西南交通大学

硕士学位论文

数据挖掘聚类算法研究

姓名：张昭涛

申请学位级别：硕士

专业：计算机应用技术

指导教师：杨燕

20050401

摘要

近年来，数据挖掘获得了快速发展，这是快速增长的数据量和日益贫乏的信息量之间矛盾运动的必然结果。国内主流的网站评比的未来十大热门技术中，数据挖掘占了一席，而且现今世界几大超级公司也早早地投入数据挖掘的研究，这其中包括 IBM、MicroSoft 等。数据挖掘技术集数理理论、专家系统、人工智能、神经网络、图形图象设计等多门学科于一身，其发展速度必将大大影响全球信息化的进程，对其进行系统、深入、全面、详尽地研究是信息化发展的客观需要。本文对数据挖掘技术，尤其是聚类分析进行了较为系统地分析和研究，提出了一些改进的算法，主要包括以下一些内容：

数据挖掘技术的概述。对数据挖掘技术的产生进行了简要的回顾，对数据挖掘的发现模式和常使用的技术进行了详细地分类、归纳和总结。对数据挖掘技术的应用进行了归纳，为本文的全面展开奠定了基础。

聚类分析的概述。聚类分析是数据挖掘的一个重要的研究方向，是一种无监督学习的方式，在许多方面发挥着重要的作用。对聚类分析的定义、使用的数据类型和主要的算法等进行了简要的介绍。

蚁群算法的概述。群体智能是模仿自然界昆虫行为的一个研究领域，同样也在许多领域取得了较为突出的成绩，而且有了一定的发展。蚁群算法是群体智能的一个典型代表算法，而且应用面比较广。对基于蚂蚁寻路和蚂蚁聚类的算法分别进行了简要介绍。

基于阈值的 T-Value 算法及蚁群聚类组合算法的研究。在研究了基本蚁群算法的基础上，结合蚂蚁寻路和觅食的习性和聚类的思想，提出一些想法和改进。首先把蚂蚁觅食原理结合 k-means 提出了一种基于信息素的 k-means 改进算法；根据密度聚类的思想，提出一种基于阈值的算法-T-Value 算法，同时引入 ε 邻域到 T-Value 中，结合基于信息素的 k-means 算法提出了一种 T-Value 聚类组合算法；结合 LF 算法和基于信息素的 k-means 算法，提出另一种聚类组合算法-蚁群聚类组合算法。最后对各种算法进行数据测试和性能分析，并把蚁群聚类组合算法用于移动客户的消费行为分析。

关键词：数据挖掘；聚类分析；群体智能；蚁群算法

Abstract

In recent years, many people in information industry attach more important to the data mining technique that gained rapid progress, which is attributed to the necessary consequence of the conflicting movement between the rapid increasing data and the poor information day by day. In some domestic primary website, data mining technique was chosen one of the most popular technologies in the future.In this days, some super companies such as IBM, Microsoft have attended the research of data mining technique. This dissertation systematically, deeply, roundly and detailedly studies and analyses the data mining technique, especially the one for clustering analysis. The main contents are listed as follows:

Description in brief of the data mining technique. The appearance of the data mining technique is reviewed in brief first. Based on the basic concepts of data mining, this dissertation classifies and summarizes the objects of data mining, the findable patterns and the common techniques in detail. In succession, the dissertation summarizes, analyses and studies the current status of the data mining technique in our native country and overseas widely and roundly and then summarizes and discusses its developmental trends and hot research fields. All of the above become the basis for this dissertation.

Description in brief of clustering analysis. As one of the most important domain of data mining, clustering analysis is a non-supervised learning method, exerted important effect in many aspects. The definition, data type and primary algorithms are briefly introduced.

Description in brief of ant colony algorithm. Swarm intelligence is a research domain which provided by simulating insects' behavior. Ant colony algorithm is a typical representative algorithm of Swarm Intelligence and has been widely applied. Two types of algorithm are briefly summarized. One is ant Routing algorithm, another is ant clustering algorithm.

Research of T-Value algorithm based on threshold and clustering algorithm combination. Firstly, an algorithm of k-means based on pheromone is presented. Then T-Value algorithm enlightened by DBSCAN algorithm is presented and a threshhod of ε is added into T-Value to improve the effect of this algorithm. Then a combination algorithm of T-Value and kmeans algorithm based on pheromone is

presented. A combination algorithm of improved LF algorithms and k-means algorithm based on pheromone is introduced detailedly, and is tested by some datasets. At last, this combination algorithm is used to analyse the mobile customer spending behavior.

Key Words: data mining; clustering analysis; swarm intelligence; ant colony algorithm

第 1 章 绪论

1.1 引言

近十几年，科学技术的飞速发展带动着经济和社会都取得了极大的进步。在各个领域产生了大量的数据，如人类对太空的探索，银行每天的巨额交易数据。显然在这些数据中存在着丰富的信息，如何处理这些数据以从中得到有益的信息，人们进行了有益的探索。计算机技术、网络技术和信息技术的迅速发展，人们生产和搜集数据的能力也得到了大幅度提高，使得数据处理成为可能，同样也推动了数据库技术的极大发展，但是面对不断增加如潮水般的数据，人们不再满足于数据库的查询功能，提出了深层次问题：能不能从数据中提取信息或者知识为决策服务，就数据库技术而言已经显得无能为力了。同样，传统的统计技术也面临着极大的挑战。这就急需有新的方法来处理这些海量般的数据。

于是，人们结合统计学、数据库、机器学习等技术，提出数据挖掘来解决这一难题。数据挖掘技术应运而生，并显示出前所未有的强大生命力，并且逐渐成为研究的热点，吸引了很多人进行研究。而作为数据挖掘技术之一的聚类分析也越来越受到研究者的关注。

1.2 国内外研究现状及未来发展趋势

近年来，数据挖掘引起了信息产业界的极大关注。国内外各研究机构纷纷开展了对数据挖掘技术的研究和探索工作。下面，本文将分别从国内和国外两个方面对数据挖掘技术的研究现状进行阐述，并对数据挖掘技术的未来发展趋势、研究方向及热点问题进行探讨。

1.2.1 国外研究现状

1989 年 8 月在美国 底特律召开的第 11 届国际人工智能联合会议的专题讨论会上首次出现 KDD(Knowledge Discovery in Databases)[8]这个术语。随后在 1991 年、1993 年和 1994 年都举行过 KDD 专题讨论会，汇集来自各个领域的研究人员和应用开发者，集中讨论数据统计、海量数据分析算法、知识表示、知识运用等问题。随着参与人员的不断增多，KDD 国际会议逐渐发展成为年会。

数据库、人工智能、信息处理、知识工程等领域的国际学术刊物也纷纷开辟了 KDD 专题或专刊[9]，包括《IEEE 知识与数据工程汇刊》(TKDE)，《ACM 数据库系统汇刊》(TODS)，《ACM 杂志》(JACM)，《信息系统》，

《VLDB 杂志》,《数据与知识工程》,《智能信息系统国际杂志》(JIIS)等，其中，IEEE 的 KnowledgeandDataEngineering 汇刊领先在 1993 年出版了 KDD 技术专刊，所发表的 5 篇论文代表了当时 KDD 研究的最新成果和动态，较全面地论述了 KDD 系统方法论、发现结果的评价、KDD 系统设计的逻辑方法，讨论了鉴于数据库的动态性冗余、高噪声和不确定性、KDD 系统与其它传统的机器学习、专家系统、人工神经网络、数理统计分析系统的联系和区别，以及相应的基本对策[59]。6 篇论文摘要展示了 KDD 在从建立分子模型到设计制造业的具体应用。

根据最近 Gartner 的 HPC 研究表明，“随着数据捕获、传输和存储技术的快速发展，大型系统用户将更多地需要采用新技术来挖掘市场以外的价值，采用更为广阔的并行处理系统来创建新的商业增长点。”所有这些均表明数据挖掘已成为当前计算机科学界的一大热点[45]。

群体智能的研究国外进行的比较早，当前主要是对蚁群算法的研究。自 1991 年 Dorigo 首次提出蚁群算法以来[76]，蚁群算法已在路由、优化组合、数据挖掘等多个领域取得了非常突出的成就。目前研究和应用主要集中在比利时、意大利、英国、法国、德国等欧洲国家，日本和美国开始启动。1998 年和 2000 年在比利时布鲁塞尔大学召开了第一届和第二届蚂蚁优化国际研讨会[8]。

1.2.2 国内现状

与国外相比，国内对 DMKD(Data mining and knowledge discovery)[15] 的研究稍晚，没有形成整体力量。许多单位也已开始进行数据挖掘技术的研究，但还没有看到数据挖掘技术在我国成功应用的案例。

1993 年国家自然科学基金首次支持对该领域的研究项目。目前，国内的许多科研单位和高等院校竞相开展知识发现的基础理论及其应用研究，这些单位包括清华大学、中科院计算技术研究所、空军第三研究所、海军装备论证中心等。其中，北京系统工程研究所对模糊方法在知识发现中的应用进行了较深入的研究，北京大学也在开展对数据立方体代数的研究，华中科技大学、复旦大学、浙江大学、中国科技大学、中科院数学研究所、吉林大学等单位开展了对关联规则开采算法的优化和改造[56]；南京大学、四川大学和上海交通大学等单位探讨、研究了非结构化数据的知识发现以及 Web 数据挖掘。国内也开始有关于蚁群算法的公开报道和研究成果，但严格理论基础尚未奠定，有关研究仍停留在实验探索阶段，多是对算法的研究和改进等。

1.2.3 未来发展趋势

当前，数据挖掘和知识发现的研究方兴未艾，其研究与开发的总体水平相当于数据库技术在 70 年代所处的地位，迫切需要类似于关系模式、DBMS 系统和 SQL 查询语言等理论和方法的指导。而且最近有国内大型网站评比未来十大热门技术，数据挖掘占了一席之地。鉴于数据挖掘任务和数据挖掘方法的多样性，研究的焦点可能会集中在以下几个方面[44]：

(1)研究专门用于知识发现的数据挖掘语言；

(2)研究在网络环境下的数据挖掘技术；

(3)加强对各种非结构化数据的挖掘；

(4)数据挖掘应用的探索；

(5)复杂数据类型挖掘；

(6)Web 挖掘，隐私保护和信息安全等；

(7)可视化数据挖掘；

(8)基于群体智能的数据挖掘方法；

(9)与数据库系统、数据仓库系统和 Web 数据库系统的集成；

(10)交互式发现；

1.3 本文研究的主要内容

本文主要研究的是数据挖掘中的基本问题之一—聚类分析技术。本文的主要内容包括如下几个方面：

第一章介绍了数据挖掘技术的发展、研究背景和发展方向，提出本文研究的主要内容。

第二章介绍了数据挖掘的基础知识。对数据挖掘技术的概念、产生、应用和发现模式进行了归纳和总结，为本文的全面展开作好铺垫。

第三章是有关聚类分析的研究。聚类分析已经被广泛的研究了许多年，主要集中在基于距离的聚类分析。对聚类分析的概念、产生和发展进行了简要的归纳和总结，同时对聚类分析中所用到的数据类型和主要聚类算法进行了简要的介绍。

第四章是蚁群算法的分析与研究。研究了基于群体智能的蚁群算法，对基于蚁群寻路和蚁群聚类算法作了较为详细的分析。

第五章是几种聚类算法的设计、分析与研究。在此阶段的研究中，首先介绍了 k-means 算法，结合蚁群觅食的习性提出一种基于信息素的 k-means 改进算法。然后提出一种结合基于信息素的 k-means 和基于阈值

的聚类组合算法。该算法启发于 DBSCAN 算法[78]，其思想是把高维数据集投影至二元数据矩阵，这样能够简化数据的复杂性，而且算法简单明了，更利于进行数据分析。其次借鉴经典的 LF 算法[36]和其相应的改进算法 CSI[23]、SACA[25]，模拟蚂蚁堆积蚁穴中的幼卵和尸体的行为进行数据聚类，并利用 SACA 算法中聚类标识的方法进行聚类的收集，然后把所产生的聚类中心和聚类个数作为输入，把聚类中心作为食物源，数据作为蚂蚁，借用蚂蚁寻路的思想，利用基于信息素的 k-means 方法进行二次聚类。

第六章选用 UCI 机器学习库[27]中的几组数据进行几种算法的测试和性能分析。最后选用取自某省移动公司的移动话费数据，利用蚁群聚类组合算法来进行客户行为分析。

第 2 章 数据挖掘概述

2.1 数据挖掘的概念

随着数据库技术的不断发展及数据库管理系统的广泛应用，数据库中存储的数据量急剧增大，在大量的数据背后隐藏着许多重要的信息，如果能把这些信息从数据库中抽取出来，将为公司创造很多潜在的利润，而这种从海量数据库中挖掘信息的技术，就称之为数据挖掘(Data Mining-DM)。1995 年以来，国外在知识发现和数据挖掘方面的论文非常多，已形成了热门研究方向。

从广义上讲，数据挖掘是指从大量的、不完全的、有噪声的、模糊的、随机的数据中，提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程。这个定义包括以下四个层次的含义：

(1) 数据源必须是真实的、大量的、含噪声的；

(2) 发现的是用户感兴趣的知识；

(3) 发现的知识要可接受、可理解、可运用，最好能用自然语言表达发现结果；

(4) 并不是要求发现放之四海皆准的知识，也不是要去发现崭新的自然科学定理和纯数学公式，更不是什么机器定理证明，所有发现的知识都是相对的，是有特定前提和约束条件、面向特定领域的。

从商业角度出发，数据挖掘可以描述为：按企业既定业务目标，对大量的企业数据进行探索和分析，揭示隐藏的、未知的或验证已知的规律性，并进一步将其模型化的先进有效的方法。

数据挖掘工具能够对将来的趋势和行为进行预测，从而很好地支持人们的决策，比如，经过对公司整个数据库系统的分析，数据挖掘工具可以回答诸如“哪个客户对我们公司的邮件推销活动最有可能作出反应，为什么”等类似的问题。有些数据挖掘工具还能够解决一些很消耗人工时间的传统问题，因为它们能够快速地浏览整个数据库，找出一些专家们不易察觉的极有用的信息。

目前，世界上比较有影响的典型数据挖掘系统有：SAS 公司的 Enterprise Miner、IBM 公司的 Intelligent Miner、SGI 公司的 SetMiner、SPSS 公司的 Clementine、Sybase 公司的 Warehouse Studio、RuleQuest Research 公司的 See5、还有 CoverStory、EXPLORA、Knowledge Discovery Workbench、DBMiner、Quest 等。

2.2 数据挖掘的产生

数据库技术在经过了 80 年代的辉煌之后，人们逐渐认识到，查询是数据库的奴隶，发现才是数据库的主人。“数据只为职员服务，不为决策者服务！”这是很多单位的领导在热心数据库建设后所发出的感叹。因此，在需求的驱动下，很多数据库学者从对演绎数据库的研究转向对归纳数据库的研究。

起初各种商业数据是存储在计算机的数据库中的，然后发展到可对数据库进行查询和访问，进而发展到对数据库的即时遍历。数据挖掘使数据库技术进入了一个更高级的阶段，它不仅能对过去的数据进行查询和遍历，并且能够找出过去数据之间的潜在联系，从而促进信息的传递。专家系统曾经是人工智能研究工作者的骄傲，但由于其在知识获取、知识表示、缺乏常识等方面的瓶颈，使得专家系统目前还停留在构造诸如发动机故障诊断一类的水平上。这自然促使人工智能学者开始正视现实生活中大量的、不完全的、有噪声的、模糊的、随机的大数据样本，走上了数据挖掘的道路。数理统计是应用数学中最重要、最活跃的学科之一，它在计算机发明之前就诞生了，迄今已有几百年的发展历史，然而，数理统计和数据库技术结合得并不算快。在人们有了从数据查询到知识发现、从数据演绎到数据归纳的要求之后，概率论和数理统计才获得了新的生命力，所以才会在 DMKD(数据挖掘和知识发现)这个结合点上，立即呈现出“忽如一夜春风来，千树万树梨花开”的繁荣景象。现在数据挖掘技术在商业应用中已经可以马上投入使用，因为对这种技术进行支持的三种基础技术已经发展成熟，他们是：海量数据搜集、强大的多处理器计算机、数据挖掘算法。

