壮性和高效性，它已经成功地应用到分类、聚类及模型选择等数据挖掘的任务中。目前，许多研究人员正致力于放松特征变量间独立性的限制，以使它适用于更大的范围。为引用方便，以后称该模型为 NBClassifier。

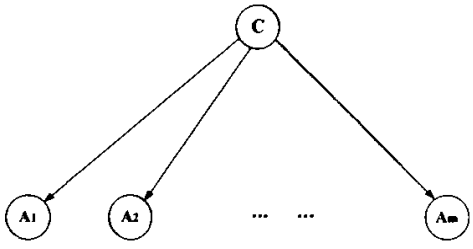

图 4.1 朴素贝叶斯分类模型

贝叶斯定理告诉我们如何通过给定的训练样本集预测未知样本的类别，它的预测依据就是取后验概率

$$p(c_i \mid x) = \frac{P(c_i) * p(x \mid c_i)}{p(x)} \tag{4.3}$$

最大的类别。这里 x 是待分类样本，$P(Y \mid X)$ 是在给定 X 的情况下 Y 的条件概率。为叙述方便，我们对一些符号作如下约定：用大写字母表示变量，C 表示类别变量，A 表示属性变量，假定共有 m 个属性变量，记 $A = < A_1, A_2, \cdots A_m >$；用小写字母表示变量的取值，记 $Val(C) = \{c_1, c_2, \cdots c_l\}$，$Val(A_i) = \{a_{11}, a_{12}, \cdots a_{ik}\}$ 分别表示类别变量和属性变量的值域，在不致混淆的情况下也用 a_i 表示 A_i 的某

个取值；用 X 表示待分样本集，$x=<a_1,a_2,\cdots a_m>$ 表示待分样本，用 T 表示训练样本集，$t_i=<a_1,a_2,\cdots a_m,c_i>$ 表示一个训练实例。由于各属性相对于类别条件独立，那么 $P(x|c_i)$ 可以分解成几个分量的积：

$$P(a_1|c_i)*P(a_2|c_i)*\cdots P(a_m|c_i) \tag{4.4}$$

从而后验概率的计算公式为：

$$P(c_i|x)=\frac{P(c_i)}{P(x)}\prod_{j=1}^{m}P(a_1|c_i) \tag{4.5}$$

4.3 基于双重高斯混合模型的 EM 学习算法

4.3.1 双重高斯 GMM 方法原理

本章的双重高斯混合模型[32][33][34]指的是在全体学习样本的概率分布中，上一重为大高斯，拟定这些样本符合高斯混合分布，并且样本的类别数就是高斯数。在每一个样本里，又都分别含有一个高斯混合模型，也就是下一层小高斯。小高斯里分样本的类别数也就是高斯数。所以，我们假设：

定义了全体学习样本的概率分布.在第一重高斯中，高斯数就是样本的类别数，设类别数是M，则第一重中需要学习的参数是 $\{\alpha_1\cdots\alpha_M\}$，即每个高斯(每个类别)的先验概率，并且满足 $\sum_{i=1}^{M}\alpha_i=1$；在第二重中，需要学习的参数是

$\{\alpha_{1|1},\cdots\alpha_{H_1|1},\cdots\alpha_{1|M},\cdots\alpha_{H_M|M}\}$ 及

$\{\mu_{11},\cdots\mu_{1H_1},\cdots\mu_{M1},\cdots\mu_{MH_M},\Sigma_{11},\cdots\Sigma_{1H_1},\cdots\Sigma_{M1},\cdots\Sigma_{MH_M}\}$，$H_i$ 表示第i个高斯包含的子高斯数；$\alpha_{j|i}$ 表示第i个高斯的条件下，第j个子高斯的条件先验概率，并且满足 $\sum_{j=1}^{H_i}\alpha_{j|i}=1,i\in\{1,2,\cdots M\}$。$\mu_{ij}$ 和 Σ_{ij} 分别表示第i个高斯的第j个子高斯的均值向量和方差矩阵。从而，双重高斯混合模型参数定义为

$\Theta=\{(\alpha_1\cdots\alpha_M),(\alpha_{1|1},\cdots\alpha_{H_1|1},\cdots\alpha_{1|M},\cdots\alpha_{H_M|M}),(\mu_{11},\cdots\mu_{1H_1},\cdots\mu_{M1},\cdots\mu_{MH_M},$

$\Sigma_{11},\cdots\Sigma_{1H_1},\cdots\Sigma_{M1},\cdots\Sigma_{MH_M})\}$

下图为双重高斯混合模型示意图，图中第1层高斯数 $M=m$，每个高斯的子高斯

数相等并且都为n

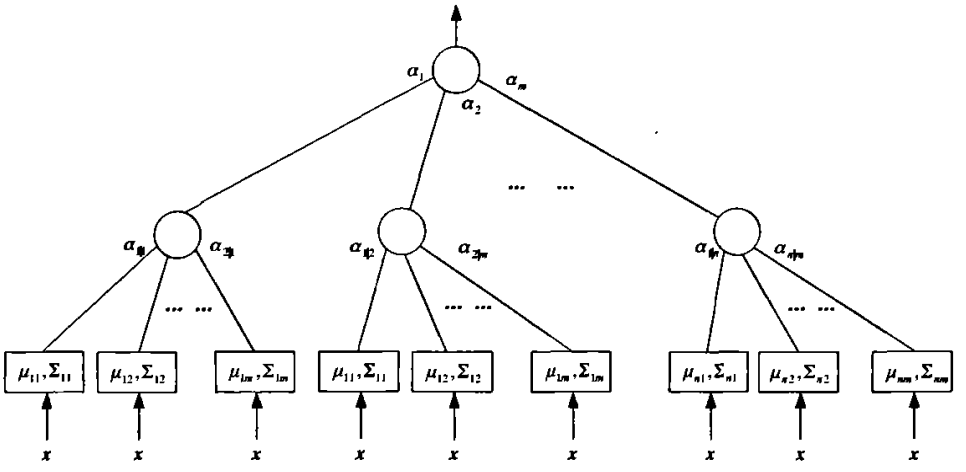

图4.2 双重高斯混合模型（$M=m, H_1=H_2=n$）

4.3.2 双重高斯混合模型的EM学习算法

给定一个有限的有类标的训练样本集，表示为：$S'=\{(X_1,y_1),.....,(X_L,y_L)\}$，其中$X_i$表示一个样本，该样本为一个高斯混合模型。$y_i$是其相应的类标。整个训练样本集$S'$共有$M$个类，所以$y_i\in\{c_1,c_2,.....,c_M\}$，每一个$X_i$可以表示为一个在$d$维空间内的由$N$个特征向量集的有$K$个成分的高斯混合模型。

$$X=<x_1,x_2,.....,x_N>, x_n\in R^d。$$

对于每个S'中的(X_i,y_i)，我们认为它们都是相互独立且同分布的。并且假设对任何一个未标记的X，通过计算最大后验概率$P(c_j|X)$，认为X属于j类，我们可以利用类条件概率密度函数$p(X|c_j)$和贝叶斯法则来计算后验概率：

$$P(c_j|X)=\frac{p(X|c_j)P(c_j)}{p(X)} \tag{4.6}$$

其中$P(c_j)$是先验概率。

$$p(X)=\sum_{j=1}^{M}p(X|c_j)P(c_j) \tag{4.7}$$

在双重高斯混合模型里，每一个特征向量$x\in R^d$的类条件概率密度可以模拟成一个高斯混合模型。即：

$$p(x|c_j)=\sum_{k=1}^{K}\partial_{jk}p(x|\mu_{jk},\Sigma_{jk}) \tag{4.8}$$

K 是第 j 类中混合成分的数目，∂_{jk} 是混合权重。$p(x|\mu_{jk},\sum_{jk})$ 是为多元高斯混合分布的概率密度函数。一个完整的高斯混合模型，即 $X=<x_1,x_2,.....,x_N>$ 的类条件概率密度可表示如下：

$$p(X|c_j)=\prod_{n=1}^{N}p(x_n|c_j)$$
$$=\prod_{n=1}^{N}(\sum_{k=1}^{K}\partial_{jk}p(x_n|\mu_{jk},\sum_{jk})) \quad (4.9)$$