数据挖掘技术是人们长期对数据库技术进行研究和开发的结果，同时，也是信息技术自然演化的结果。从机器学习到知识工程，从知识工程到专家系统，80 年代人们又在新的神经网络理论的指导下重新回到机器学习，随后又进入到数据库中的知识发现，接着又相辅相成地产生数据挖掘，在此期间，数据仓库技术的出现和逐步成熟为数据挖掘技术的繁荣注入了强劲的动力，最近人们又认识到把统计分析方法和数据挖掘有机地结合将是最好的策略。因此，数据挖掘是一门交叉学科，其发展是一个螺旋上升的过程。

2.3 数据仓库与数据挖掘

2.3.1 数据仓库的定义

数据仓库[54]的概念始于 20 世纪 80 年代中期，其创始人--号称“数据仓库之父”的 William H.Inmon 在他的《建立数据仓库》一书中对数据仓库是这样定义的：数据仓库就是面向主题的(subject-oriented)、集成的(integrated)、*时变的*

(time-variant)、非易失的(nonvolatile)数据集合。数据仓库(Data Warhouse)是计算机应用领域里的一个崭新方向，它是一种信息管理技术，其研究的主要宗旨是通过通畅、合理、全面的信息管理，来达到对管理决策的支持。与联机事物处理(OLTP)相比，它完全是另一种类型的信息管理方式。

2.3.2 数据库中的知识发现和数据挖掘

数据库中的知识发现(Knowledge Discovery In Database-KDD)一词首次出现在 1989 年 8 月举行的第 11 届国际联合人工智能学术会议上[58]。许多人把数据挖掘视为数据库中的知识发现(KDD)的同义词，而另一些人只是把数据挖掘视为数据库中知识发现过程的一个基本步骤。到底怎样看，其实并不重要。如果把数据挖掘视为数据库中知识发现过程的一个基本步骤，则知识发现过程如图 1-1 所示。

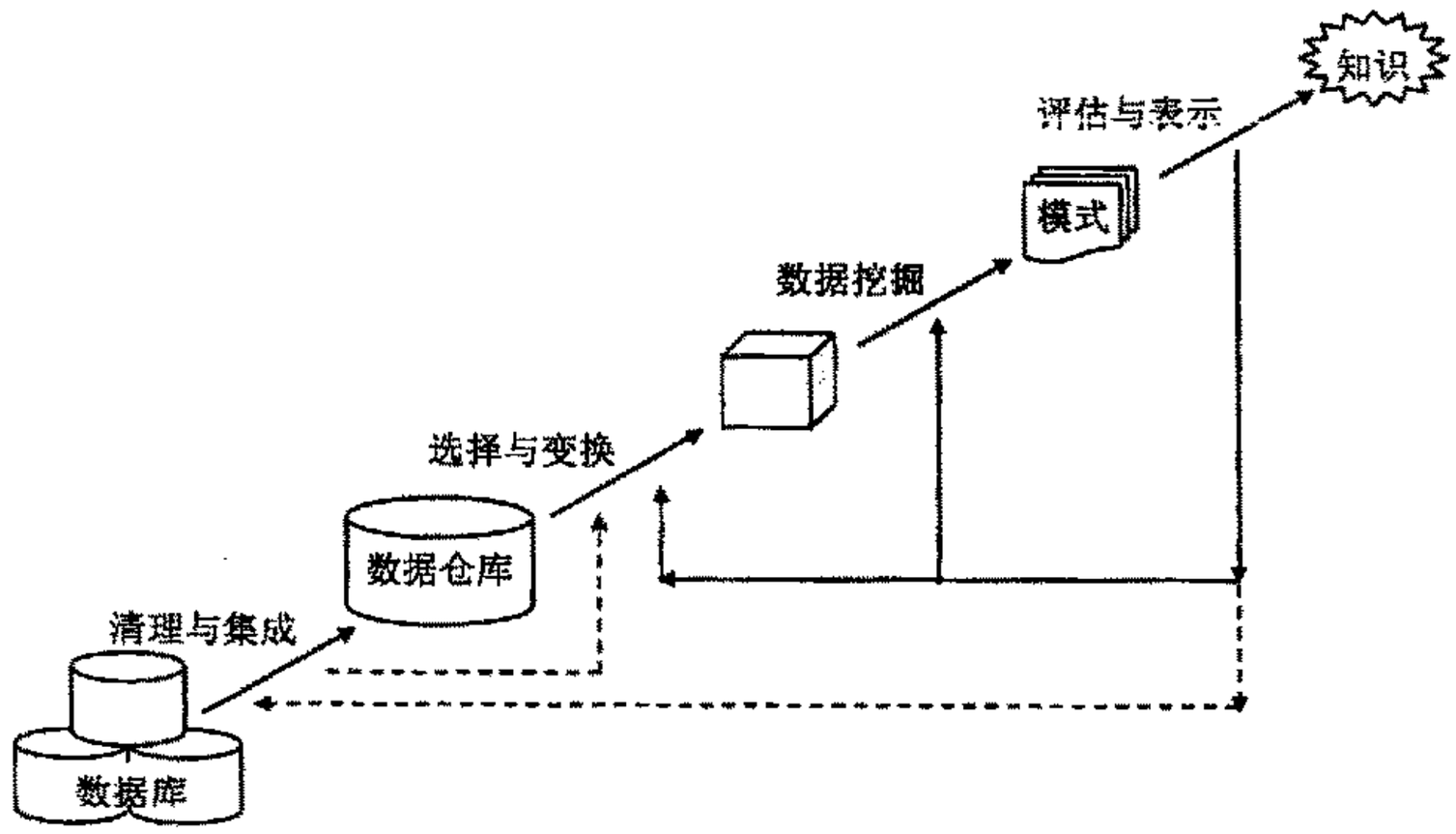

图 2-1 数据挖掘视为知识发现过程的一个基本步骤

2.4 数据挖掘的对象

原则上讲，数据挖掘可以在任何类型的信息存储上进行，这包括关系数据库、数据仓库、事务数据库、WWW、面向对象数据库、对象-关系数据库、时间序列数据库、空间数据库、文本数据库、多媒体数据库等。挖掘的原始数据可以是结构化的，如关系数据库、数据仓库中的数据；也可以是半结构化的，如文本、图形和图像数据；甚至是分布在网络上的异构型数据。

2.5 数据挖掘的应用

数据挖掘的应用非常广泛，而且数据挖掘技术从一开始就是面向应用的。目前，在很多领域，数据挖掘都是一个很时髦的词，尤其是在如银行、电信、保险、交通、零售(如超级市场)等商业领域。数据挖掘所能解决的典型商业问题包括：数据库营销(Database Marketing)、客户群体划分(Customer Segmentation & Classification)、背景分析(Profile Analysis)、交叉销售(Cross-selling)等市场分析行为，以及客户流失性分析(Churn Analysis)、客户信用记分(Credit Scoring)、欺诈发现(Fraud Detection)等等。下面介绍的一些应用可以说明数据挖掘的用途。

2.5.1 商品销售

商业部门把数据视作一种竞争性的财富可能比任 何其他部门显得更为重要，为此需要把大型市场营销数据库演变成一个数据挖掘系统。科拉福特(Kraft)食品公司(KGF)是应用市场营销数据库的公司之一，该公司搜集了购买它商品的3000 万个用户的名单，这是(KGF)通过各种促销手段得到的。KGF 定期向这些用户发送名牌产品的优惠券，介绍新产品的性能和使用情况。该公司体会到了解自己商品的用户越多，则购买和使用这些商品的机会也就越多，公司的营业状况也就越好。

2.5.2 加工制造

许多公司不仅将决策支持系统用于支持市场营销活动， 而且，由于市场竞争越演越烈，这些公司已使用决策支持系统来监视制造过程，有制造商声称已经指示它的各个办事机构，在三年内把制造成本每年降低 25%。不言而喻，该制造商经常收集各部件供应商的情况。因为，它们也必须遵循该制造商降低成本的战略。为了对付来自各方的挑战，该制造商已拥有一套“成本”决策支持系统，可以监视各供应商提供的零部件成本，以实现所制定的价格目标，这种应用需要收集有关各厂商连续一年来的产品成本信息，以便确定这种组织方式能否满 足原先制定的有关降价的战略目标。

2.5.3 金融服务/信用卡

通用汽车公司(General Motors)已经采用信用卡-GM 卡，在该公司的数据库中已拥有 1200 万个持有信用卡的客户。公司通过观察，可以了解他们正在驾驶什么样的汽车，下一步计划购买什么样的汽车及他们喜欢哪一类车辆。譬如说，一个持有信用卡的客户表示对一种载货卡车感兴趣，公司就可以向卡车部门发出一个电子邮件，并把该客户的信息告诉有关部门。

2.5.4 远程通讯

许多远程通讯的大公司近来突然发现它们面临极 大的竞争压力，这在几年前是不存在的。在过去，业务上并不需要他们密切注视市场动向，因为顾客的挑选余地有限，但是这种情况近来发生很大变化。各公司当前 都在积极收集大量的顾客信息，向他们现有的客户提供新的服务，开拓新的业务项目，以扩大他们的市场规模。从这些新的服务中，公司在短期内就可以取得更大的效益。

2.6 数据挖掘的发现模式

随着数据挖掘和知识发现(DMKD)的研究逐步走向深入，其研究已经形成了三个强大的技术支柱：数据库、人工智能和数理统计。因此，KDD 大会程序委员会曾经由这三个学科的权威人物同时来任主席。目前 DMKD 的主要研究内容包括基础理论、发现算法、数据仓库、可视化技术、定性定量互换模型、知识表示方法、发现知识的维护和再利用、半结构化和非结构化数据中的知识发现以及网上数据挖掘等。下面就数据挖掘所发现的最常见的知识模式(类型)作一简单介绍，所介绍的这些知识都可以在不同的概念层次上被发现，随着概念树的提升，从微观到中观再到宏观，以满足不同用户、不同层次决策的需要。

2.6.1 广义模式

广义模式(Generalization Pattern)是指类别特征的概括性描述知识。根据数据的微观特性发现其表征的、带有普遍性的、较高层次概念的、中观和宏观的知识，反映同类事物共同性质，是对数据的概括、精炼和抽象。

广义模式的发现方法和实现技术包括数据立方体、加拿大 SimonFraser 大学提出的面向属性的归约(attribute-oriented induction, AOI)方法等。基于数据立方体的概化技术由 E.F.Codd、S.B.Codd 和 C.T.Salley[64]提出，数据立方体中计算聚集的操作由 Gray、Chaudhuri、Bosworth、Layman、Reichart、Venkatrao、Pellow 和 Pirahesh [65]提出，面向属性的归约方法由 Cai、Cercone 和 Han[66]提出等。

2.6.2 关联模式

关联模式(Association Pattern)是反映一个事件和其他事件之间相互依赖或关联的知识。经典的关联规则发现方法可分为两步：第一步是迭代识别所有的频繁项集，要求频繁项集的支持率不低于用户设定的最低值；第二步是从频繁项集中构造可信度不低于用户设定的最低值的规则。识别或发现所有频繁项集是关联规则发现算法的核心，也是计算量最大的部分。在进行关联知识发现时，需要由用户输入最小置信度 c 和最小支持度 s。

2.6.3 序列模式

序列模式(Sequence Pattern)分析和关联分析相似，其目的也是为了挖掘数据之间的联系，但序列模式分析的侧重点在于分析数据间的前后序列关系。它能发现数据库中形如“在某一段时间内，顾客购买商品 A，接着购买商品 B，而后购买商品 C，即序列 A→B→C 出现的频度较高”之类的知识，序列模式分析描述的问题是：在给定交易序列数据库中，每个序列是按照交易时间排列的一组交易集，挖掘序列函数作用在这个交易序列数据库上，返回该数据库中出现的高频序列。在进行序列模式分析时，同样需要由用户输入最小置信度 c 和最小支持度 s。

Agrawal 和 Srikant 提出了一种类 Apriori 技术的序列模式挖掘算法[25]，Srikant 和 Agrawal 提出了挖掘序列模式的 GSP 算法[37]，Zaki、Lesh 和 Ogihara 提出了对 plan failure 的序列模式挖掘[60]，Guha、Rastogi 和 Shim 提出了一种基于约束的序列模式挖掘方法[61]，Han、Pei、Mortazavi-Asl、Chen、Dayal 和 Hsu 提出了序列模式挖掘的 FreeSpan 方法[67]，Yi、Sidiropoulos、Johnson、Jagadish、Faloutsos 和 Biliris 给出了针对时间序列的联机挖掘方法[68]等。

2.6.4 分类模式

设有一个数据库和一组具有不同特征的类别(标记)，该数据库中的每一个记录都赋予一个类别的标记，这样的数据库称为示例数据库或训练集。分类分析就是通过分析示例数据库中的数据，为每个类别做出准确的描述或建立分析模型或挖掘出分类规则，然后用这个分类规则对其它数据库中的记录进行分类。导出模型是基于对训练数据集(即其类标记已知的数据对象)的分析而产生的。分类模式的预测值可以是离散的(如根据某种动物的特征来判断这种动物是两栖动物还是哺乳动物)，也可以是连续的(如根据某人的受教育情况和工作经验来判断这个人的工资范围)。

分类模式(Classification Pattern)的实现技术包括决策树、统计、粗糙集(RoughSet)、神经网络方法等。最为典型的决策树学习系统是 ID3，它采用自顶向下不回溯策略，能保证找到一个简单的树。算法 C4.5 和 C5.0 都是 ID3 的扩展，它们将分类领域从类别属性扩展到数值型属性。线性回归和线性辨别分析是典型的统计模型分析方法。

2.6.5 聚类模式

聚类分析和分类分析是一个互逆的过程。在统计方法中，聚类分析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)。在机器学习，中聚

类分析被称作无监督或无教师归纳。在人工智能文献中，聚类也称概念聚类。与分类分析不同，聚类分析输入的是一组未分类记录，并且这些记录应分成几类事先也不知道。聚类分析就是通过分析数据库中的记录数据，根据一定的分类原则，合理地划分记录集合，确定每个记录所在类别。数据库中的记录被化分为一系列有意义的子集叫做聚(簇)类。分类原则采用最大化类内的相似性、最小化类间的相似性原则，即使得一个簇中的对象具有很高的相似性，而与其他簇中的对象很不相似。

聚类模式(Clustering Pattern)包括统计方法、机器学习方法、神经网络方法和面向数据库的方法，比如，系统聚类法、分解法、加入法、动态聚类法、有序样品聚类、有重叠聚类、模糊聚类法、运筹方法等。主要的聚类算法的类型可分为基于划分方法、基于层次的方法、基于密度的方法、基于网格的方法、基于模型的方法等。一个聚类算法通常包含了多种聚类方法的思想。在神经网络中，有一类无监督学习方法：自组织神经网络方法，如 Kohonen 自组织特征映射网络、竞争学习网络等。在数据挖掘领域里，神经网络聚类方法主要是自组织特征映射方法，IBM 在其发布的数据挖掘白皮书中就特别提到了使用此方法进行数据库聚类分割。

2.6.6 预测模式

预测模式(Prediction Pattern)根据时间序列型数据，由历史的和当前的数据去推测未来的数据，也称作时间序列模式。可以认为预测型模式是以时间为关键属性的关联模式。

时间序列预测的方法有经典的统计方法、神经网络方法和机器学习方法等。

2.6.7 偏差模式

偏差模式(Deviation Pattern)是对差异和极端特例的描述，揭示了事物偏离常规的异常现象。偏离常规的数据有时被叫做孤立点(outlier)，因此偏差模式有时被叫做孤立点模式。

偏差模式的发现方法可以分为统计的方法、基于距离的方法和基于偏移的方法等。

2.7 数据挖掘的技术

前面在介绍数据挖掘的发现模式时，针对每一种知识类型，都介绍了相应的数据挖掘技术。通常，数据挖掘技术可分为机器学习方法、统计方法、神经网络方法、数据库方法等。下面，将对这些技术作一简要的介绍。

2.7.1 机器学习方法

机器学习方法可分为：归纳学习方法(决策树、规则归纳等)、基于范例的学习、遗传算法等。

2.7.2 统计方法

统计方法可分为：回归分析(多元回归、自回归等)、判别分析(贝叶斯判别、费歇尔判别、非参数判别等)、聚类分析(系统聚类、动态聚类、分解法、加入法、模糊聚类法、运筹方法等)、探索性分析(主元分析法、相关分析法等)等。

2.7.3 神经网络方法

神经网络方法可分为：前向神经网络(BP 算法等)、自组织神经网络(自组织特征映射、竞争学习等)等。

2.7.4 数据库方法

数据库方法可分为：多维数据分析、OLAP 方法、面向属性的归约方法等。

第 3 章 聚类分析

3.1 聚类分析的定义

在日常生活、生产、科研工作中，经常要对被研究对象进行分类。研究和处理给定对象的分类常用的数学方法是聚类分析(Clustering Analysis)。聚类分析是一种重要的人类行为，它的目的是把相似的东西归为一类，使得类内具有较大的相似性，而类间具有较小的相似性。聚类分析是多元统计分析的方法之一，它试图根据数据集的内部结构将数据集分成若干个不同的子类，使得同一子类中的样本尽可能的相似，不同子类的样本尽可能的不相似。换句话说，如果将含有 n 个样本 $x_1,\ldots,x_n$ 的数据集 X 聚集成 c 个子类 $X_1,\ldots,X_c$ 则要求 $X_1,\ldots,X_c$ 满足：

$$X_1 \cup X_2 \cup \cdots \cup X_c = X$$

$$X_i \cap X_j = f \qquad (1 \leqslant i \neq j \leqslant c)$$

聚类分析已经广泛地用在许多应用中，包括模式识别，数据分析，图象处理，以及市场研究。通过聚类，人能够识别密集的和稀疏的区域，因而发现全局的分布模式，以及数据属性之间的有趣的相互关系。

与分类不同的是，它要划分的类是未知的。在分类模式中，对于目标数据库中存在哪些类这一信息我们是知道的，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。聚类分析源于许多研究领域包括数据挖掘、统计学、生物学、以及机器学习等。是在预先不知道目标数据库到底有多少类的情况下，希望将所有的纪录组成不同的类或者说“聚类”(cluster)，使得在同一聚类之间具有较大的相似度，而在不同聚类之间具有较大的相似度。聚类就是根据描述对象的属性值计算得出的相异度，将数据对象分组成为多个类或簇，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较大。距离是经常采用的度量相似度的重要方式。

数据聚类正在蓬勃发展，有贡献的研究领域包括数据挖掘、统计学、机器学习、空间数据库技术、生物学、以及市场营销。由于数据库中收集了大量的数据，聚类分析已经成为数据挖掘研究领域中一个非常括跃的研究课题。聚类分析已经被广泛地研究了许多年，主要集中在基于距离的聚类分析。基于 k-means(k-平均值)、k-medoids(k-中心点)和其他一些方法的聚类分析工具已经被加入到许多统计分析软件包或系统中，例如 S-P1us，

SPss 以及 SAS。在机器学习领域，聚类是无指导学习的一个例子。由于这个原因，聚类是观察式学习，而不是示例式学习。在概念聚类(concept clustering)中，一组对象只有当它们可以被一个概念描述时才形成一个簇。这不同于基于几何距离来度量相似度的传统聚类。概念聚类由两个部分组成：

(1) 发现合适的簇；

(2) 形成对每个簇的描述

随着计算机的发展和实际问题的需要，基于目标函数的聚类方法已成为聚类分析的主流。一方面是由于将聚类问题表述成优化问题易于与经典数学的非线性规划领域联系起来，可用现代数学方法来求解；另一方面是由于算法的求解过程比较容易用计算机来实现围绕着目标函数的优化问题。目前主要形成三大方向：一是建立合适的目标函数表达式，用数学规划方法求解最优值，如硬 c-均值聚类方法、模糊 c-均值聚类算法及其推广形式等。这类方法的主要缺陷是对初始化较敏感，易于陷入局部极小点，收敛速度慢。二是将传统的聚类方法与神经网络相结合，借助神经网络的并行实现以提高算法的收敛速度和性能。目前，针对普通聚类已提出 Kohonen 聚类神经网络等，针对模糊聚类技术已提出 Kohonen 模糊聚类神经网络等。三是将传统的聚类技术与现代化方法相结合，以克服算法对初始化敏感，易于陷入局部极小点的问题。如与模拟退火算法结合，与遗传算法结合等。

聚类算法中确定数据集中样本相似性的常用方法是欧氏距离，而且由于现实数据库中数据类型的多样性，关于如何度量两个含有非数值型字段的记录之间的距离的讨论有很多，并提出了相应的算法。在很多应用中，由聚类分析得到的每一个聚类中的成员都可以被统一看待。

聚类的用途是很广泛的。在商业上，聚类可以帮助市场分析人员从他们的消费者数据库中区分出不同的消费群体来，并且概括出每一类消费者的消费模式或者说习惯；在生物学中，它可以被用来辅助研究动、植物的分类，可以用来分类具有相似功能的基因，还可以又来发现人群中的一些潜在的结构等等；聚类还可以用来从地理数据库中识别出具有相似土地用途的区域；可以从保险公司的数据库中发现汽车保险中具有较高索赔概率的群体；还可以从一个城市的房地产信息数据库中，根据房型、房价及地理位置分成不同的类；还可以用来从万维网上分类不同类型的文档等。同时，聚类分析作为数据挖掘中的一个模块，它既可以作为一个单独的工具以发现数据库中数据分布的一些深入的信息，并且概括出每一类的特点，

或者把注意力放在某一个特定的类上以作进一步的分析；聚类分析也可以作为数据挖掘算法中其他分析算法的一个预处理步骤。

聚类分析是一个具有很强挑战性的领域，它的一些潜在的应用对分析算法提出了特别的要求，下面列出一些典型的要求：

- 可伸缩性：这里的可伸缩性是指算法要能够处理大数据量的数据库对象，比如处理上百万条纪录的数据库。这就要求算法的时间复杂度不能太高，最好当然是多项式时间的算法。值得注意的是，当算法不能处理大数据量时，用抽样的方法来弥补也不是一个好主意，因为这通常导致歪曲的结果。

1. 处理不同字段类型的能力：也即算法不仅要能处理数值性的字段，还要有处理其它类型字段的能力，例如：布尔型、枚举型、序数型、以及混合型等。
2. 发现具有任意形状的聚类的能力：很多聚类分析算法采用基于欧几里德距离的相似性度量方法，这一类算法发现的聚类通常是一些球状的、大小和密度相近的类；但可以想见，现实数据库中的聚类可以是任意的形状，甚至是具有分形维度的形状，故要求算法有发现任意形状的聚类的能力。
3. 输入参数对领域知识的弱依赖性：很多聚类算法都要求用户输入一些参数，例如需要发现的聚类数，结果的支持度、置信度等，聚类分析的结果通常都对这些参数很敏感，但另一方面，对于高维数据，这些参数又是相当难以确定的。这样就加重了用户使用这个工具的负担，使得分析的结果很难控制。一个好的聚类算法应该针对这个问题，给出一个好的解决方法。
4. 能够处理异常数据：现实数据库中常常包含有异常数据，或者数据不完整，缺乏某些字段的值，甚至是包含错误数据的现象，有一些聚类算法可能会对这些数据很敏感，从而导致错误的分析结果。
5. 结果对输入记录顺序的无关性：有些分析算法对纪录的输入顺序是敏感的，也即，对同一个数据集，将它以不同的顺序输入到分析算法，得到的结果会不同，这是我们不希望的。
6. 处理高维数据的能力：一个数据库或者数据仓库都有很多的字段或者说维，一些分析算法在处理维数比较少的数据集时表现不错，例如两、三维的数据；人的理解能力也可以对两、三维数据的聚类分析结果的质量作出较好的判别，但对于高维数据就没有那么直观了。所以对于高维数据的聚类分析是很具有挑战性的，特别是考虑到在高维空间中，

数据的分布是极其稀疏的，而且形状也可能是极其不规则的。

7. **增加限制条件后的聚类分析能力**：现实的应用中总会出现各种其它限制，我们希望聚类算法可以在考虑这些限制的情况下，仍旧有较好的表现。
8. **结果的可解释性和可用性**：聚类的结果最终都是要面向用户的，所以结果应该是容易解释和理解的，并且是可应用的。这就要求聚类算法必须与一定的语义环境，语义解释相关联。领域知识是如何影响聚类分析算法的设计是很重要的一个研究方面。

3.2 聚类分析中的数据类型

3.2.1 数据矩阵

数据矩阵是一个对象-属性结构。它是由 n 个对象组成，设聚类问题中有 n 个对象组成：$x_i(i=1,2,...,n)$，每个对象有 p 个属性，第 i 个对象第 j 个属性的观测值为 x_{ij}。数据矩阵采用关系表形式或 n×p 矩阵来表示[60]。

$$\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \tag{3-1}$$

(3-1)常称为数据矩阵，其中第 i 个对象 p 个变量的观测值记为：

$$x_i = (x_{i1}, x_{i2}, \cdots, x_{ip})^T \tag{3-2}$$

3.2.2 相似性矩阵

相似性矩阵是一个对象-对象结构。它存放所有 n 个对象彼此之间所形成的相似性。它一般采用 n×n 矩阵来表示。

$$\begin{bmatrix} 0 & & & \\ d(2,1) & 0 & & \\ \vdots & \vdots & \ddots & \\ d(n,1) & d(n,2) & \cdots & 0 \end{bmatrix} \tag{3-3}$$

其中 d(i,j)=d(j,i)且 d(i,i)=0，d(i,j)表示对象 i 和对象 j 之间的相似性的量化表示，通常 d(i,j)为一个非负数。当对象 i 和对象 j 非常相似或彼此“接近”时，该数值接近 0；该数值越大，就表示对象 i 和对象 j 越不相似。

数据矩阵经常被称为二模(two-mode)矩阵，而相似性矩阵被称为单模

(one-mode)矩阵。许多聚类算法以相似性矩阵为基础。如果数据是用数据矩阵的形式表现的，在使用该类算法之前要将其转化为相似性矩阵。

3.2.2 区间标度变量

区间标度变量是一个粗略线性标度的连续度量，典型的例子包括质量和高度，经度和纬度座标(如聚类房屋)，以及大气温度等。采用的测量单位会对聚类分析产生不同程度的影响，导致不同的聚类结构。一般而言，选用的度量单位越小，变量可能的值域就越大，这样对聚类结果的影响也越大。为了避免对度量单位选择的依赖，数据应当标准化。所谓标准化就是给所有属性相同的权值。这一做法在没有任何先验知识的情况下是非常有用的，但在一些应用中，用户会有意识地赋予某些属性更大权值以突出其重要性。

为了实现标准化测量，一种方法就是将初始测量值转换为无单位变量。给定一个属性(变量)f，可以利用以下计算公式对其进行标准化：

1) 计算绝对偏差均值 s_f

$$s_f = \frac{1}{n}(|x_{1f} - m_f| + |x_{2f} - m_f| + \cdots + |x_{nf} - m_f|) \tag{3-4}$$

其中 $x_{1f}, x_{2f}, \ldots, x_{nf}$ 是变量 f 的 n 个测量值。m_f 是变量 f 的平均值，即 $m_f = (x_{1f} + x_{2f} + \cdots + x_{nf})/n$。

2)计算标准化的测量值，或 z-score：

$$z_{if} = \frac{x_{if} - m_f}{s_f} \tag{3-5}$$

其中绝对偏差均值 s_f 要比标准偏差 s_f 更为鲁棒(对含有噪声数据而言)。在计算绝对偏差均值时，对均值的偏差 $|x_{1f} - m_f|$ 没有进行平方运算，因此异常数据的作用被降低；还有一些关于针对分散数据更鲁棒的处理方法，如：中间值绝对偏差方法。但是利用绝对偏差均值的好处就是：异常数据(outlier)的 z- 分值不会变得太小，从而使得异常数据仍是可识别的。

3.2.3 相似性度量

在标准化处理后，或在无需标准化的特定应用中，对象间的相似性(或相似度)是基于对象间的距离来计算的。最常用的距离度量方法是明氏(Minkowski)距离、马氏(Mahalanobis)距离、Cosine距离等。

1) 明氏(Minkowski)距离

$$d_{ij}(q)=(\sum_{k=1}^{p}|x_{ik}-x_{jk}|^{q})^{1/q}$$

其中q>0，$d_{ij}(q)$为对象i和对象j之间的距离，常用于欧氏空间中。

当q为1时，明氏距离即为绝对值距离：

$$d_{ij}(1)=(\sum_{k=1}^{p}|x_{ik}-x_{jk}|)$$

当q为2时，明氏距离即为欧氏(Euclid)距离：

$$d_{ij}(2)=(\sum_{k=1}^{p}|x_{ik}-x_{jk}|^{2})^{1/2}$$

2) 马氏(Mahalanobis)距离

考虑到对象的各变量的观测值往往为随机值，因此第 i 个对象的 p 个分量的观测值 $x_i=(x_{1i},x_{2i},\cdots,x_{pi})^T$ 为 p 维随机向量。由于随机向量有一定的分布规律，各分量之间又具有一定的相关性，因此两个对象作为两个随机向量的个体，则第 i 个与第 j 个对象间的马氏距离的平方表示为：

$$d_{ij}^{2}(M)=(x_i-x_j)^{T}\sum{}^{-1}(x_i-x_j)$$

其中 Σ 是随机变量的协方差矩阵。

3)Cosine 距离

在聚类分析中，每个对象可以看作 n 维空间中的向量，第 i 个对象可表示为：$x_i=(x_{1i},x_{2i},\cdots,x_{pi})^T$。它们的相似系数可用两个向量间的夹角余弦来表示，于是第 i 个与第 j 个对象的相似系数表示为：

$$q_{ij}=\cos(\theta_{ij})=\frac{\sum_{k=1}^{p}X_{ik}X_{jk}}{\sqrt{\sum_{k=1}^{p}X_{ik}^{2}\sum_{k=1}^{p}X_{jk}^{2}}}$$

3.3 聚类分析中的主要算法

聚类分析的算法可以分为以下几大类：分裂法、层次法、基于密度的方法、基于网格的方法和基于模型的方法等。

3.3.1 分裂法

给定一个有 N 个元组或者纪录的数据集，分裂法将构造 K 个分组，每一个分组就代表一个聚类，K<N。而且这 K 个分组满足下列条件：(1) 每一个分组至少包含一个数据纪录；(2)每一个数据纪录属于且仅属于一个分组(注意：这个要求在某些模糊聚类算法中可以放宽)；对于给定的 K，算法首先给出一个初始的分组方法，以后通过反复迭代的方法改变分组，使得每一次改进之后的分组方案都较前一次好，而所谓好的标准就是：同一分组中的记录越近越好，而不同分组中的纪录越远越好。使用这个基本思想的算法有：K-MEANS 算法、K-MEDOIDS 算法、CLARANS 算法；