于是，我们可以通过上式，计算出后验概率 $P(c_j|X)$，使得：

$$j^{*}=\arg\max_{j}P(c_j|X) \quad (4.10)$$

这样就可以给未标记的样本一个类标 j。

4.3.3 半监督学习的 EM 算法

高斯混合模型的参数 $\theta_j=\{\partial_{jk},\mu_{jk},\sum_{jk}\}$ 是使用 EM 算法来估计的。在半监督学习中，我们假设学习的样本集是由部分已标记的和大部分未标记的样本组成，即 $S=S^{l}\cup S^{u}$

$$S=\{(X_1,y_1),.....,(X_L,y_L),X_{L+1},.....,X_{L+U}\},$$

这里 $y_{L+1},.....,y_{L+U}$ 对应于这些未标记的样本 $X_{L+1},.....,X_{L+U}$。对于每个未标记的样本 X_i，我们定义表示类别数目的 M 个隐含变量 z_{ij}，$j=1,.....,M$。

$$z_{ij}=\begin{cases}1 & if \quad y_i=c_j\\ 0 & otherwise\end{cases} \quad (4.11)$$

EM 算法可以用来寻找高斯混合模型的参数 θ_j。利用已标记的样本，计算出的初始的迭代参数 θ^{0}。然后再按照如下的步骤迭代计算。

[E-step] $Set \quad \hat{z}^{(t+1)}=E[z|S;\hat{\theta}^{(t)}]$

[M-step] $Set \quad \theta^{(t+1)}=\arg\max_{\theta}P(S,z^{(t+1)}|\theta)$

在 E-step 中，根据最大后验概率，给每一个未分类的样本一个类别标签。我们可以用如下式子来计算求得 Z

$$E[z_{ij}|S;\theta]=P(y_i=c_j|S;\theta)=P(c_j|X_i;\theta_j)$$
$$=\frac{p(X_i|c_j)P(c_j)}{p(X_i)}$$
$$=\frac{[\prod_{n=1}^{N}(\sum_{k=1}^{K}\partial_{jk}p(x_n;\mu_{jk},\sum_{jk}))]*P(c_j)}{\sum_{j=1}^{M}[\prod_{n=1}^{N}(\sum_{k=1}^{K}\partial_{jk}p(x_n;\mu_{jk},\sum_{jk}))]*P(c_j)] \quad (4.12)}$$

在M-step中，根据新分配的类标和原有的类标，按照最大似然，重新计算参数 θ_j。

该算法可以用如下流程表示：

表4.1 双重高斯混合模型EM算法流程图

1，令 t=0

2，初始化，M-step：$\hat{\theta}^{(0)} = \arg\max_{\theta} P(S^l \mid \theta)$

3，E-step：令 $\hat{z}^{(t+1)} = E[z \mid S; \hat{\theta}^{(t)}]$

4，对每一个 $i = L+1, \ldots\ldots, L+U$，令：

1，$j^* = \arg\max_j \hat{z}_{ij}^{(t+1)}$

2，$z_{ij}^{(t+1)} = \begin{cases} 1 & if \quad j = j^* \\ 0 & otherwise \end{cases}$，

$j = 1, 2, \ldots\ldots, M$

5，M-step：令 $\theta^{(t+1)} = \arg\max_{\theta} P(S, z^{(t+1)} \mid \theta)$

6，t=t+1

7，循环步骤3-6，直到收敛。

8，输出 $\theta^{(t)}$

在这里，每一个 j 类的 GMM 参数 θ_j 都是使用样本集中属于第 j 类的样本估计的。

4.3.4 实验

我们使用上述方法，模拟生成符合条件的双重高斯混合模型的数据。在matlab7.0的环境下运行。实验数据为包含500个样本的样本集，每个样本都是由100个2维的且分为3类的数据组成。该样本集又是5个成分（类）拟高斯混合分布组成。我们用表 4.3 来表示我们的数据集。横向为五类数据，纵向为每一个样本由100个2维的数据。分为三类，由：个数，维数，均值，协方差形式表示如下：

表4.2 模拟实验数据

	第一类	第二类
1	30,2,[1,4]',[9,0;0,9]	30,2,[2,2]',[2,0;0,2]
2	30,2,[2,1]',[9,0;0,9]	30,2,[3,1]',[2,0;0,2]

3	40,2,[3,0]',[9,0;0,9]	40,2,[3,2]',[2,0;0,2]
	第三类	第四类
1	30,2,[4,1]',[4,0;0,4]	30,2,[1,2]',[6,0;0,6]
2	30,2,[2,5]',[4,0;0,4]	30,2,[2,4]',[6,0;0,6]
3	40,2,[6,3]',[4,0;0,4]	40,2,[6,3]',[6,0;0,6]
	第五类	
1	30,2,[4,5]',[1,0;0,1]	
2	30,2,[2,3]',[1,0;0,1]	
3	40,2,[1,1]',[1,0;0,1]	

我们依次取前 50，100，150，200，250 个样本作为已标记的样本，用本文算法观察样本的每个类的识别率，即错误聚类的个数。实验结果用下图表示：

表 4.3：半监督学习模型下的样本识别错误率

已标记样本	50	100	150	200	250
data1	19.2	17.4	14.2	10.8	8.3
data2	16.1	15.6	9.8	8.9	9.1
data3	17.1	12.1	9.7	9.1	6.2
data4	15.9	14.4	12.2	10.4	5.2
data5	17.1	14.3	12.4	7.1	6.1

将应用双重 GMM 识别的错误率随着样本数量的变化做出错误率的折线图，如图所示：

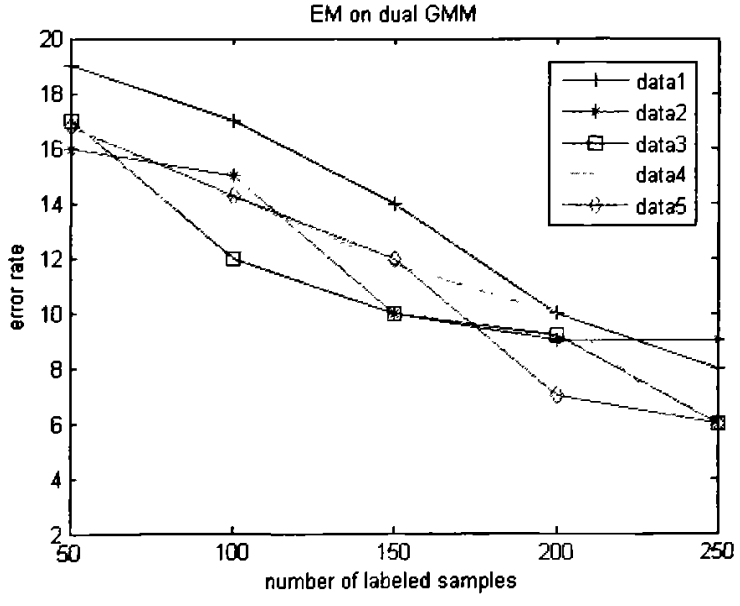

图 4.3 双重高斯混合模型的半监督学习实验结果

实验分析：

图中的每一条线都代表一个类。通过实验，我们看出，双重高斯混合模型里，在原本的无监督的学习中，加入了少量的监督信息，即有标记的样本，每一类样本的识别率也在增加。用类别识别的错误率来表示。线条呈下降趋势。