3.3.2 分层聚类法

这种方法对给定的数据集进行层次似的分解，直到某种条件满足为止。具体又可分为“自底向上”和“自顶向下”两种方案。例如在“自底向上”方案中，初始时每一个数据纪录都组成一个单独的组，在接下来的迭代中，它把那些相互邻近的组合并成一个组，直到所有的记录组成一个分组或者某个条件满足为止。代表算法有：BIRCH 算法、CURE 算法、CHAMELEON 算法等；自下而上聚合层次聚类方法和自顶而下分解层次聚类方法中，用户均需要指定所期望的聚类个数作为聚类过程的终止条件。

3.3.3 基于密度的方法

基于密度的方法与其它方法的一个根本区别是：它不是基于各种各样的距离的，而是基于密度的。这样就能克服基于距离的算法只能发现“类圆形”的聚类的缺点。这个方法的指导思想就是，只要一个区域中的点的密度大过某个阈值，就把它加到与之相近的聚类中去。代表算法有：DBSCAN 算法、OPTICS 算法、DENCLUE 算法等；

3.3.4 基于网格的方法

基于网格的方法首先将数据空间划分成为有限个单元(cell)的网格结构,所有的处理都是以单个的单元为对象的。这么处理的一个突出的优点就是处理速度很快，通常这是与目标数据库中记录的个数无关的，它只与把数据空间分为多少个单元有关。代表算法有：STING 算法、CLIQUE 算法、WAVE-CLUSTER 算法；

3.3.5 基于模型的方法

基于模型的方法给每一个聚类假定一个模型，然后去寻找能够很好地满足这个模型的数据集。这样一个模型可能是数据点在空间中的密度分布函数或者其它。它的一个潜在的假定就是：目标数据集是由一系列的概率分布所决定的。通常有两种尝试方向：统计的方案和神经网络的方案。

第 4 章 蚁群算法

4.1 群体智能

4.1.1 群体智能的定义

我们经常能够看到成群的鸟、鱼或者浮游生物。这些生物的聚集行为有利于它们觅食和逃避捕食者。它们的群落规模动辄以十、百、千甚至万计，并且经常不存在一个统一的指挥者。它们是如何完成聚集、移动这些功能呢？受这些社会性昆虫行为的启发，通过对其行为的模拟，产生了一系列解决复杂优化问题的新思路和方法，这些研究被称为对群体智能[56] (Swarm Intelligence)的研究。

群居昆虫以集体的力量，进行觅食、御敌、筑巢的能力。这种群体所表现出来的通过大量数目的智能体群来实现的智能方式，就称之为群体智能。如蜜蜂采蜜、筑巢、蚂蚁觅食、筑巢等，从群居昆虫互相合作进行工作中，得到启迪，研究其中的原理，以此原理来设计新的求解问题的算法。群体智能中的群体(Swarm)指的是“一组相互之间可以进行直接通信或者间接通信(通过改变局部环境)，并且能够合作进行分布问题求解的主体”。作为实现群体智能的每一个个体，它们的功能相对于整个问题的求解是有限的，而且每个智能个体在整个智能系统中只能实现总体功能的一小部分，其智能寻优方式实现是通过智能群体的总体优化特征来实现的。

群体智能的特点是最小智能但自治的个体利用个体与个体和个体与环境的交互作用实现完全分布式控制，并具有自组织、可扩展性、健壮性等特性。但群体智能的研究还处于萌芽阶段，还存在很多问题。首先，它们都是概率算法，从数学上对于它们的正确性与可靠性的证明还是比较困难的，所做的工作也比较少。其次，这些算法都是专用的算法，一个算法只能解决某一类问题，各种算法之间的相似性很差。并且系统的高层次的行为是需要通过低层次的昆虫之间的简单行为交互涌现(Emerge)产生的。单个个体控制的简单并不意味着整个系统设计的简单，我们必须能够将高层次的复杂行为(也就是系统所要执行的功能，例如 TSP 问题[24]、数据聚类、搬运箱子)映射到低层次的简单个体的简单行为(例如信息素的遗留、物体的拾起与放下)上面，而这二者之间是存在较大差别的。并且我们在系统设计时也要保证多个个体简单行为的交互能够涌现出我们所希望看到的高层次的复杂行为。这可以说是群体智能中一个极为困难的问题。

群体智能具有协作性、分布性、鲁棒性和快速性等特点，其优点可以描述如下：

（1） 群体中相互合作的个体是分布式的，这样的分布模式更适合于网络环境下的工作状态。

（2） 系统中没有集中的控制指令与数据存储，这样的系统更具有鲁棒性，不会由于某一个或几个个体的故障而影响到整个问题的求解进程。

（3） 系统不通过个体间的直接通信，而通过非直接通信方式进行信息的传输与合作，这样的系统具有更好的可扩充性，而且由于系统中个体的增加而增加的通信开销也比较小。

（4） 系统中每个个体的能力十分简单，每个个体的执行时间也比较短，并且实现较为方便，具有简单性的特点

由于以上这些优点，群体智能作为智能计算与智能控制等领域发展的一个重要方向，越来越受到研究者的关注。一些启发于群居性生物的觅食、打扫巢穴等行为而设计的算法已经成功地解决了许多组合优化、通信网络和机器人等领域的实际问题。当前对群体智能的研究主要集中在对模拟生物蚁群智能寻优能力的蚁群算法、模拟鸟群运动的微粒群等算法上。

4.1.2 蚁群算法的提出

蚁群算法是最近几年才提出的一种新型的模拟进化算法。蚂蚁是大家司空见惯的一种昆虫，而他们的群体合作的精神令人钦佩。他们的寻食、御敌、筑巢(蚂蚁的筑窝、蜜蜂建巢)之精巧令人惊叹。蚂蚁是自然界中常见的一种生物，人们对蚂蚁的关注大都是因为“蚂蚁搬家，天要下雨”之类的民谚。然而随着近代仿生学的发展，这种似乎微不足道的小东西越来越多地受到学者们的关注。1991 年 M.Dorigo、V.Maniezzo 等人首先提出了蚁群算法[74](Ant Colony Algorithms)，人们开始了对蚁群的研究：相对弱小，功能并不强大的个体是如何完成复杂的工作的(如寻找到食物的最佳路径并返回等)。在此基础上一种很好的优化算法逐渐发展起来。

人们经过大量研究发现，蚂蚁个体之间是通过一种称之为信息素(Pheromone)的物质进行信息传递，从而能相互协作，完成复杂的任务。蚂蚁在运动过程中，能够在它所经过的路径上留下该种物质，而且蚂蚁在运动过程中能够感知这种物质的存在及其强度，并以此指导自己的运动方向，蚂蚁倾向于朝着该物质强度高的方向移动。因此，由大量蚂蚁组成的蚁群的集体行为便表现出一种信息正反馈现象：某一路径上走过的蚂蚁越多，则后来者选择该路径的概率就越大。蚂蚁个体之间就是通过这种信息的交流达到搜索食物的目的。

蚁群算法的主要特点是通过正反馈、分布式协作来寻找最优解，这是一种基于种群寻优的启发式搜索算法，它充分利用了生物蚁群能通过个体间的信息传

递，用蚁群在搜索食物源的过程中所体现出来的寻优能力来解决一些离散系统优化中困难问题。已经用该方法求解了旅行商问题、二次指派问题、Job-Shop调度问题等，并取得了一系列较好的实验结果，充分体现了该算法适用于组合优化类问题求解的优越特征。

4.2 基于蚂蚁寻路的蚁群算法

4.2.1 蚂蚁觅食行为

蚁群寻找食物时会派出一些蚂蚁分头在四周游荡，如果一只蚂蚁找到食物，它就返回巢中通知同伴并沿途留下"信息素"(Pheromone)作为蚁群前往食物所在地的标记。信息素会逐渐挥发，如果两只蚂蚁同时找到同一食物，又采取不同路线回到巢中，那么比较绕弯的一条路上信息素的气味会比较淡，蚁群将倾向于沿另一条更近的路线前往食物所在地。

4.2.2 基本的蚂蚁算法

在自然界中，单个的蚂蚁个体行为极为简单，但由多个蚂蚁所组成的群体却成功地在搜寻食物等方面表现出复杂的行为。蚂蚁在搜索食物源时能很快地找到通向食物的最短路径，通过研究发现，蚂蚁会在走过的路径上留下信息素(pheromone)。蚂蚁个体之间通过这种信息素物质进行信息传递，蚂蚁在移动过程中通过感知遗留在路上的该种物质来指导自己的运动方向，并在自己经过的路径上留下该类物质，信息素多的地方经过的蚂蚁就多。大量蚂蚁所组成的群体便构成了一种信息正反馈，经过一段时间的正反馈过程，从而成功地实现了食物搜索，最短路径选择等行为。蚁群算法正是通过模拟蚂蚁的这种行为来达到目的。

蚁群算法是受蚂蚁的行为启发而产生的一种“自然”算法。它是从对蚁群行为的研究中产生的。M.Dorigo 等人在关于蚁群算法的第一篇文章中指出的：蚁群中的蚂蚁以“信息素”为媒介的间接的异步的联系方式是蚁群算法的最大的特点。蚂蚁在行动(寻找食物或者寻找回巢的路径)中，会在它们经过的地方留下一些化学物质。这些物质能被同一蚁群中后来的蚂蚁感受到，并作为一种信号影响后到者的行动(具体表现在后到的蚂蚁选择有这些物质的路径的可能性，比选择没有这些物质的路径的可能性大得多)，而后到者留下的信息素会对原有的信息素进行加强，并如此循环下去。这样，经过蚂蚁越多的路径，在后到蚂蚁的选择中被选中的可能性就越大(因为残留的信息素浓度较大的缘故)。由于在一定的时间内，越短的路径会被越多的蚂蚁访问，因而积累的信息素也就越多，在下一个时间内被其他的蚂蚁选中的可能性也就越大。这个过程会一直持续到

所有的蚂蚁都走最短的那一条路径为止。

图 4-1 中有一条蚂蚁经过的路径，我们假设 a 点是食物，而 e 点是蚂蚁的巢穴，如图 1a)所示。在某一个时刻忽然有一个障碍物出现在蚂蚁经过的路径中，原有的路径被切断，从 a 点到 e 点的蚂蚁就必须在 b 点决定应该往左还是往右走。而从 e 点到 a 点的蚂蚁也必须在 d 点决定选择哪条路径。这种决定会受到各条路径上以往蚂蚁留下的信息素浓度的影响。如果向右的路径上的信息素浓度比较大，那么向右的路径被蚂蚁选中的可能性也就比较大一些。但是对障碍出现后第一个到达 b 点或 d 点的蚂蚁而言，因为没有信息素的影响，所以它们选择向左或者向右的可能性是一样的，如图 4-1(b)所示。

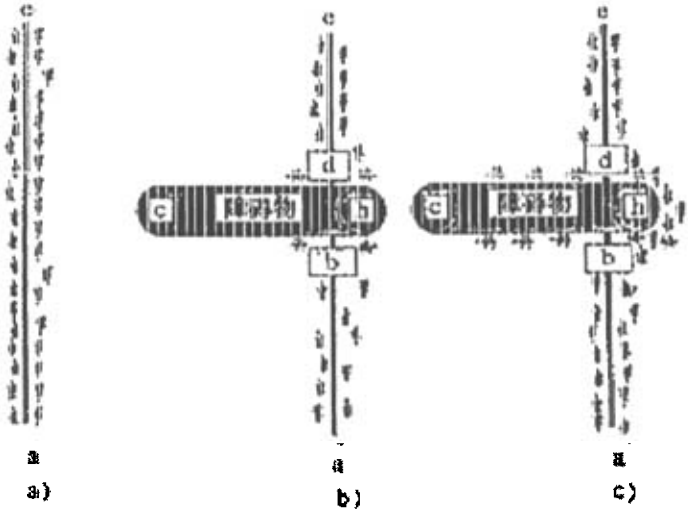

图 4-1 蚂蚁觅食路径示意图

若以从 a 点到 e 点的蚂蚁为例进行说明，对于从 e 点到 a 点的蚂蚁而言过程基本是一样的。由于路径 bhd 比路径 bcd 要短，因此选择 bhd 路径的第一只蚂蚁要比选择 bcd 的第一只蚂蚁早到达 d 点。此时，从 d 点向 b 点看，指向路径 dhb 的信息素浓度要比指向路径 dcb 的信息素浓度大。因此从下一时刻开始，从 e 点经 d 点达到 a 点的蚂蚁选择 dhb 路径比选择 dcb 路径的可能性要大得多从而使路径 bhd(或 dhb)上信息素浓度与路径 bcd (或 dcb)上信息素浓度的差变大。而信息素浓度差变大的结果是选择路径 bhd(或 dhb)路径的蚂蚁进一步增加，这又导致信息素浓度差进一步加大。

在自然界中，蚁群的这种寻找路径的过程表现为一种正反馈的过程，与人工蚁群的寻优算法极为一致。如我们把只具备了简单功能的工作单元视为“蚂蚁”，

那么上述寻找路径的过程可以用于解释人工蚁群的寻优过程。

由以上分析可知，人工蚁群和自然界蚁群的相似之处在于，两者优先选择的都是含“信息素”浓度较大的路径；这在两种情况下，较短的路径上都能聚集比较多的信息素；两者的工作单元(蚂蚁)都是通过在其所经过的路径上留下一定信息的方法进行间接的信息传递。而人工蚁群和自然界蚁群的区别在于，人工蚁群具有一定的记忆能力。它能够记忆已经访问过的节点；另外，人工蚁群在选择下一条路径的时候并不是完全盲目的，而是按一定的算法规律有意识地寻找最短路径。

4.2.3 TSP 问题的蚁群算法

由于蚂蚁寻找从蚁巢到食物源的最短的路径与 TSP 问题相似，基本的蚁群算法是与旅行商(TSP)[77]问题的求解联系在一起的。TSP 问题属于一种典型的组合优化问题，其定义为：给定 n 个城市的集合，寻找一条只经过各城市一次的具有最短长度的闭合路径。设(x_i, y_i)是城市 i 的坐标值，d_{ij}为城市 i 和城市 j 之间的距离，用欧几里德空间距离表示：

$$d_{ij}=\sqrt{(x_i-x_j)^2+(y_i-y_j)^2} \tag{4-1}$$

一个 TSP 问题可由图(N，E)给定，其中 N 是城市的集合，E 是城市之间的支路集合(欧几里德空间中 TSP 意义下的一个全连接图)，令 $b_i(t)$(i=1，2，…，n)为 t 时刻位于城市 i 的蚂蚁个数，则 $m=\sum_{i=1}^{n} b_i(t)$ 为蚁群中蚂蚁的总个数。每个蚂蚁可认为具有下列特征的简单智能体：

1） 其选择城市的概率是城市之间的距离和连接支路所包含的当前信息素余量的函数。

2） 为了强制蚂蚁进行合法的周游，直到周游完一次所有的城市，才允许蚂蚁游走已访问过的城市，设置禁忌表来进行控制。

3） 当完成一次周游后，它在每条访问过的支路上都会留下信息素。

设 $\tau_{ij}(t)$为 t 时刻在 ij 连线上残留的信息量，而初始时刻各条路径上的信息量相等，即 $\tau_{ij}(0)=C$。如果在时间间隔(t，t+1)中 m 个蚂蚁都从当前城市选择下一个城市，则经过 n 个时间间隔。为了避免残留信息过多引起的残留信息淹没启发信息的问题，在每一只蚂蚁完成对所有 n 个城市的访问后(也即一个循环结束后)，必须对残留信息进行更新处理，模仿人类记忆的特点，对旧的信息进行削弱。同时，必须将最新的蚂蚁访问路径的信息加入 τ_{ij}，此时按如下方法修改各条路径上的残留信息。

$$\tau_{ij}(t+n) = \rho\tau_{ij}(t) + \Delta\tau_{ij} \tag{4-2}$$

$$\Delta\tau_{ij} = \sum_{k=1}^{m} \Delta\tau_{ij}^{k}$$

上式中， ρ 为信息残留系数，

1- ρ 表征了从时刻 t 到 t+n 路径 ij 上残留信息的挥发程度。

$\Delta\tau_{ij}^{k}$ 为本次循环第 k 只蚂蚁在 t 与 t+n 时刻，留在路径 ij 上的单位长度上的信息量。

根据 Dorigo 的 Ant-Cycle System 模型，有

$$\Delta\tau_{ij}^{k} = \begin{cases} \dfrac{Q}{L_k}, & \text{若第}k\text{只蚂蚁在其周游过程的}t\text{和}t+n\text{时刻经过支路}L_k \\ 0, & \text{其它} \end{cases} \tag{4-3}$$