4.4 本章小结

本章对 EM 算法在双重高斯混合模型中进行了研究，通过加入有标记的样本，作为监督信息给出算法迭代的初始值。由 E－step，在整个样本集中，通过计算最大后验概率，给所有未标记的样本一个类标。然后在 M－step，按照最大似然重新计算参数θ，重复 EM 两个步骤，多次迭代，直到收敛。最后即为我们的聚类结果。

该方法有着一定的应用领域，除了用于普通的文本识别，我们还可以在语音识别里。下一章，我们将详细讨论高斯混合模型在说话人识别中的应用。

第五章 GMM在说话人识别中的应用

5.1 说话人识别

在过去20年里，语音识别[38][39]神秘不可思议的学术研究发展成为新世纪人机信息交互最时髦的界面技术之一。现代语音识别系统将信号处理、模式识别、语言学、语音学等多领域技术有机地融入统计数学方法的框架，并通过算法和计算机技术相结合的方式来实现。目前，这样的系统能够做到识别理解数十万条词汇的连续语音信号。这种现代模式识别系统除了在语音领域的应用外，可以广泛应用于信号处理和模式识别的其它领域，代表着信号与信息处理技术从曾经以解析结论或数值模拟占主导地位的方法论和系统工程向现代以大规模科学数据积累为基础，以复杂系统或过程中局部与整体交互演化的功能实现为主要目标的方法论和系统工程的革命性转变。说话人识别就是从说话人的一段语音中提取出说话人的个性特征，通过对这些个人特征的分析和识别，从而达到对说话人进行辨认或者确认的目的。说话人识别不同于语音识别，前者利用的是语音信号中说话人的个性特征，不考虑包含在语音中的字词的含义，强调的是说话人的个性；而后者的目的是识别出语音信号中的语义内容，并不考虑说话人的个性，强调的是语音的共性。说话人识别系统的结构框图，它由预处理、特征提取、模式匹配和判决等几大部分组成。除此之外，完整的说话人识别系统还应包括模型训练和判决阈值选择等部分。

5.2 说话人识别的特征提取技术

在说话人识别系统[40]中，语音信号通常被看作是短时平稳的序列，语音特征提取的第一步是语音信号的分帧处理，并利用窗函数来减少由截断处理导致的Gibbs效应；同时利用高频预加重来提升高频信息，压缩语音的动态范围。然后对每帧语音信号进行频谱处理，得到各种不同的特征参数。目前语音特征抽取有线性预测编码导出的倒频谱参数(Short-Time Cepstral Coefficient，简称LPCC)；Mel尺度式倒频谱参数(Mel-Frequency Cepstral Coefficient，简称MFCC)、感知线性预测系数(Perceptual Linear Predictive，简称PLP)以及差值倒谱等等。

其中，LPCC根据语音信号产生的全极点模型得到，MFCC和PLP则根据人耳对不同频率的语音信号的敏感程度不同，提取参数时利用Mark刻度对语音频谱进行了刻度转换，模拟人的听觉特性。Reynolds[41]的研究表明，在说话识别中，MFCC比LPCC和PLP更具有优越的识别性能。MFCC是目前应用最广的特征参

数。

5.2.1 信号特征提取-----MFCC

在语音辨识（Speech Recognition）和说话人辨识（Speaker Recognition）方面，最常用到的语音特征就是 Mel 尺度式倒频谱参数(Mel-Frequency Cepstral Coefficient，简称 MFCC)，MFCC[42][43]根据人耳对不同频率的语音信号的敏感程度不同，提取参数时利用 Mark 刻度对语音频谱进行了刻度转换，模拟人的听觉特性，因此特别适合用在语音识别中。

5.2.2 MFCC 参数的提取过程

MFCC 参数的提取过程[44][45][49][50]如下：

(1) 预处理（Pre-emphasis）：将语音信号 s(n) 通過一個高通滤波波器：

$$H(z) = 1 - a * z^{-1} \tag{5.1}$$

其中 a 介于 0.9 和 1.0 之间。若以时域的运算式来表示，预强调后的信号 $s_2(n)$ 为 $s_2(n) = s(n) - a * s(n-1)$

这个目的就是为了消除发声过程中声带和嘴唇的效应，来补偿语音信号受到发音系统所压抑的高频部分。(另一种说法则是要突现在高频的共振峰)。下面就是示范预强调所产生的效果，如图所示（注意：本小节的所有图示均为参考参考文献和网上的示范数据，用 matlab 仿真出的）

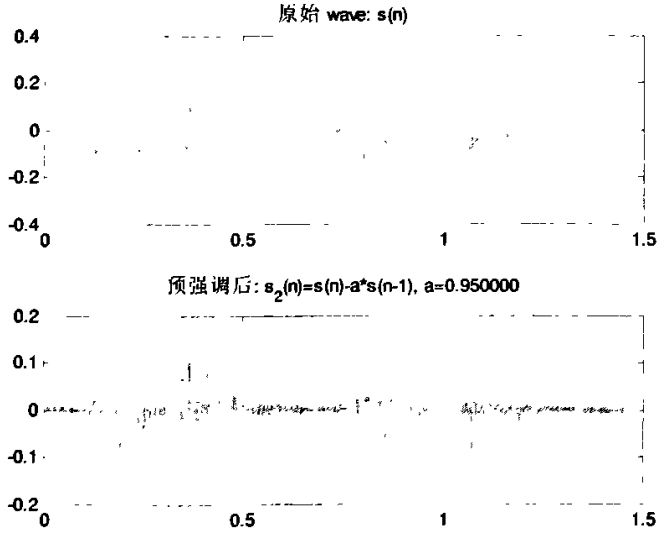

图 5.1：预处理示意图

很明显地，经过了预强调之后，声音变的比较尖锐清晰，但是音量也变小了。

(2) 音框化（Frame blocking）：先将 N 个取样点集合成一个观测单位，称为音框（Frame），通常 N 的值是 256 或 512，覆盖的时间约为 20ms～30 ms 左右。为了避免相邻两音框的变化过大，所以我们会让相邻音框之间有一段重叠区域，此重叠区域包含了 M 个取样点，通常 M 的值约是 N 的一半或 1/3。通常说话人识别所用的采样频率为 8 KHz 或 16 KHz，以 8 KHz 来说，若音框长度为 256 个取样点，则对应的时间长度是 256/8000*1000 = 32 ms。

(3) 汉明窗（Hamming window）：将每一个音框乘上汉明窗，以增加音框左端和右端的连续性。假设音框化的信号为 S(n), n = 0,...N-1。那么乘上汉明窗后为 S'(n) = S(n)*W(n)，此 W(n) 形式如下：

$$W(n,\alpha) = (1-\alpha) - \alpha\cos(2\pi n/(N-1)), \quad 0 \le n \le N-1 \tag{5.2}$$

不同的α值产生不同的汉明窗，如图所示：

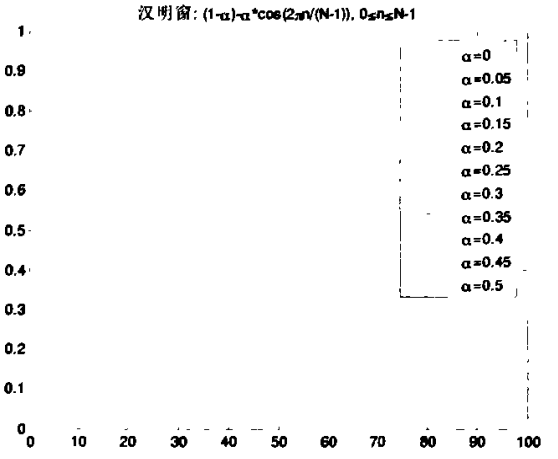

图 5.2：汉明窗示意图

一般我们都取$\alpha = 0.46$。

4) 快速傅立叶转换（Fast Fourier Transform, or FFT）：

由于 DFT(离散傅里叶变换)的运算量较大,可以采用高效的 FFT (快速傅里叶变换)把语音帧由时域变换到频域。由于信号在时域（Time domain）上的变化通常很难看出信号的特性，所以通常将它转换成频域（Frequency domain）上的能量分布来观察，不同的能量分布，就能代表不同语音的特性。所以在乘上汉明窗后，每个音框还必需再经过 FFT 以得到在频域上的能量分布。乘上汉明窗的主要目的，是要加强音框左端和右端的连续性，这是因为在进行 FFT 时，都是假设一个音框内的信号是代表一个周期性信号，如果这个周期性不存在，FFT 会为了要符合左右端不连续的变化，而产生一些不存在原信号的能量分布，造成了分析上的误差。当然，如果我们在取音框时，能够使音框中的信号就已经包含基本周期的整数倍，这时候的音框左右端就会是连续的，那就可以不需要乘上汉明窗