上式中，Q 为常量，L_k 为第 k 只蚂蚁在本次循环中所走路径的长度。则 t 时刻蚂蚁 k(k=1，2，3，…，n)由城市 i 到城市 j 的选择概率定义如下：

$$P_{ij}^{k}(t) = \frac{\tau_{ij}^{\alpha}(t)\eta_{ij}^{\beta}}{\sum_{k \in N_i^{allowed}} \tau_{ik}^{\alpha}(t)\eta_{ik}^{\beta}} \tag{4-4}$$

定义 $tabu_k$ 为一动态增长的列表，其中记录了蚂蚁 k 所经过的所有城市号。$N_i^{allowed}$ 为允许第 k 只蚂蚁访问的城市列表，则 $N_i^{allowed} = \{n - tabu_k\}$。$\eta_{ij}$ 为 t 时刻蚂蚁由城市 i 选择城市 j 的某种启发信息，在 TSP 问题中，通常取 $\eta_{ij} = \frac{1}{d_{ij}}$，如果 $d_{ij} = 0$，则 η_{ij} 取一个比较大的数值。而 α 和 β 则分别为残留信息和启发信息的相对重要程度系数，其中 Q， α， β， ρ 的最佳组合可以由实验确定。

Ant cycle 算法流程如图 4-2 所示:

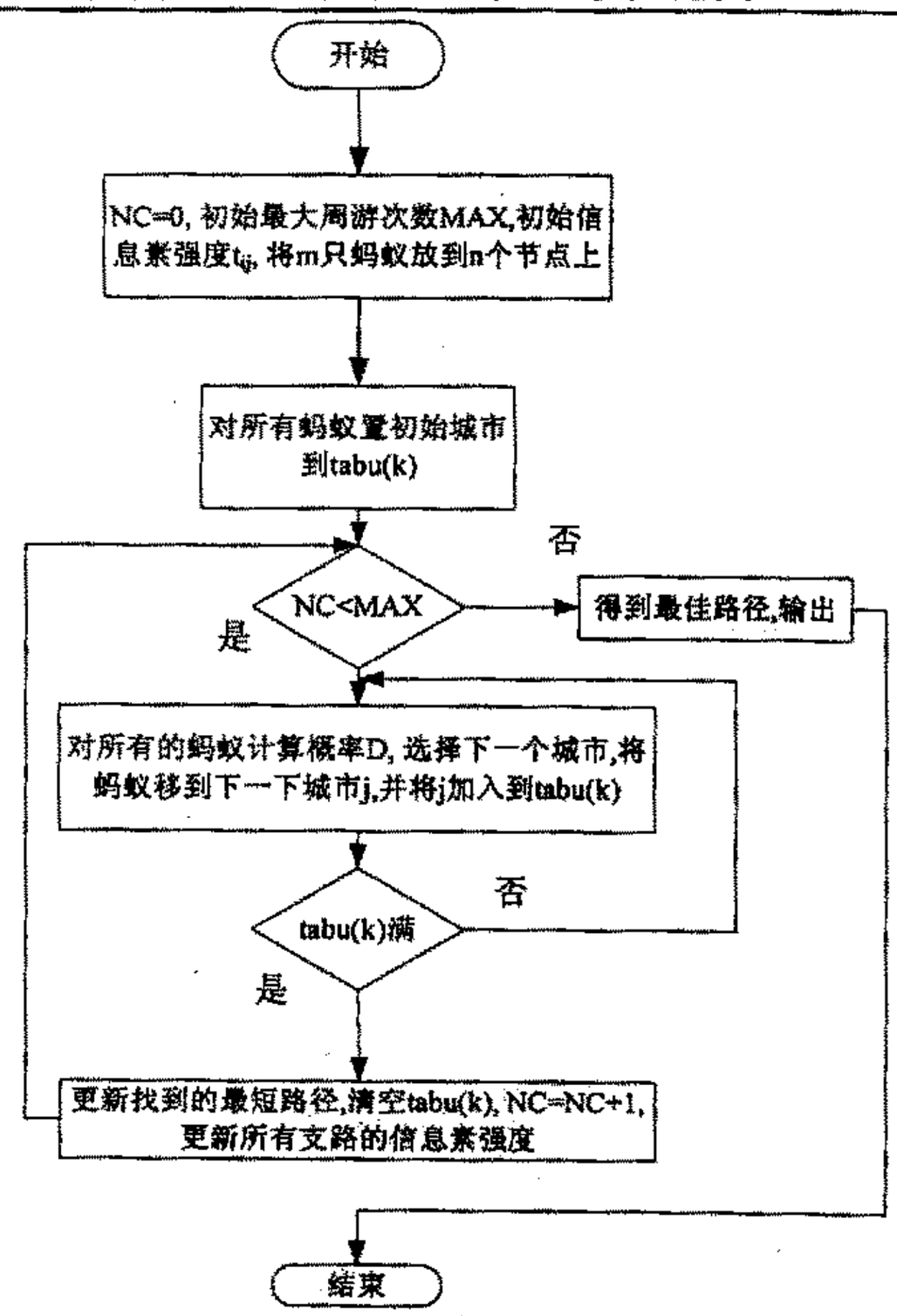

图 4-2 Ant-cycle 算法流程图

而后 Dorigo 等人又提出了蚁群算法的另外两个版本：蚁密算法(ant density)和蚁量算法(ant quantity)，这两种算法在信息素更新的方式上与蚁周算法是有差异的。这两种算法的模型中，每只蚂蚁在每一步后都留下了它的信息素，而不必等到周游结束。在蚁密算法中，蚂蚁每次从 i 到 j 都会在支路(i,j)上留下数量为 Q 的信息素;在蚁量算法中,一只从 i 至 j 的蚂蚁在支路(i,j)上留下数量为 Q/d_{ij} 的信息素。其更新方式定义如下：

蚁密算法：

$$\Delta\tau_{ij}^{k}=\begin{cases}Q, & \text{若第}k\text{只蚂蚁在其周游过程的}t\text{和}t+n\text{时刻从}i\text{至}j\\ 0, & \text{其它}\end{cases}$$

蚁量算法：

$$\Delta\tau_{ij}^{k}=\begin{cases}\dfrac{Q}{d_{ij}}, & \text{若第}k\text{只蚂蚁在其周游过程的}t\text{和}t+n\text{时从}i\text{至}j\\ 0, & \text{其它}\end{cases}$$

4.2.4 蚁群算法的改进模式

在以旅行商问题为代表的组合优化类问题的求解中，蚁群算法表现出了非常优越的求解特征。当前对蚁群算法的研究结果表明，蚁群算法利用了正反馈的原理，在一定程度上加快了进化过程；而且该算法是一种本质并行的算法，个体间不断地进行信息交流和传递，有利于较优解的发现。虽然单个个体容易收敛于局部最优，但多个个体通过合作，会很快收敛于解空间的某一子集。实验表明，蚁群算法具有通用性、鲁棒性、群体性和并行性等特点。但这种算法也存在着一些缺陷：需要较长的搜索时间，当问题规模较大时也易陷入过早收敛的问题等。

作为一种体现群体智能协作寻优特征的启发式优化算法，当前对蚁群算法的研究改进主要从以下两个方面来进行的：

（1） 算法的总体结构和组织模式

常见的算法如多蚁群算法和混合型蚁群算法等。这些算法利用多个蚁群来协同求解，在这些算法的信息联系模式中，蚁群群体层的交互还利用到了正负两种信息素效应等。由于引入了群体层的交互，蚁群能更好地交换问题解决过程中的规划信息，并保持它们在搜索过程中的多样性。

（2） 算法的具体参数设定和调整策略

这里，可以对参数的选择和变化模式进行设定，常见的算法如最大最小蚁群算法、带变异特征的蚁群算法、ANT-Q 算法、带禁忌搜索策略的蚁群算法等。

4.3 基于蚂蚁聚类的蚁群算法

4.3.1 蚁群聚类的仿生原理

Chretien 用 Lasiusniger 蚂蚁做了蚁群蚁穴墓地组织的实验[20]，后来 Deneubourg 等人也用 Pheidole palliduia 蚂蚁做了类似的实验[29]。实验证实：某些种类的种群的确能够组织蚁穴中的墓地，也就是将分散在蚁穴各处的蚂蚁尸体堆垒起来。另外观察还发现蚁群在安排不同蚁卵的位置时，按照蚁卵大小不同而分别堆放在蚁穴周边和中央位置。通过对这些实验的观察和研究，Deneubourg 等人提出了一种解释蚁群聚类现象的基本模型(Basic Model，BM)，并模拟实现了蚁群的聚类过程。这一基本模型认为单独的对象将被拾起并放到

其它有更多同一类型对象的地方。假设环境中只有一种类型的对象，由一个当前没有负载对象、随机移动的蚂蚁拾起一个对象的概率是：

$$p_p = (\frac{k_1}{k_1 + f})^2 \tag{4-5}$$

其中 f 是在蚂蚁附近对象观察分数(Perceived Fraction)，反映蚂蚁附近同类对象的个数，k_1 是阈值常数：若 f << k_1，则 p_p 接近 1(即当周围没有多少对象时，拾起一个对象的概率很大)；若 k_1 << f，则 p_p 接近 0(即在一个稠密的聚类中，一个对象不大可能被移动)。一个随机移动的有负载的蚂蚁放下一个对象的概率是：

$$p_d = (\frac{f}{k_2 + f})^2 \tag{4-6}$$

其中，k_2 是阈值常数：若 f <<k_2，则 p_d 接近 1；若与 k_2<< f ， p_d 接近 0。拾起和放下的行为大致遵守相反的规则。

为了跟踪聚类的动态过程，Gutowitz 提出了采用空间熵的方法[18]。空间熵用于度量对象聚集的效果。设 s-Patches 为一个空间区域(例如 s=5 表示一个 5X 5 的区域)，空间熵定义为：

$$E_s = \sum_{l \in (s-patches)} P_l \log P_l \tag{4-7}$$

其中，P_l 是在区域 s-Patches 内对象个数与总对象个数的比值，E_s 随着聚类过程而减小。

4.3.2 LF 算法

Lumer 和 Faieta 将 BM 推广应用到数据分析，提出了 LF 算法[36]。主要思想是定义一个在对象属性空间里对象之间的距离(或者是不相似性)d。例如，在 BM 中，两个对象 O_i 和 O_j 不是相似就是不同，所以可以定义一个二进制矩阵，若 O_i

和O_j是相同的对象，则$d(O_i,O_j)=0$；若O_i和O_j是不同的对象，则$d(O_i,O_j)=1$。显然，这一思路可以扩展到有更多复杂对象的情况，即对象具有更多的属性或者更复杂的距离。n 维对象可认为是R^n空间的点，$d(O_i,O_j)$表示对象间的距离。

Lumer 和 Faieta 的 LF 算法将属性空间投影到一些低维空间，如二维空间，并且使得聚类具有聚内距离小于聚间距离的特性。

LF 算法沿用了基本模型，其相似度函数定义如下：

$$f(O_i)=\begin{cases}\dfrac{1}{s^2}\displaystyle\sum_{O_j\in Neigh_{s\times s}(r)}\left[1-\dfrac{d(O_i,O_j)}{\omega}\right] & if \quad f>0,\\ 0 & otherwise\end{cases} \tag{4-8}$$

其中，$f(O_i)$是对象O_i与出现在它邻近范围内的其它对象O_j的平均相似度，对应基本模型中的 f，Neigh(r)表示局部环境,在两维网格环境中通常表示以 r 为半径的圆形区域，s 表示邻近范围的半径，$d(O_i,O_j)$为两对象间的距离，参数ω为群体相似性系数，它是群体相似度测量的关键系数,它直接影响聚类中心的个数,同时也影响聚类算法的收敛速度。

概率转换函数是将群体相似度转换为简单个体移动待聚类模式(对象)概率的函数。它是以群体相似度为变量的函数，此函数的值域是[0，1]。同时概率转换函数也可称为概率转换曲线。它通常是两条相对的曲线，分别对应模式拾起转换概率p_p和模式放下转换概率p_d。概率转换函数制定的主要原则是群体相似度越大，模式拾起转换概率越小，群体相似度越小，模式拾起转换概率越大，而模式放下转换概率遵循大致相反的规律。拾起概率p_p和放下概率p_d的计算公式分别定义为：

$$p_p(O_i)=\left(\frac{k_1}{k_1+f(O_i)}\right)^2 \tag{4-9}$$

$$p_d(O_i)=\begin{cases}2f(O_i), & if \quad f(O_i)<k_2\\ 1, & if \quad f(O_i)\geq k_2\end{cases} \tag{4-10}$$

为了改进 BM 模型的性能，他们在系统上增加了三个特性：

(l)蚂蚁具有不同的移动速度，设定蚂蚁的速度 v 均匀分布在[1，V_{max}] 之

间。速度 V，通过修正公式 4-11 所示的相似度函数 $f(O_i)$ 决定蚂蚁是拾起一个对象还是放下一个对象；

$$f(O_i)=\max\left\{0,\frac{1}{s^2}\sum_{O_j\in Neigh_{s\times s}(r)}\left[1-\frac{d(O_i,O_j)}{\omega(1+((v-1)/v_{\max}))}\right]\right\} \tag{4-11}$$

(2)蚂蚁具有一个短时间的记忆；

(3)行为转换，如果在一个给定的时间内没有进行任何拾起或者放下的行动，蚂蚁能够消除这些聚类中心。

这些特性在减少相同的聚类中心、避免局部非优化结构等方面对原模型进行了改进。

LF 聚类算法的主要思想是将待聚类对象随机分布在一个环境中(一般是一个两维网格)，简单个体如蚂蚁测量当前对象在局部环境的群体相似度，并通过概率转换函数得到拾起或放下对象的概率，以这个概率拾起或放下对象，这样经过群体大量的相互作用，最终得到若干聚类。

在基本模型与 LF 算法中，采用了相似的概率转换函数。后来的研究者对此进行了许多改进。吴斌提出的基于 LF 算法的改进算法-CSI 算法中提出了一种比基本模型简单的概率转换函数[23]，该算法将 Deneubourg 提出的一种二次曲线简化成斜率为 k 的直线，其定义如下：

$$p_p=\begin{cases}1-\varepsilon & f(O_i)\le 0\\ 1-k\times f(O_i) & 0<f(O_i)\le 1/k\\ 0+\varepsilon & f(O_i)>1/k\end{cases} \tag{4-12}$$

$$p_d=\begin{cases}1-\varepsilon & f(O_i)\ge 1/k\\ k\times f(O_i) & 0<f(O_i)<1/k\\ 0+\varepsilon & f(O_i)\le 0\end{cases} \tag{4-13}$$

其中 ε 是一个很小的数，以便于算法的收敛，群体相似度函数 $f(o_i)$ 采用基本的测量公式， 其定义如下：

$$f(o_i)=\sum_{O_j\in Neigh(r)}\left[1-\frac{d(o_i,o_j)}{\omega}\right] \tag{4-14}$$

在基本模型中，概率转换函数的参数包括两个阈值，常数 k_1 和 k_2，而且阈值常数 k_1 和 k_2 的选取与数据密切相关，而简化后的概率转换函数，只有一个参数 k 并且，k 不随着实验数据的变化而变化，因此简化后的概率转换函数减轻了算法参数选取的复杂度，从而提高了算法的实用性。