了。但是在实际中，由于基本周期的计算会需要额外的时间，而且也容易算错，因此我们都用汉明窗来达到类似的效果。下图显示汉明窗的效果：

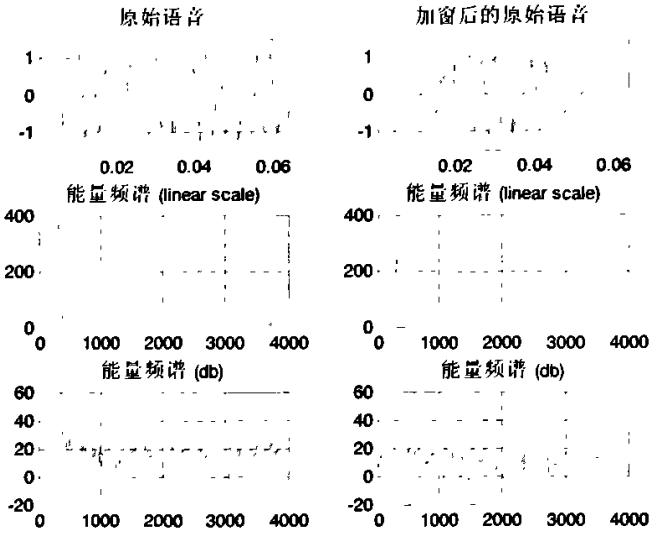

图 5.3：汉明窗效果图

在上图中，音框中的信号是一段弦波加上杂音，若不乘上汉明窗，音框的左端和右端并不连续，因此在频谱上，代表弦波的高峰比较不明显。若乘上汉明窗后，杂音在能量频谱上面的强度就会比较弱，代表弦波的高峰也相对比较突出。

(5) 三角滤波器（Triangular Filters）：将能量频谱能量乘以一组 20 个三角带通滤波器，求得每一个滤波器输出的对数能量（Log Energy）。必须注意的是：这 20 个三角带通滤波器在梅尔频率（Mel Frequency）上是平均分布的，而梅尔频率和一般频率 f 的关系式如下：

$$mel(f) = 2595 * \log_{10}(1 + f / 700) \tag{5.3}$$

梅尔频率代表一般人耳对于频率的感受度，由此也可以看出人耳对于频率 f 的感受是呈对数变换的：

- 在低频部分，人耳感受是比较敏锐
- 在高频部分，人耳的感受就会越来越粗燥

画出一般频率对于梅尔频率的特性曲线：

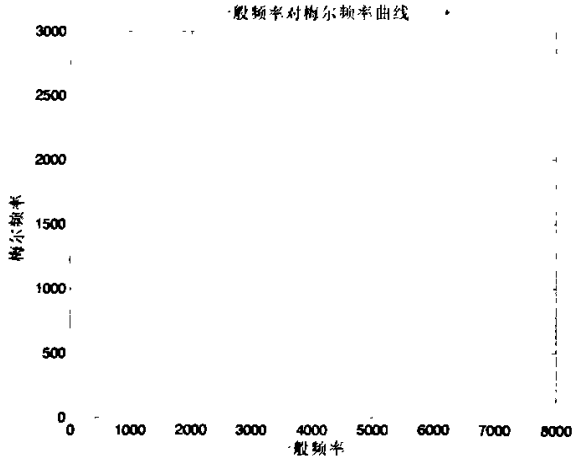

图 5.4 梅尔频率的特性曲线

三角形滤波器组如下图所示：

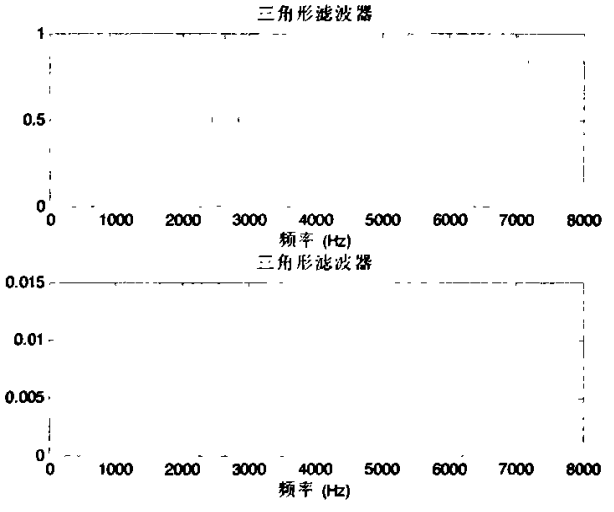

图5.5：三角形滤波器

三角形滤波器对频谱进行平滑化，並消除谐波的作用，突显原先语音的共振峰。由此可见，以 MFCC 为特征的说话人识别系统，并不会受到输入语音的音调不同而有所影响。

(6) 离散余弦变换（DCT）将上述的 20 个对数能量 E_k 带入 DCT，求出 L 阶的 Mel- scale Cepstrum 参数，这里 L 通常取 12。求出 L 阶的 Mel- scale Cepstrum 参数，这里 L 通常取 12。DCT 公式如下：

$$C_m = \sum_{k=1}^{N} \cos[m*(k-0.5)*\pi/N]*E_k, \quad m = 1,2,\cdots,L \tag{5.4}$$

其中 E_k 是由前一个步骤所算出來的三角形滤波器河频谱能量的内积值，N 是三角形滤波器的个数。由于之前作了 FFT，所以采用 DCT转换是期望能转回类

似 Time Domain 的情况来看，又称Quefrency Domain，其实也就是 Cepstrum。又因为之前采用 Mel- Frequency 来转换至梅尔频率，所以才称为Mel-scale Cepstrum。

(7) 对数能量（Log energy）：一个音框的音量（即能量），也是语音的重要特征，而且非常容易計算。因此我们通常再加上一个音框的对数能量（定义为一個音框内信号的平方和，再取以 10 为底的对数值，再乘以 10），使得每一個音框基本的语音特征就有 13 维数，包含了 1 个对数能量和 12 个倒频谱参数。（若要加入其他语音特征以测试识别率，也可以在此阶段加入，這些常用的其他语音特征，包含音高、过零率、共振峰等。）

(8) 差分倒频谱参数（Delta cepstrum）：虽然已经求出 13 个特征参数，然而在实际应用于语音识别中，我们通常会再加上差分倒频谱参数，以显示倒频谱参数对时间的变化。它的意义为倒频谱参数相对于时间的斜率，，也就是代表倒频谱参数在时间上的动态变换，公式如下：

$$\Delta C_m(t) = \left[\sum_{\tau=-M}^{M} C_m(t+\tau)\right] / \left[\sum_{\tau=-M}^{M} \tau^2\right] \tag{5.5}$$

這裡 M 的值一般是取 2 或 3。因此，如果加上差分运算，就会产生 26 维的特征向量；如果再加上差差分运算，就会产生 39 维的特征向量。一般我们在 PC 上进行的语音识别，就是使用 39 维的特征向量。