第 5 章 聚类组合算法

5.1 基于信息素的 k-means 算法

5.1.1 k-means 算法简介

k-means 算法是一种基于划分的聚类方法，也是最常用和最知名的聚类算法。基于划分的聚类算法描述为：已知 d-维空间 R^d，在 R^d 中定义一个评价函数 $c:\{X:X\subseteq S\}\rightarrow R^+$ 给每个聚类一个量化的评价，输入 R^d 中的对象集合 S 和一个整数 k，要求输出 S 的一个划分：S_1，S_2，…，S_k，这个划分使得 $\sum_{i=1}^{k} c(S_i)$ 最小化。不同的评价函数将产生不同的聚类结果，最常用的评价函数定义如下：

$$c(S_i)=\sum_{r=1}^{|S_i|}\sum_{s=1}^{|S_i|}(d(x_r^i,x_s^i))^2 \tag{5-1}$$

其中，S_i 为划分形成的簇，x_r^i、x_s^i 分别为 S_i 的第 r 个和第 s 个元素，$|S_i|$ 表示簇中的元素个数，$d(x_r^i,x_s^i)$ 为 x_r^i 和 x_s^i 的距离。

k-means 算法不断计算每个聚类 S_i 的中心，也就是聚类 S_i 中对象的平均值，作为新的聚类种子。实际使用的评价函数为：

$$c(S_i)=\sum_{r=1}^{|S_i|}(d(\vec{x}^i,x_r^i))^2 \tag{5-2}$$

其中，$\vec{x}^i$ 为 S_i 的中心，其它符号含义同公式(5-1)。利用(5-2)式能够产生和利用(5-1)式同样的聚类结果。

k-means 算法具体描述如下：

(1)按一定原则选择 k 个对象作为初始的聚类种子；

(2)重复执行(3)、(4)两步，直到各个簇不再发生变化；

(3)根据聚类种子的值，将每个对象重新赋给最相似的簇；

(4)更新聚类种子，即重新计算每个簇中对象的平均值，用对象均值点

作为新的聚类种子。

k-means 算法除生成 k 个聚类外，还生成每个聚类的中心。k-means 算法具有较好的可伸缩性和很高的效率，当结果簇密集并且各簇之间的区别明显时，采用 k-means 算法的效果较好。

5.1.2 基于信息素的 k-means

蚂蚁觅食的过程，是以信息素来决定蚂蚁的运动方向。借鉴这一原理，将数据视为具有不同属性的蚂蚁，聚类中心看作是蚂蚁所要寻找的“食物源”，所以，数据聚类过程就看作是蚂蚁寻找食物源的过程。算法的思想是将蚂蚁从食物源 i 到食物源 j 的转移概率 P_{ij} 引入 k-means 中，根据概率来决定数据的归属；在下次的循环中，更新聚类中心，计算聚类的偏差，再次判断，直至偏差无变化，算法结束。k-means 算法是以距离为判断的标准来进行聚类的，改进算法中是以转移概率为标准进行聚类。改进的算法具体描述如下：

设 X 是待进行聚类分析的数据集合，

令：$X=\{X_i|X_i=(x_{i1},x_{i2},x_{i3},\ldots,x_{in}), i=1,2,\ldots, n\}$ 是待聚类的数据集合。

令：
$$d_{ij}=\|X_i-C_j\|=\sqrt{\sum_{r=1}^{n}(x_{ir}-c_{jr})^2} \tag{5-3}$$

其中 C_j 表示聚类中心的坐标，初始值为任意分配不相同的数据值。d_{ij} 表示 X_i 到 C_j 之间的欧氏距离，设 R 为聚类半径，ε 为统计误差，$\tau_{ij}(t)$ 是 t 时刻蚂蚁 X_i 到聚类中心 C_j 路径上残留的信息素，设 $\tau_{ij}(0)=0$，即在初始时刻各条路径上的信息量相等且为 0。$L(i,j)$ 表示蚂蚁从 i 到食物源(聚类中心)j 的路径矢量。则 $L(i,j)$ 上的信息素由下式给出：

$$\tau_{ij}(t)=\begin{cases}1, d_{ij}\le R\\ 0, d_{ij}>R\end{cases} \tag{5-4}$$

判断数据 X_i 是否归并到 C_j，由式 5-5 给出：

$$p_{ij}(t)=\frac{\tau_{ij}^{\alpha}(t)\eta_{ij}^{\beta}(t)}{\sum_{s\in S}\tau_{sj}^{\alpha}(t)\eta_{sj}^{\beta}(t)} \tag{5-5}$$

其中S={X_s| $d_{sj} \le R$,s=1,2,…,n}，若 $p_{ij}(t) \ge p_0$（p_0为一设定值），则 X_i归并到C_j。令CS_j={X_k|$d_{kj} \le R$, k=1,2,…,J}，CS_j表示所有归并到C_j邻域的数据集合，J 为 C_j邻域的数据个数。根据下式求出理想的聚类中心：

$$\overline{C}_J = \frac{1}{J}\sum_{k=1}^{J} X_k \tag{5-6}$$

其中：$X_k \in CS_J$。

聚类的偏离误差可由式 5-7，5-8 计算：

$$\varepsilon_J = \frac{1}{J}\sum_{i=1}^{J}(X_i - \overline{C}_J) \tag{5-7}$$

$$\varepsilon = \sum_{J=1}^{k} \varepsilon_J \qquad (i=1, 2, \ldots, k) \tag{5-8}$$

其中ε_J为第 J 个聚类的偏离误差，ε为所有聚类的总的偏差。

算法流程图如图 5-1 所示：

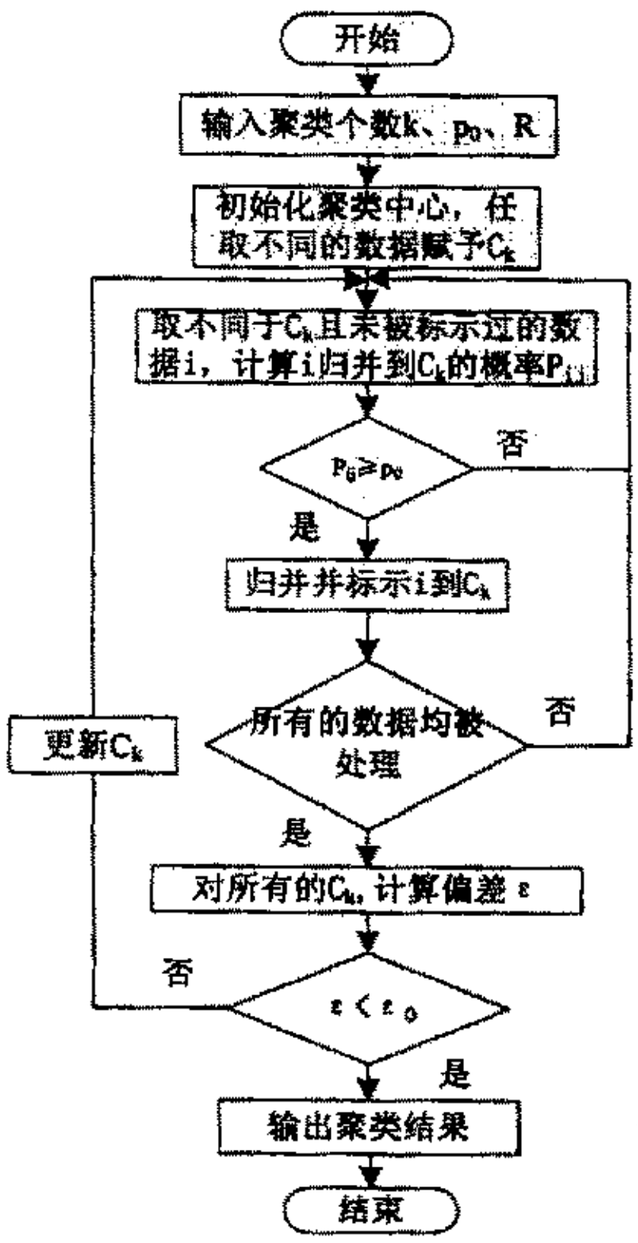

图 5-1 基于信息素的 k-means 算法流程图

5.2 基于阈值的聚类组合算法

一般在一个数据空间中，高密度的对象区域被低密度(稀疏)的对象区域(通常就认为是噪声数据)所分割。DBSCAN 算法是一种基于密度的聚类算法，它利用类的密度连通特性，可以快速发现任意形状的类。其基本思想是:对于一个类中的每一对象，在其给定半径的邻域中包含的对象不能少于某一给定的最小数目。在 DBSCAN 中，发现一个类的过程是基于这样的事实：一个类能够被其中的任意一个核心对象所确定。

5.2.1DBSCAN 算法

DBSCAN 是一种基于密度的算法，它利用类的高密度连通性，快速发现任意形状的类。发现一个类的过程是基于这样的事实：一个类能够被其中的任意一个核心对象所确定[78]。其基本思想是：对于一个类中的每个对象，在其给定半径的领域中包含的对象不能少于某一给定的最小数目。DBSCAN 为了发现一个类，先从数据库对象集 D 中找到任意一对象 P，并查找 D 中关于 R 和 minPts 的从 P 密度可达的所有对象(其中 R 为半径，minPts 为最小对象数)。如果 P 是核心对象，也就是说，半径为 R 的 P 的领域中包含的对象不少于 minPts，则根据算法，可以找到一个关于参数 R 和 minPts 的类。如果 P 是一个边界点，则半径为 R 的 P 领域包含的对象数小于 minPts，即没有对象从 P 密度可达，P 被暂时标注为噪声点，然后，DBSCAN 处理 D 中的下一个对象。

5.2.2 T-Value 算法

受 DBSCAN 算法的启发，提出一种基于阈值的聚类方法(T-Value)。该方法能够帮助发现具有任意形状的聚类。算法要求输入合适的距离阈值 T，该算法就能聚出较好的类，使得类间所有数据都大于阈值 T，而类内则都小于阈值 T，类的个数也会自动生成。首先给出如下三个定义：

定义5.1：p维数据x与y的相似性是由欧式距离来定义的，其定义如下：

$$d(x,y)=\sqrt{\sum_{i=1}^{p}(x_i-y_i)^2} \tag{5-9}$$

定义 5.2：假设 S 是 p 维数据集合，阈值 T 为一实数，基于阈值的聚类问题就是把数据集 S 分成 k 个类 C_1、C_2、...、C_k，使得：

$$\forall z_1,z_2,\ldots,z_n \in C_i,\ z_i \neq z_{i+1},\ \text{则 } d(z_i,z_{i+1})<T$$

其中$\cup_i C_i = S$, and $C_i \cap C_j = \varnothing$ $(i \neq j)$。

定义 5.3：定义一 NxN 关系矩阵 A[N][N]，对于数据集中的两个点 i 和 j，如果 d(i,j)<T，则 A[i][j]=1；否则 A[i][j]= 0。

该算法首先要把p维的数据投影至二维平面。具体做法就是：计算两个不同数据i，j之间的距离dist(i,j)(欧式距离)，与阈值T进行比较，如果小于阈值，则关系数组A[i][j]=1；否则A[i][j]=0。最终生成如图表5-1所示的只有0和1的二元关系矩阵。在聚类分析时，只对关系矩阵进行分析，而不用考虑模式的维数和数值，这样能够简化了算法流程和复杂性。

算法思想是任取未标记归类的任一数据i，判断数据i与数据j的关系数组值是否为1，如果是则归并i、j为第一个类，标示i、j为第一类。通过递归收集完相邻的所有数据归为一类，然后再取另外一个未被标示过的数据，依上述方法找出与其相似的数据归为第二类，依次类推完成所有的数据的归类。

表5-1 数据样本的二元数据矩阵

j / i	i1	i2	…	in
i1	1	0	…	1
i2	0	1	…	0
…	…	…	…	…
in	1	0	…	1

5.2.3 基于 e-邻域的 T-Value 算法

考虑到具有大量分散点的数据集，T-Value算法的聚类效果就不是很好，因此在T-Value算法的基础上引入e-邻域的概念。首先给出两个定义：

定义 5.4：一个给定对象的 e 半径内的近邻就称为该对象的 e-邻域；

定义 5.5：若一个对象的 e-邻域至少包含一定数目(MinPts)的对象，该对象就称为核对象；否则称为散列点。如图 5-2 所示，假设 MinPts=3，则 M、P、O 为核心对象，A 和 B 为散列点。

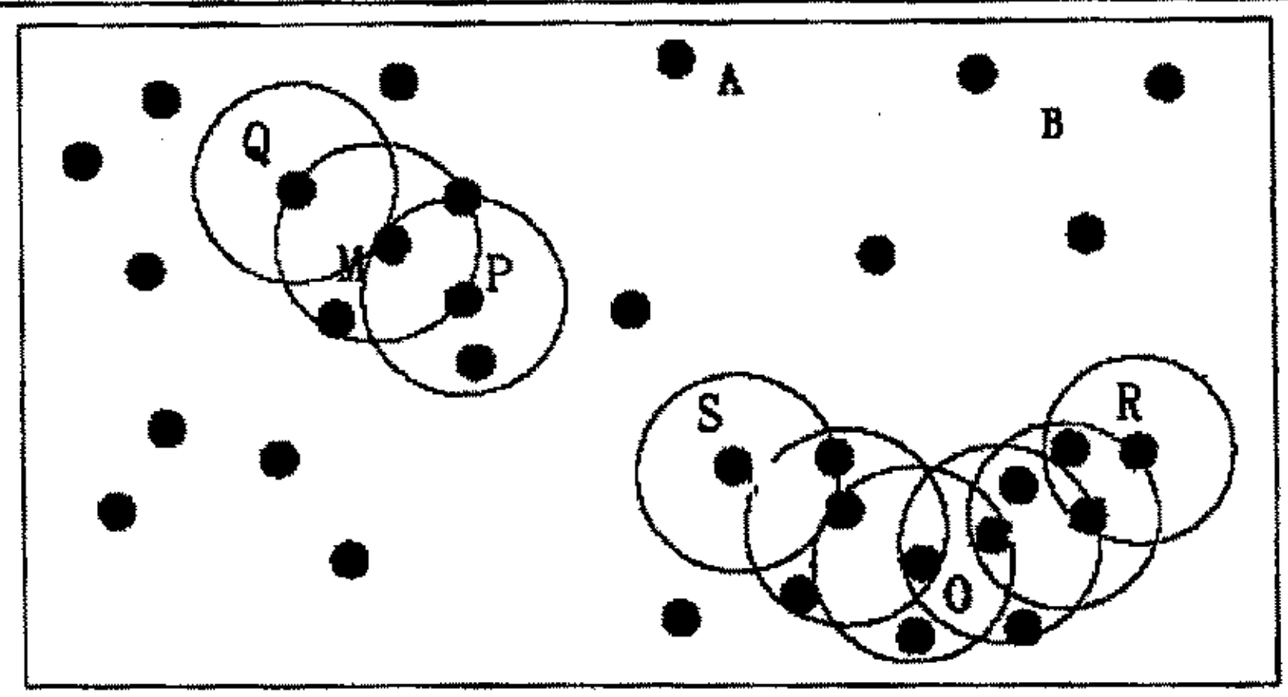

图 5-2 核心对象示意图

对于如 5-2 所示的图中的数据样本，基于改进的 T-Value 的思想是只对核心对象进行聚类处理。首先进行标识散列点，然后把所有的核心对象进行 T-Value 处理。