几乎所有的语音识别试验都可以证明在一般安静环境下 Mel 的优异特性，而其动态特性，比如简单的相隔四帧的一阶差分、二阶差分虽然会增加语音识别的特征维数，但对当前语音识别的识别率起着极其重要的作用。缺点：在上面介绍的 MFCC 特征分析中，Mel 滤波器是下面所示的一组滤波器，在计算出每个滤波器的能量后，通过 log 压缩和 DCT 来去掉各个频带之间的相关性，再完成倒谱的计算。假设某个语音信号含有某个噪声，该噪声的能量集中在某个频段上，在完成 DCT 的计算后，就把该噪声扩散到了所有例谱参数中去了，造成识别率急剧下降。

5.3 GMM 模型的训练和识别方法

基于高斯混合模型(GMM)的说话人识别[46][47][48]研究始于上世纪九十年代，1992 年 Dauglas A. Reynolds 在其博士论文里系统地提出了应用高斯混合模型的说话人识别方法。基于高斯混合模型的说话人识别的基本原理是对说话人集合中的每一个说话人建立一个概率模型(高斯混合模型)，该概率模型中的参数是由说话人的特征参数分布决定的，因此表征了说话人的身份。为了使处理简单，令每一个说话人的概率密度函数形式相同，所不同的只是函数中的参数，这时说话人模型则是在特定概率密度函数形式下的一组参数。研究表明，说话人的特征分布并非严格服从某一特定分布(比如高斯分布)，然而任何分布都可以由高斯分布的

加权和来逼近，这样就得到了 GMM 模型.

给定一个语音样本，说话人辨认的目的是要决定这个语音属于 N 个说话人中的哪一个。在一个封闭的说话人集合里，只需要确认该语音属于语音库中的哪一个说话人。在辨认任务中，目的是找到一个说话者i^*，其对应的模型λ_i。使得待识别语音特征矢量组X具有最大后验概率$P(\lambda_i / X)$。基于 GMM 的说话人辨认系统结构框图如图所示。

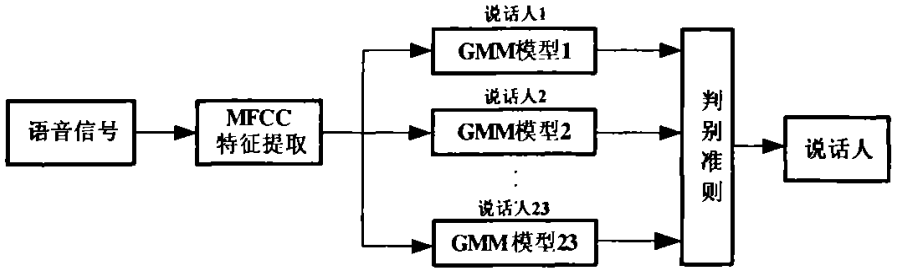

图 5.6: 基于 GMM 模型的语音辨识系统

由上图，在识别阶段，设系统有 N 个说话人，N 个候选说话人分别对应得高斯混合模型参数为：$\lambda_1, \lambda_1, \cdots, \lambda_N$。且给定待识别说话人的语音特征矢量集为$X = \{x_1, x_2, \cdots, x_T\}$，根据 Bayes 理论，则该说话人为第 n 个人的后验概率为

$$P(\lambda_n \mid X) = \frac{p(X \mid \lambda_n)P(\lambda_n)}{p(X)} = \frac{p(X \mid \lambda_n)P(\lambda_n)}{\sum_{m=1}^{N} p(X \mid \lambda_m)P(\lambda_m)} \tag{5.6}$$

其中，$P(\lambda_n)$为第 n 个说话人的先验概率：$P(X)$为所有说话人条件下，特征矢量集X的概率密度：$P(X \mid \lambda_n)$为第 n 个人条件下，特征矢量集X的概率密度：$P(\lambda_n \mid X)$是后验概率密度，表示在特征矢量集为X的条件下，说话人为第 n 个人的概率。识别判别结果可以由最大后验概率准则给出，即

$$n^* = \arg\max_{1 \le n \le N}[p(\lambda_n \mid X)] \tag{5.7}$$

其中n^*表示识别判决结果。

一般情况下，每个人说话的先验概率是未知的，可设每个人说话的概率相同，则有

$$p(\lambda_n) = \frac{1}{N} \tag{5.8}$$

其中$n = 1, 2, \cdots, N$。另外，对每个说话人，式(5.6)中的$p(X)$都相等。这样，式(5.7)等价于

$$n^* = \arg\max_{1 \le n \le N}[p(X \mid \lambda_n)] \tag{5.9}$$

这时，最大后验概率准则就转化成了最大似然准则。辨认该语音属于语音库中的

哪一个说话人可以表示为：

$$n^* = \arg\max_{1 \le n \le N}[L(X \mid \lambda_n)] \tag{5.10}$$

对于一个长度为T的测试语音时间序列 $X = (\vec{x}_1, \vec{x}_2, \cdots, \vec{x}_T)$，它的 GMM 概率可以写作：

$$P(X / \lambda_i) = \sum_{t=1}^{T} p(\vec{x}_t / \lambda) \tag{5.11}$$

5.4 说话人识别的实验及分析

设计采用高斯混合模型 GMM 的语音识别方法，使用 MFCC 倒谱系数作为特征矢量集。本实验运行在 PC 机上的 WindowsXP 操作平台上。PC 机的主频为 P4 2.93G，内存为 512M。编程主要使用 Matlab7。

5.4.1 实验用的语音库

目前国内尚没有应用于说话人识别的标准库，而国外的语音识别的标准库又不是免费和公开的，所以本次实验采用台湾清华大学的学者张智星录制并公开的“2004-义隆电子厂训”语音数据集。该数据集下载网址为 http://neural.cs.nthu.edu.tw/jang2/dataSet/childSong4public/QBSH-corpus/waveFile/ 。该数据集中参加录音人数为 23 人(全是男性)，录入的语言全部转化为 wav 格式的语音文件。录音环境比较安静，说话人离麦克风也比较近，音质比较清晰。每个人的录音段数各不相等，但是每段语音的录音时间都相等，为 8 秒。实验中，分别采用了每个说话人的五段语音和十段语音，每次实验中均用此单数句用来训练 GMM，双数句来测试，即识别说话人。对个别人的录音段数小于十段的，我们依选用了张智星的“2005-音讯与音乐处理专题”的语音数据集中的部分语音者作了替换，语音数据完全符合实验条件。

5.4.2 实验的基本过程和具体参数

说话人语音识别分为训练阶段和识别阶段，在训练阶段系统对输入的训练语料进行预处理，提取特征参数，根据特征参数用 EM 算法为每个说话人建立 GMM 模型。在识别阶段，对测试语音提取特征参数，与每个人说话人的 GMM 模型进行对比而得出识别结果。其中训练和识别阶段的特征提取算法是一样的。

实验所用的主要参数如下表：

表 5.1：实验的主要参数

预加重	$1\text{-}0.97z^{-1}$

采样率	16000Hz
量化数	8bit
加窗	Hamming Window
帧长	32ms （512 个采样点）
帧移	11ms （171 个采样点）
特征向量	12 维 MFCC

5.4.3 实验结果和分析

（1）GMM 的阶数对识别性能的影响

本节主要研究 GMM 的阶数对识别性能的影响，从而确定最佳的 GMM 混合数。实验中所用的阶数分别为 2，4，8，16，32，64，128。训练语音长度为 8 秒，测试语音长度为 8 秒，识别人数为 23 人，采用 MFCC，维数为 12，即原始的 MFCC 特征取前 12 维特征分量作为最终的特征向量。分别在 3 次，5 次以及 10 次迭代下，观察 GMM 的阶数对系统识别性能的影响。实验结果如下图所示：

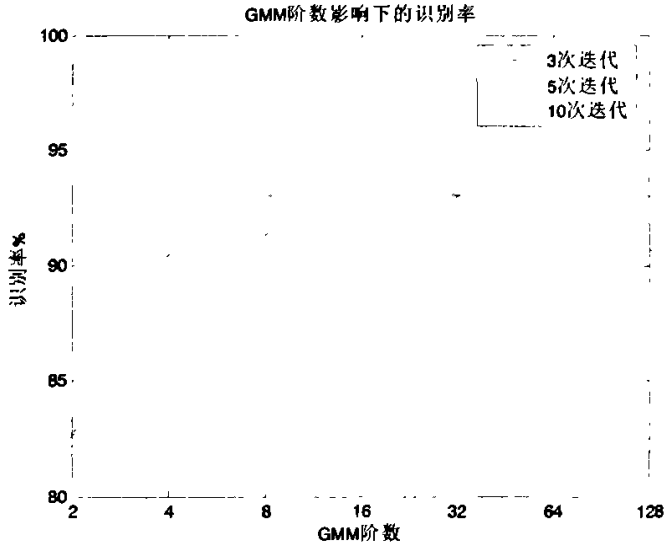

图 5.7：GMM 阶数影响下的识别率

从实验结果可以看出，对每一次迭代来说，当迭代次数不变时， GMM 的阶数从 2 增加到 16 时，识别率逐渐上升，然后，随着阶数的增加，识别性能逐渐下降。造成识别率下降的原因主要是当 GMM 阶数较大时，模型比较复杂，因此估计模型参数所需的数据量较大，而此次实验所用的训练语音时间仅为 8 秒。数据量过小，从而造成了为每一个说话人所建立的 GMM 模型不够精确，最终导致了识别性能的严重下降。

经过上面的分析可知，对于 GMM 阶数的选择，当阶数较小时，识别性能较差，而当阶数较大时，又要求有较多的训练数据，即要保证有较长的训练时间。

如果训练时间达不到一定的长度，由于模型不能精确建立，反而会使识别性能显著下降。另一方面，显著增加 GMM 的阶数使得模型更为复杂，运算量增大。因此在实际应用中，应考虑多方面的因素，譬如能够获取的语音长度，运算量的要求，整个系统说话人的数量等等。由此可见，GMM 阶数的选择对于整个系统的性能十分重要，在本论文的实验中，GMM 的阶数选取为 16 时识别率相应较高。

（2）EM 迭代次数对识别性能的影响

在应用对角协方差矩阵的 GMM 模型训练过程中，利用 EM 算法求取最大似然估计，EM 算法在迭代中去估计一个新的模型，并利用新的模型对于下一次重复运算来说成为初始模型，该过程反复执行直到达到收敛门限。而在迭代过程中，我们发现 EM 迭代的次数对于识别率的也有着一定的影响。在第一次实验中，识别人数为 23，每人的 5 段语音作为训练语音，每段的时间为 8 秒。先采用与文本有关的训练方式，即训练语音与测试（识别）语音一致，然后采用与无本无关的训练方式，即训练语音与测试（识别）语音不一致。改变 GMM 的阶数和 EM 算法的迭代次数，实验结果如下：

与文本有关的说话人识别：

表：5.2　5 句训练下的说话人识别

GMM阶数	2	4	8	16	32	64	128
0 次迭代	4.35	11.59	14.49	23.19	24.64	62.32	73.91
1 次迭代	95.65	94.20	97.10	100	100	100	100
2 次迭代	94.20	100	100	100	100	100	100
3 次迭代	94.20	100	100	100	100	100	100
5 次迭代	95.65	98.55	100	100	100	100	100
10 次迭代	95.65	98.55	100	100	100	100	100
50 次迭代	95.65	98.55	100	100	100	100	100
100 次迭代	95.65	98.55	100	100	100	100	100

与文本无关的说话人识别：

表：5.3　5 句训练下的说话人识别

GMM阶数	2	4	8	16	32	64	128
0 次迭代	4.35	13.04	15.22	23.91	26.09	47.83	56.52
1 次迭代	73.91	76.09	84.78	86.96	86.96	86.96	89.13
2 次迭代	76.09	82.61	89.13	89.13	89.13	91.30	91.30
3 次迭代	78.26	84.78	89.13	89.13	89.13	91.30	89.30
5 次迭代	82.61	84.78	86.96	86.96	86.96	80.43	73.91
10 次迭代	82.61	86.96	86.96	86.96	86.96	80.43	73.91
50 次迭代	80.43	86.96	91.30	86.96	89.13	80.43	71.74
100 次迭代	80.43	86.96	89.13	86.96	89.13	80.43	71.74

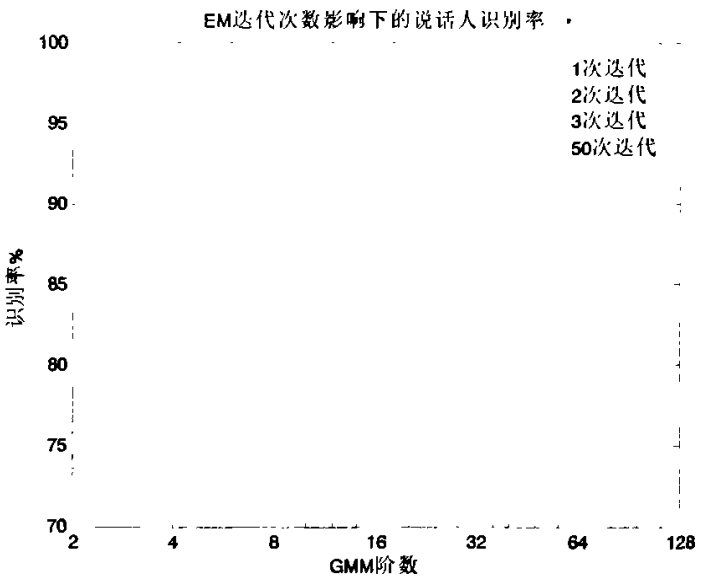

图 5.8：EM 迭代次数影响下所对应 GMM 阶数的识别率

与文本有关的说话人识别：

表 5.4：10 句训练下的说话人识别（与文本有关）

GMM阶数	2	4	8	16	32	64	128
0 次迭代	4.35	12.17	13.91	16.52 .	30.43	52.17	66.96
1 次迭代	91.30	91.30	92.17	98.26	99.13	99.13	99.13
2 次迭代	95.65	96.52	99.13	100.00 .	100	100.00	100.00
3 次迭代	95.65	98.26	99.13	100.00 .	100.00	100.00	100.00
5 次迭代	93.04	99.13 .	99.13	100.00 .	100.00	100.00	100.00
10 次迭代	95.65	99.13	99.13	100.00 .	100.00 .	100.00	100.00
50 次迭代	95.65	99.13	99.13	100.00 .	100.00 .	100.00	100.00
100 次迭代	95.65	99.13 .	99.13	100.00	100.00	100.00	100.00

与文本无关的说话人识别：

表 5.5：10 句训练下的说话人识别（与文本有关）

GMM阶数	2	4	8	16	32	64	128
0 次迭代	4.35	10.43	13.04	15.65	28.70	44.35	51.30
1 次迭代	80.87	81.74	80.87	84.35 .	89.57	93.04	93.91
2 次迭代	84.35	86.09	87.83 .	91.30	93.04	93.91 .	93.91
3 次迭代	82.61	90.43	91.30	94.78	93.04	93.91 .	93.04
5 次迭代	82.61	91.30 .	93.04 .	93.91 .	91.30	92.17	90.43
10 次迭代	82.61	91.30	93.04	93.04	92.17	92.17	90.43
50 次迭代	82.61	91.30	93.04	93.04	93.04	92.17	91.30
100 次迭代	82.61	91.30	93.04	92.17	93.04	92.17	90.43