基于 e-邻域的 T-Value 算法流程：

输入:N×p数据矩阵，距离阈值T，

输出：聚类和聚类个数

算法描述：

1. 输入N个p维数据，阈值T, k=0,
2. 对所有的数据，计算其 ε 邻域内的数据样本个数，如果小MinPts, 则标记该数据为散列点，并统计散列点的个数outlier,outlier<N，于是核心对象个数为n= N-outlier；
3. 计算所有未被标识为散列点的数据样本间的距离，
   如果i与j的距离小于阈值T，则置A[i][j]等于1
   否则等于0，最终生成n×n的数据矩阵
4. 取核心对象中的任一未被标识归类的数据样本i，归为类C_k
   for j = 1 to n
   if (A[i][j]=1)
   $C_k = C_k \cup \{j\}$，标示j为已归类，然后利用递归，把所有关系矩阵中与j的关系值为1的数据样本加入到聚类C_k中；
5. 如果所有的数据都被标识过，则跳到步骤6
   否则，k=k+1，重复步骤4
6. 计算并输出结果$\{C_1,C_2,\ldots,C_k\}$

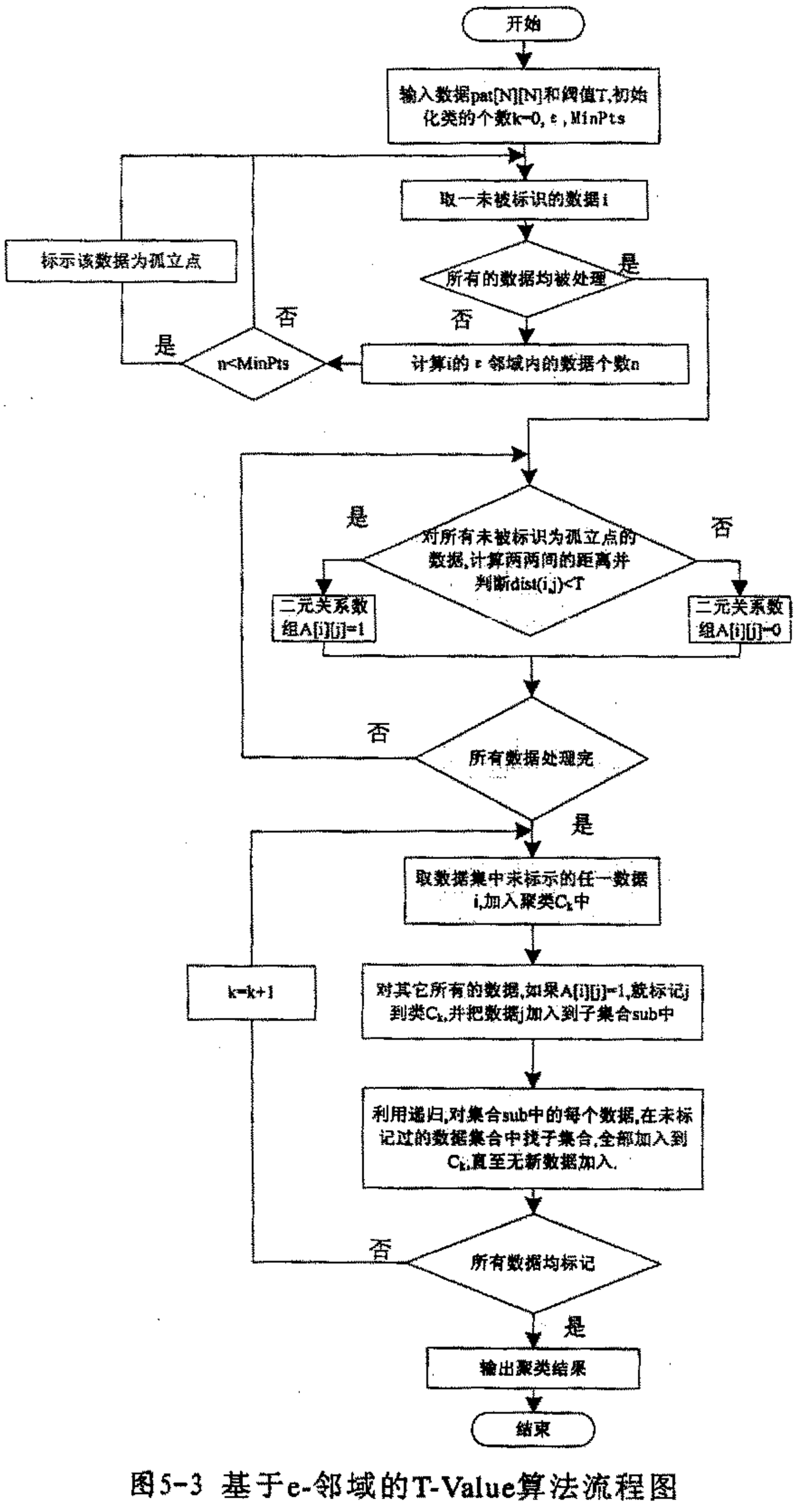

图5-3 基于e-邻域的T-Value算法流程图

5.2.4 T-Value 聚类组合算法

由上分析可知，基于 e-邻域的 T-Value 算法能够自动生成聚类，但仍存在许多散列点没有加入到聚类中。因此提出一种组合算法，先由 T-Value 算法进行处理，生成聚类个数和聚类中心，然后利用基于信息素的 k-means 算法进行二次归类，最后输出聚类结果。

该算法首先把p维的数据投影至二维平面。具体做法就是：首先对每个数据寻找其 ε 邻域内的数据个数，如果小于MinPts，则标识为散列点。其次计算两个不为散列点的不同数据i，j之间的距离dist(i,j)(欧式距离)，与阈值T进行比较，如果小于阈值，则关系数组A[i][j]=1；否则A[i][j]=0。最终生成如图表5-1所示形式的只有0和1的二元关系矩阵。然后任取关系矩阵A中的一数据i，判断数据i与数据j的关系数组值是否为1，如果是则归并i、j为一个类，标示为第一类。递归收集完相邻的所有数据加入类中，然后再取另外一个未被标示过的数据，依上述方法找出与其相似的数据加入第二类，依次类推完成所有的数据的处理。把所产生的聚类个数和聚类中心作为输入,对所有散列点，利用改进的基于信息素的k-means的方法加入到已有的聚类中。

组合算法的流程如图 5-4 所示

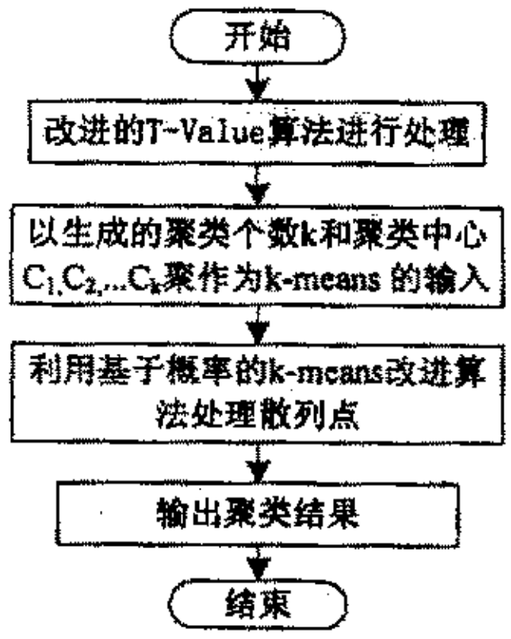

图5-4 聚类组合算法的流程图

5.3 基于蚁群算法的聚类组合算法

5.3.1 算法的思想

聚类组合算法的描述可以看成是基于蚁群聚类算法和基于信息素的 k-means 算法的组合。基于蚁群算法的聚类方法(Clustering based on Ant Colony

Algorithm)来源于 Deneubourg 提出的基本模型和 Lulner 和 Faieta 的数据分析 LF 算法，借鉴 CSI 的概率计算公式 4-12 和 4-13 计算蚂蚁拾起和放下数据的概率。基于蚁群算法的聚类算法是一种自组织聚类算法，具备健壮性、可视化等特点，并能生成一些新的有意义的聚类模式。但是由于算法有时收敛时间较长，而且常常出现一些由一个模式组成的聚类中心。虽然这些模式可以用于孤立点(outlier)分析，但是对于一般要求的聚类，聚类中心过多、过散并没有益处。因此，本章在基于蚁群聚类算法的基础上，结合基于信息素的 k-means 算法，提出了一种聚类组合算法。该算法可分为如下二个阶段进行：

1. 基于改进的 LF 算法的聚类过程；

此阶段又可分为两步进行，第一步通过蚂蚁的随机运载，得到初始聚类。这一步骤用 4-12 和 4-13 的概率计算公式，计算蚂蚁拾起或放下数据样本的概率，进行训练，最终得到较为合理的聚类。第二步是对数据进行收集和标识。这一步借鉴了 SACA[25]的收集和标示数据样本的方法，这种方法是对样本在二维平面上的一区域进行搜索，然后利用递归收集完与样本相关的所有数据，这种方法具有较小的搜索盲目性，能够有效提高算法的效率。对于标示为孤立点的数据样本，不再列为下一阶段的初始聚类中心模板。

2. 基于信息素的 k-means 算法的聚类优化过程

以第一阶段得到的聚类中心均值和聚类个数为参数，把数据集作为蚂蚁，把聚类中心作为食物源，进行后处理(Post-processing)。该阶段具体流程可描述为：基于蚁群算法的聚类分析中，将数据视为具有不同属性的蚂蚁，聚类中心看作是蚂蚁所要寻找的“食物源”。所以，数据聚类过程就可看作是蚂蚁寻找食物源的过程。因此利用基于信息素的 k-means 算法进行后处理，把所有未归类的数据样本和孤立点进行二次聚类。

5.3.2 算法描述

5.3.2.1 算法主流程图

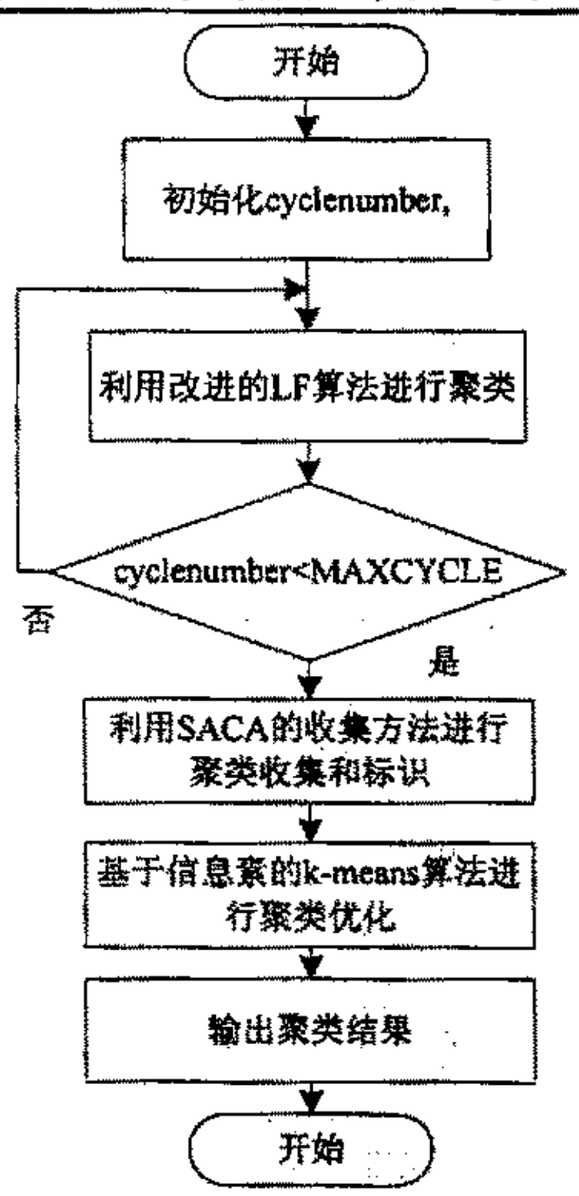

图 5-5 基于蚁群算法的聚类组合算法的主流程图

5.3.2.2 算法的流程描述

(1)参数初始化，预设的循环次数 cycle_number，蚂蚁个数 ant_number，a，ß，方形区域半径 R，聚类的最小个数 minNumer。

(2)将待聚类对象随机分布于平面上，即随机赋予每个对象一对坐标 (a_i,b_i)，其中 $a_i,b_i \in [Xwight,Yweight]$ $(i=1,2,...,n)$。

(3)给一组蚂蚁设置初始状态值，初始状态为无负载。

(4)随机选取一只蚂蚁，以其初始模式对应坐标为中心，R 为观察半径，利用公式 4-14 计算此模式在观察半径范围内的群体相似度。

(5)若本只蚂蚁无负载，则用公式 4-12 计算拾起概率 p_p：若 $p_p < p_p^{'}$ ($p_p^{'}$ 为一随机概率)，则蚂蚁不拾起此对象，再随机赋给蚂蚁一个模式值，否则蚂蚁拾起此对象，蚂蚁状态改为有负载，随机给蚂蚁赋予一个新坐标。

(6)若本只蚂蚁有负载，则用公式 4-13 计算放下概率：若 $p_d > p_d^{'}$ (为一随机概率)，则蚂蚁放下此对象，将蚂蚁的坐标赋给此对象，蚂蚁状态改为无负载，再随机赋给蚂蚁一个模式值，否则蚂蚁继续携带此对象，蚂蚁状态仍为

有负载，再次随机给蚂蚁赋予一个新坐标。

(7)借鉴SACA标识数据样本的方法进行数据标识。若数据样本未被标注类别，则标注此对象的类别。即如果数据样本是孤立的或者它的邻域数据样本个数小于minNumber则标识该数据为孤立点，否则给该对象分配一个聚类序列号，并递归地将其邻域对象标记为同样的聚类序列号。

(8)利用基于信息素的 k-means 算法进行聚类优化

(a)生成聚类中心模板，即计算不包括孤立点的每一个聚类中心的平均值；

(b)利用基于信息素的 k-means 算法的概率计算公式 5-5，计算每一个孤立数据样本 i 归并到到聚类中心 j 的概率 p_{ij}，若 $p_{ij} \geq p_0$(其中 p_0 为设定数据)，则将数据划分到所属的聚类中心；

(c)计算每个聚类中的偏离误差 ε_j 和总误差 ε，并更新聚类中心；

(d)直至误差 $\varepsilon \leq \varepsilon_0$，程序结束。

可以看出(1)~(3)是算法初始化阶段，其主要作用是程序初始化和在平面上随机分布对象；(4)~(6)是基于 LF 算法的聚类过程；(7)是模式类别标注过程，也就是聚类结果收集过程；(8)基于信息素的 k-means 聚类过程，以上面的聚类结果为初始条件。

5.4 本章小节

本章首先介绍经典的k-means算法，把蚂蚁觅食的原理引入k-means算法中，提出了一种基于信息素的k-means改进算法。在研究了基于密度的DBSCAN算法的基础上，提出了一种基于阈值的T-Value聚类算法，由于对于具有大量散列点的数据集，T-Value算法无法体现它的优势，因此引入了 ε -邻域的概念，结合基于信息素的k-means算法，提出基于T-Value聚类组合算法。而后研究了LF算法及其相关的改进，提出了一种基于LF和基于信息素的蚁群聚类组合算法。在下一章将采用数据集进行算法分析。

第 6 章 测试与性能评价

6.1 测试数据集

测试数据取自 UCI 机器学习数据库，选取两组测试数据集：鸢草(IRIS)和酒(Wine)。表 6-1 列出了各数据集的记录数、条件属性个数及类别数。

表 6-1 测试数据集

数据集名称	记录个数	属性个数	聚类个数
Iris	150	4	3
Wine	178	13	3

6.2 性能评价

一般来说，质量评价方法分为外部和内部两种，其依据是有无关于数据集的先验知识。首先介绍一种常用的外部评价方法：F-measure 方法。

F-Measure 组合了信息检索中查准率(precision)与查全率(recall)的思想，是信息检索领域的一种系统性能测试指标。F-Measure 参数由 Van Rijsbergen 于 1979 年提出，当前，它主要用于信息的检索，是搜索引擎性能评价的重要标准[52]。