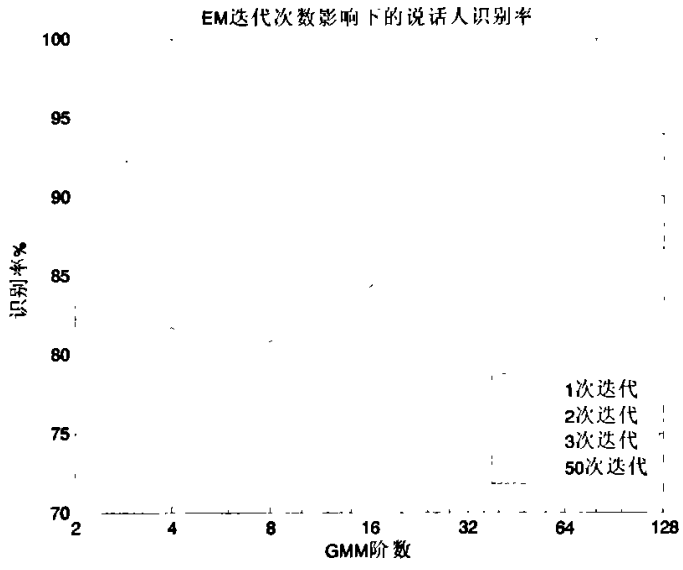

图 5.9：EM 迭代次数影响下所对应 GMM 阶数的识别率

从以上实验结果可以看出以下几点：

1，与无本有关的识别率要远远高于与无本无关的识别率。这是因为在与无本有关的识别中，每个人的测试的语音和训练的语音完全一致，即每个人的语音模型可以被精确的建立，所以，一般达到了较高的识别率。而与文本无关的识别中，测试的语音不同于训练的语音，所以相对识别率较低。

2，10 句语音的训练识别率高于 5 句语音识别率。增加的训练时间，系统可以得到从更多的语音中建立个人的语音模型，增加了信息量，即提高了相对识别率。

3，随着 EM 迭代次数的增加，在阶数一定的情况下，系统得到的个人相对率有所增加。但是，当阶数达到 64 时，迭代次数对识别率的影响的变化不再明显。因为在有些人的语音模型中，迭代到一定次数后，已经达到了收敛条件，不再迭代，所以改变迭代次数没有影响。

5.5 本章小结

本章基于 EM 算法的高斯混合模型对说话人识别技术应用的进行研究，讨论了语音信号特征提取问题，详细介绍了 MFCC 参数的特征提取过程。采用底层 MFCC 参数作为其特征描述信息，MFCC 特征在语音识别、音频分类和检索研究领域应用十分广泛，应用高斯混合模型(GMM)的说话人识别方法。实验证明，在 GMM 的说话人识别系统中，GMM 阶数的选择对于整个系统的性能十分重要。在 EM 算法求最大似然值的时候，EM 算法的迭代次数对系统的识别率也有着一定的影响。在未来的工作中，我们会研究 GMM 模型在其他语音特征提取的中的识别性能。

第六章 总结与展望

聚类就是根据事物之间的相似性把事物聚集成不同类别的一种技术，得到的聚类结果中同类之间相似度较高，而不同类之间的相似度较低。这样，聚类技术就可以把大数据集合中相似度较高的对象聚集在一起，而把相似度较低的对象区分开来。从而使得获得的聚类结果与人们的判断相一致。

聚类是数据挖掘中的一种重要技术，是分析数据并从中发现有用信息的一种有效手段。基“物以类聚”的朴素思想，它将数据对象分组成为若干个类或簇，使得在同一个簇中的对象之间具有较高的相似度，而不同簇中的对象差别很大，通过聚类，人们能够识别密集和稀疏的区域，发现全局的分布模式以及数据属性之间有趣的相互关系。聚类分析在客户分类、基因识别、WWW 文本分类、空间数据处理、卫星照片分析、医疗图象自动检测等领域有着广泛的应用，而其本身的研究也是一个蓬勃发展的领域，数据挖掘、统计学、机器学习、空间数据库技术、生物学和市场学的发展推动着聚类分析研究的进展，使它已成为数据挖掘研究中的一个热点。

本文的前一部分简要介绍了聚类的基础知识，论文研究的相关知识。在此基础上，我们深入讨论了 EM 算法和高斯混合模型的知识。并着重从一下几个方面来展开我的研究工作：

文章的第三部分，讨论了高斯混合模型的 EM 算法，实现了该算法，并和 Kmeans 算法作了比较。然后，深入 EM 算法，讨论了其初始化的方法。比较了随机初始化，层次聚类初始化，Kmeans 初始化。并提出用于密度估计的 bin 法，作为初始化 EM 算法的一种尝试。通过实验，我们发现，bin 法在聚类结果和实验的可操作性上都有较好的表现。

文章的第四部分，在讨论了本文提出的双重高斯混合模型的 EM 算法。算法可作为半监督学习的一种探讨，在实验中加入了有标记的样本，并观察其个数对整改实验聚类效果的影响。

文章的最后一部分研究了基于 MFCC 系数和高斯混合模型（GMM）的说话人识别系统，并通过实验研究了 EM 算法迭代次数和 GMM 模型阶数对识别性能的影响。

致谢

时间如梭，岁月如流，在攻读硕士学位期间，我得到了导师王士同教授的悉心指导，王教授无论是学业上还是生活上都给予了我无微不至的关怀，处处为我们着想，教导我们先学做人，再学做学问。本课题从选题，调研，开题，直至每一篇论文的写作完成是在王士同教授的细心指导下完成的。导师在学术上敏锐的观察力，严谨求实的治学态度，忘我的工作精神，以及对科学研究和教育事业高度热情和无私奉献的态度将在我今后的学习，生活，工作中产生深刻的影响，时刻激励着我不断努力攀登人生新的高峰。在此，谨向王士同教授致以我最诚挚的谢意和敬意!

其次，要感谢我实验室一起度过三年时光的同学，他们在生活和学习上都给了我很大的帮助。最后，感谢我的家人，没有他们的关心和支持，就没有我现在的成绩，我会把一直带着你们的支持，永不停止前进的步伐。

参考文献

[1] 边肇祺 张学工. 模式识别[M]. 清华大学出版社 2000.

[2] Andrew R.Webb. 统计模式识别（第二版）[M]. 电子工业出版社 2004

[3] Richard O. Duda, Peter E. Hart, David G.Stork. 模式分类[M]. 机械工业出版社 2003.

[4] Diebolt, J.and Robert, C.P. Estimation of finite mixture distributions through Bayesian sampling[J]. Journal of the Royal Statistical Society, series B,56,363-375

[5] Stephens, M. Dealing with label switching in mixture models[J]. J.R.Statist Soc.B.62,795-809,2000

[6] Jiawei Han, Micheline Kamber 著, 范明，孟小峰译,数据挖掘概念与技术[M],机械工业出版社,223-262，2001

[7] Cluster Analysis Applied to a Study of Race Mixture in Human Population[J]s, Classification and clustering, Academic Press.

[8] Jiawei Han, Micheline Kamber. 数据挖掘概念与技术[M]。机械工业出版社，2001，8

[9] Mehmed Kantardzic. 数据挖掘——概念，模型，方法和计算[M]。清华大学出版社，2003，8

[10] J. Larsen, A. Szymkowiak, and L.K. Hansen, Probabilistic Hierarchical Clustering with labeled and Unlabeled Data, invited submission for Int. Journal of Knowledge Based Intelligent Engineering Systems, 2001. http://citeseer.nj.nec.com/larsen01probabilistic.html

[11] Banfield, J. D. and Raftery, A. E. Model-based Gaussian and non-Gaussian clustering. *Biometrics*[M], 1993,49, 803–821

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

[13] Pilla, R. S. and Lindsay, B. G. Alternative EM methods for Nonparametric Finite Mixture Models[J], Biometrika, 2001, 88, 535-550.

[14] Fraley, C. and A. E. Raftery. Model-based clustering, discriminant analysis and density estimation[J]. Journal of the American Statistical Association 97,2002, 611–631

[15] J.MacQueen, Some methods for classification and analysis of multivariate observations. In proceedings of the Fifth Berkeley symposium on mathematical statistics and probability. [J]Vol1, statistics, L.M. Le Cam and J.Neyman. University of Californis Press, 1967

[16] Fraley, C. Algorithms for model-based Gaussian hierarchical clustering. [J] *SIAM J. Sci Comput*1999., 20, 270–281

[17] Kaufman L,Rousceuw P. Finding groups in data: An introducation to cluster analysis[J], New York, John Wiley and Sons, 1990

[18] McLachlan, G.J. and Krishnan, T. The EM Algorithm and Extensions[J]. John wiley and Sons, NewYork,1997

[19] C.F. Jeff WU. On the convergence properties of the EM algorithm[J]. The Annals of Statistics,1983, 11(1):95-103

[20] Christophe Biernacki, Initializing EM Using the Properties of its Trajectories in Gaussian Mixtures[J],2005

[21] Biernacki,C,Celeux,G,and Govaert,G.Choosing Starting Values for the EM Algorithm for Getting the Highest Likelihood in Multivariate Gaussian Mixture Models[J]. Computational

Statistics and Data analysis,2002

[22] D.W.Scott. On optimal and data-based histograms. Volume[J] 1979,66, 605-610

[23] Jeff A. Bilmes and Computer Science Division. A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models[J]. 1998

[24] Fraley, C. and A. E. Raftery. Model-based clustering, discriminant analysis and density estimation[J]. Journal of the American Statistical Association 97,(2002), 611–631

[25] Patricia McKenzie, Michael Alder. Initializing the EM Algorithm for use in Gaussian Mixture Modelling[J]

[26] Martin Szummer and Tommi Jaakkola. Kernel expansions with unlabeled examples. In Advances in Neural Information Processing Systems[J] (NIPS) [NIP01], 2001. 626~632

[27] Kamal Nigam, Andrew Mccallum, Sebastian Thrun, Tom Mitchell, Text Classification from Labeled and Unlabeled Documents using EM, Machine Learning, 2000, 39: 103~134

[28] B Sugato. Semi-supervised clustering by seeding. [J]. The 19th Int'l Conf on Machine Learning, Sydney,2002

[29] Cheeseman, P., and Stutz, J.. Bayesian Classification (AutoClass): Theory and Results.Knowledge Discovery in Data Bases II[M]. Menlo Park, Calif.: AAAI Press / The MIT Press.1995

[30] Fraley, C.Algorithms for model-based Gaussian hierarchical clustering. [J] SIAM Journal on Scienti_c Computing,1998.20, 270–281

[31] A. McCallum, R. Rosenfeld, T. Mitchell, and A. Ng. Improving text classification byshrinkage in a hierarchy of classes[J]. In Proc. 15th Intl. Conf. on Machine Learning (ICML) [ICML98], 1998. 359~367

[32] Liu, B., Lee, W.S., Yu, P.S. and Li, X., Partially Supervised Classification of Text Documents[J], Proc. 19th Intl. Conf. on Machine Learning, Sydney, Australia, July2002.387~394

[33] Adrian Corduneanu and Tommi Jaakkola, Stable Mixing of Complete and Incomplete Information[J], AI Memo 2001-030, November 2001.

[34] Ion Muslea, Steven Minton, Craig A.Knoblock, Active + Semi-Supervised Learning =Robust Multi-View Learning[J], ICML2002, 2002.

[35] A.Blum and S. Chawla. Learning from Labeled and Unlabeled Data using Graph Mincuts. [J].ICML, 2001.

[36] Doughtergy, J. etc. Supervised and unsupervised discretization of continuous features.Proceedings of Twelfth International Conference on Machine Learning[J], Tahoe City, CA.Morgan Kaufmann, 1995.192-202.

[37] Avrim Blum and Tom Mitchell. Combining Labeled and Unlabeled Data with Co-Training.[J] Proceedings of the 11th Annual Conference on Computational Learning Theory,1998.92-100

[38] 拉宾纳 L.R. 语音识别的基本原理[M]，清华大学出版社，2002

[39] 胡航．语音信号处理[M]．哈尔滨：哈尔滨工业大学出版社，2000

[40] K. Markov and S. Nakagawa. Text-independent speaker recognition system using frame Level likelihood processing[J] .Technical Report of IEICE,SP96-17,p37-44,1996

[41] D. A. Reynolds, and R. C. Rose, Robust Text -independent Speaker Identification Using Gaussian Mixture Speaker Models[J], IEEE Transactions on Speech and Audio Processing,

vol. 3, no. 1, Jan. 1995, 72-83.

[42] Tomi Kinnunen, "Spectral Features for Automatic Text-independent Speaker Recognition," of Joensuu, Dec. 2003. ftp://cs.joensuu.fi/pub/PhLic/2004_PhLic_Kinnunen_Tomi.pdf

[43] Minh N. Do. An Automatic Speaker Recognition System. Digital Signal Processing Mini-Project[J]. Audio Visuat Communications Laboratory. Swiss Federal Institute of Technology, Lausanne, Switzerland

[44] 武健、郑方等.基于音调的特征提取在非特定人语音识别中的应用[J].第三届全国计算机智能接口与智能应用学术会议(NCCIIIA): 93-97, 1997-8-16

[45] 郭春霞.裘雪红.基于 MFCC 的说话人识别系统.电子科技[J].2005. Vo1:11

[46] 赵力．语音信号处理[M]．北京：机械工业出版社，2003

[47] Bing Xiang and Toby Berger. Efficient text-independent speaker verification with structural gaussian mixture models and neural network[J]. IEEE Transaction on Speech and Audio Processing.2003, 11(5):447–456

[48] Zhenyu Xiong, Thomas Fang Zheng, Zhanjiang Song, and Wenhu Wu. Combining selection tree with observation reordering pruning for efficient speaker identification using gmm-ubm[J].International Conference on Acoustics, Speech and Signal Processing. 2005, 625–628.

[49] Douglas A. Reynolds. Experimental evaluation of features for robust speaker identification. [J] IEEE Transactions on Speech and Audio Processing. October 1994, 2(4):639–644.

[50] 张万里、刘桥.Mel 频率倒谱系数提取及其在声纹识别中的作用[J].贵州大学学报.Vol 22.No.2 .2005-5

发表论文

[1] 高斯混合模型聚类中 EM 算法及其初始化问题的研究. 微计算机信息.2006 12.第一作者

[2] 基于双重高斯混合模型的 EM 算法的聚类问题研究. 计算机仿真.2007 10. 第一作者

基于EM算法的模型聚类的研究及应用

作者： 岳佳
学位授予单位： 江南大学
被引用次数： 1次

本文读者也读过(4条)

1. 史鹏飞 基于改进EM算法的混合模型参数估计及聚类分析[学位论文]2009
2. 房彦兵 α-EM算法及其若干应用[学位论文]2004
3. 连军艳 EM算法及其改进在混合模型参数估计中的应用研究[学位论文]2006
4. 岳佳.王士同.YUE Jia.WANG Shi-tong 双重高斯混合模型的EM算法的聚类问题研究[期刊论文]-计算机仿真 2007,24(11)

引证文献(1条)

1. 李广水.李杨.马青霞.宋丁全 基于频繁集的图像特征抽取[期刊论文]-计算机工程与应用 2010(20)