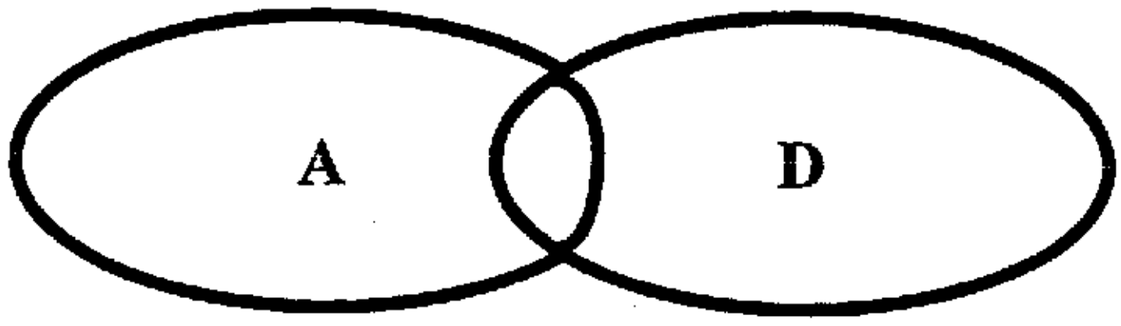

图 6-1 返回结果集和期望结果集图示

如图 6-1 所示，A 代表一次搜索返回的所有结果集，D 代表所有期望得到的结果集。查准率 p(precision)是指已查出的有效结果占所有搜索到结果的百分比。分别定义如下：

$$p = p(D/A) = \frac{P(A \cap D)}{P(A)}$$

$$r = p(A/D) = \frac{P(A \cap D)}{P(D)}$$

这两个参数反映了搜索引擎的检索效果，按照 F-Measure 参数对其属性的定义，我们可以将这两个参数结合成一种简单的标量尺度 f，定义如下：

$$f=\frac{2\cdot p\cdot r}{p+r}=\frac{2\cdot p(A/D)\cdot p(D/A)}{p(A/D)+p(D/A)}=\frac{2\cdot P(A\cap D)}{P(A)+P(D)}$$

把 F-Measure 引入聚类分析[25]，就是组合了信息检索中查准率(precision)与查全率(recall)的思想。一个聚类 j 及与此相关的分类 i 的 precision 与 recall 定义为：

$$P=precision(i,j)=\frac{N_{ij}}{N_i}$$

$$R=recall(i,j)=\frac{N_{ij}}{N_j}$$

其中 N_{ij} 是在聚类 j 中分类 i 的数目，N_j 是聚类 j 中所有对象的数目，N_i 是分类 i 中所有对象的数目。则分类 i 的 F-measure 定义为：

$$F(i)=\frac{2PR}{P+R}=\frac{2\cdot N_{ij}}{N_i+N_j}$$

对分类 i 而言，哪个聚类的 F-measure 值高，就认为该聚类代表分类 i 的映射，换句话说，F-measure 可看成分类 i 的评判分值。对聚类结果 p 来说，其总 F-measure 可由每个分类 i 的 F-measure 加权平均得到：

$$F_p=\frac{\sum_i(|i|\times F(i))}{\sum_i|i|} \tag{6-1}$$

其中其中|i|为分类 i 中所有对象的数目。

6.3 试验结果分析

本节对各种算法用取自 UCI 的数据集进行测试，记录其测试的聚类结果。

1．k-means 算法：

1）对 Iris 数据集，取 k=3 进行聚类，其运行时间 t=18ms。

表 6-2 Iris 数据集 k-means 算法测试结果

类别	聚类中的样本个数	正确的个数	不正确的个数
第 1 类	50	50	0
第 2 类	62	48	14
第 3 类	38	36	2

2）对 Wine 数据集，取 k=3 进行聚类，其运行时间 t=42ms。

表 6-3 Wine 数据集 k-means 算法测试结果

类别	聚类中的样本个数	正确的个数	不正确的个数
第 1 类	43	34	9
第 2 类	101	64	37
第 3 类	34	23	8

2. 基于信息素的 k-means 算法：

1）对 Iris 数据集，取 k=3 进行聚类，其运行时间 t=18 ms。

表 6-4 Iris 数据集基于信息素的 k-means 算法测试结果

类别	聚类中的样本个数	正确的个数	不正确的个数
第 1 类	50	50	0
第 2 类	58	46	12
第 3 类	42	38	4

2) 对 Wine 数据集，取 k=3 进行聚类，其运行时间 t=40ms。

表 6-5 Wine 数据集基于信息素的 k-means 算法测试结果

类别	聚类中的样本个数	正确的个数	不正确的个数
第 1 类	45	36	9
第 2 类	95	60	35
第 3 类	38	25	13

3. T-Value 聚类组合算法：

对 Iris 数据集，取 MinPts=2，分别取 $\varepsilon = T = 0.4$、0.5、0.57、0.6、0.7 进行聚类；对 Wine 数据集，MinPts=4，分别取 $\varepsilon = T = 30$、33、34.7、36、39，进行聚类。它们的总 F-measure 测量值如表 6-6 所示，图 6-2、6-3 是 T 取不同值时的总 F-measure 测量值图示。

表 6-6 Iris 和 Wine 数据集的 T-Value 聚类组合算法的聚类测试结果

数据集	阈值 T	总 F-measure 测量值
Iris	0.4	0.431
	0.5	0.681
	0.57	0.914
	0.6	0.751
	0.7	0.523
Wine	30	0.408
	33	0.487
	34.7	0.734
	36	0.582
	39	0.420

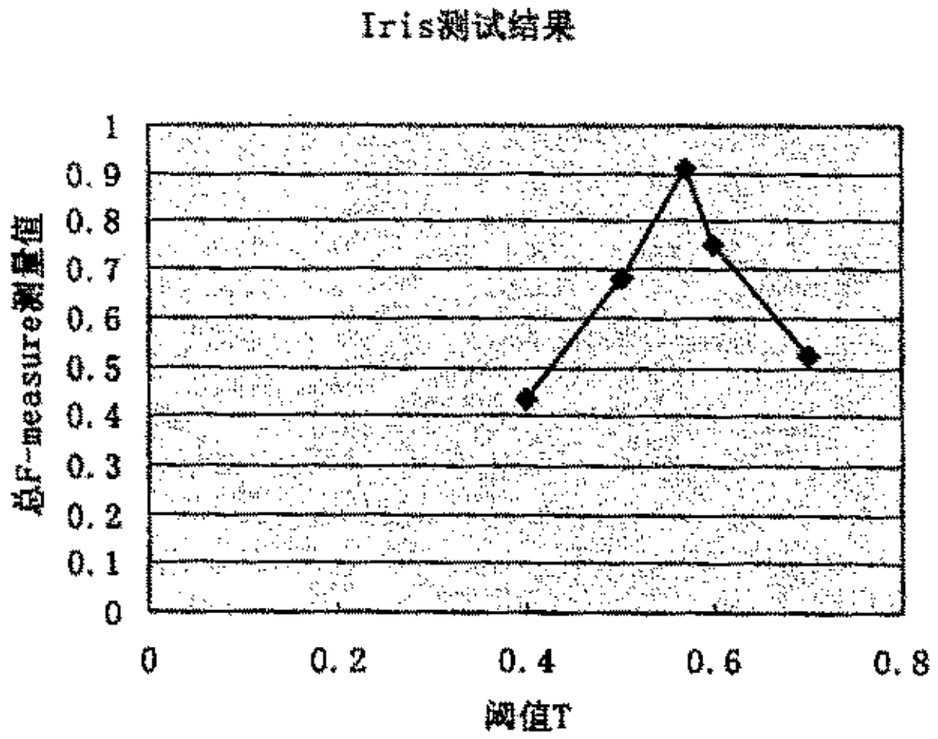

图 6-2 Iris 数据集 T-Value 聚类组合算法测试结果图

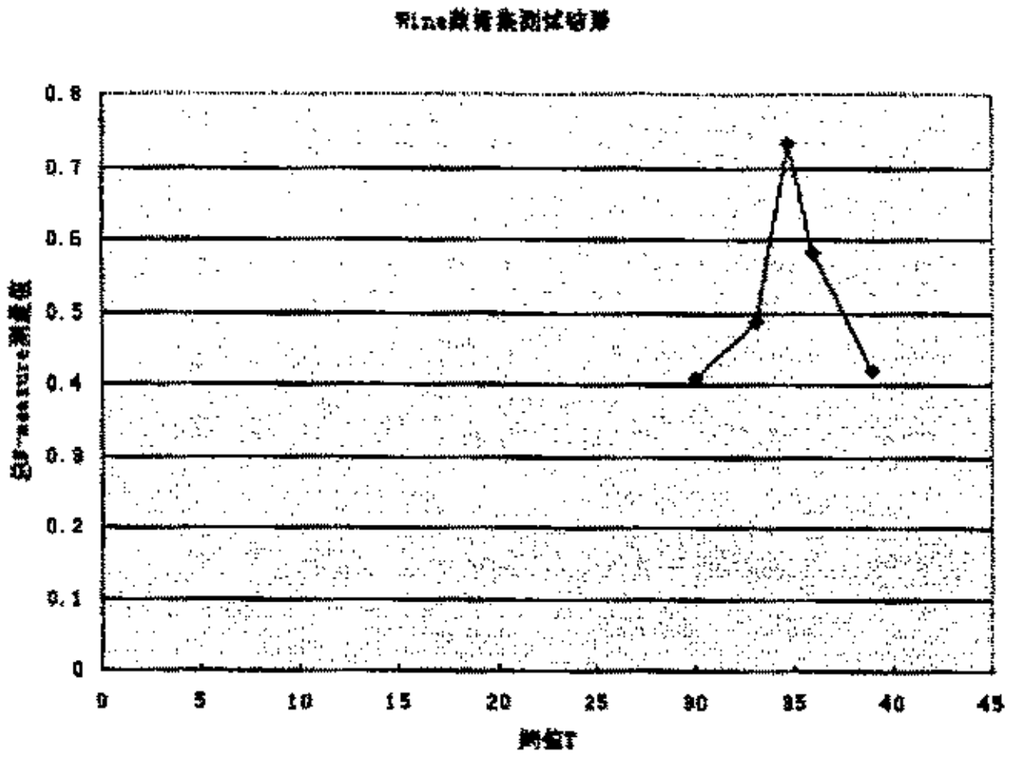

图 6-3 Wine 数据集 T-Value 聚类组合算法测试结果图

由图 6-2、6-3 可以看出，对 Iris 数据集，当 T=0.57 时，T-Value 聚类组合算法具有最大的总 F-measure 测量值；对 Wine 数据集，当 T=34.7 时，T-Value 聚类组合算法具有最大的总 F-measure 测量值；

4. 蚁群聚类组合算法：

1）对 Iris 测试数据集，取 cycle_number=1000，antnumber=30，R=3，a=1，ß=1，其运行时间 t=3341ms。

表 6-7 Iris 数据集蚁群聚类组合算法测试

类别	聚类中的样本个数	正确的个数	不正确的个数
第 1 类	50	50	0
第 2 类	53	48	3
第 3 类	47	45	2

2）对 Wine 测试数据集，取 cycle_number=1500，antnumber=20，R=3，a=1，ß=1，运行时间 t=3975ms。

表 6-8 Wine 数据集蚁群聚类组合算法测试

类别	聚类中的样本个数	正确的个数	不正确的个数
第 1 类	59	52	7
第 2 类	87	63	24
第 3 类	42	29	13

5. 几种算法的 F-measure 测量值比较

对 T-Value 聚类组合算法、k-means 算法、基于信息素的 k-means 改进算法和蚁群聚类组合算法,利用 F-measure 方法进行性能分析。利用公式 6-1 进行计算总 F-measure 值，结果如表 6-9 所示：

表 6-9 几种算法的总 F-measure 测量值比较

数据集	k-means	基于信息素的 k-means 算法	T-Value 聚类组合算法	基于蚁群算法的聚类组合算法
Iris	0.907	0.910	0.914	0.943
Wine	0.70	0.718	0.734	0.738

由上分析可知，相比 k-means 算法，基于信息素的 k-means 的聚类算法具有较高的总 F-measure 值，性能较优；基于 T-Value 的组合算法的总 F-measure 值要比前两种算法大一些，而且在运行时间和聚类质量上也比较高。因此该组合算法能够在一定程度上的改善聚类效果。

对蚁群聚类组合算法算法,利用 F-measure 方法进行性能分析。对 Iris 数据集其 F-measure 测量值为 0.943，对 wine 数据集则达到 0.738，因此是几种算法中效果最好的一个。但其运行时间要比前几种算法长。由以上分析，对比基于 T-Value 的组合算法、基于信息素的 k-means 算法，蚁群聚类组合算法，可以看出在计算时间上，比前两个算法都要长，但正确率要高。与基于 T-Value 的组合算法一样，蚁群聚合算法能够自动生成聚类个数。

6.4 客户行为中的聚类

6.4.1 客户行为分析

客户行为分析是客户关系管理中决策分析的一个重要部分。客户行为分析是通过对客户整体、特定客户群体(大客户等)进行分析，了解客户在自然社会特征、消费行为等方面的分布和随时间的变化趋势。客户关系管理中的客户行为分析可以划分为整体行为分析和群体行为分析。整体行为分析用来发现企业所有客户的行为规律，但仅有整体行为分析是不够的，企业的客户千差万别，根据客户行为的不同可以将他们划分为不同的群体，各个群体有着明显的行为特征。通过客户群体行为分析，CRM用户可以更好地理解客户，发现群体客户的行为规律。基于这些理解和规律，市场专家可以制定出相应的市场策略，同时还可以针对不同客户群进行交叉分析，帮助CRM用户发现客户群体间的变化规律。

在客户分析中，显然对大客户分析十分重要，另外，客户分类中还有一种极端情况，每个类别里的客户只有一个，即一对一营销(One To One)。一对一营销是指了解每一个客户，并与之建立起长期持久的关系。

目前国内一些行业如金融、电信行业已经开始启动并积极推广CRM项目，为客户提供更为优质的服务的同时，深度挖掘客户的潜在价值，从而为提供服务者带来更大的收益。而CRM服务的实现要求企业对客户的各种行为均有一定的数据资料积累，并要求企业具有一套行之有效的管理工具，使企业可以从这些数据资料中获得客户的分类，并提供对不同类型客户消费习惯的参考数据，使得一个企业可以了解自己的客户，了解自己的市场因素，从而制定有效的服务方案，做到有的放矢。同时，对客户数据的分析，也为企业的发展方向提供一个有力的决策依据。中国移动的BOSS系统,银行开展的数据大集中等都为客户关系管理的实施提供了充足的挖掘数据。

由于客户的消费能力、消费习惯、消费周期等诸方面都不尽相同。这便为企业的客户服务增加了许多不定的因素，要求企业将客户服务内容细分:针对不同的客户群，采取不同的服务策略，最大可能的为客户实现个性化、专业化服务。国外一些银行早就将客户分为一般客户(对银行没有特别要求)、年轻而富有的客户(要求快捷方便的服务)、成熟而富裕的富户(要求个性化的服务)，分别提供服务。国内工商银行等金融机构最近兴起的个性化服务，正是对这种先进经营理念的学习。当然，这种创新丝毫不会损害普通客户的利益。

6.4.2 实验分析

本节将聚类算法应用于对移动客户话费数据集进行的实验,将前面所述的聚类组合算法用于实现客户细分。数据取自于某省移动用户话费某月的消费大于100元的部分用户数据。话费包括基本月租、本地通话费、长途、信息费和漫游费，样本个数选择1000个。由于聚类组合算法具有的并行性、鲁棒性和自动生成聚类个数的特性，结合移动用户的消费数据进行分析，从中发现不同特征的客户群体，从而为移动公司推广新产品、提高客户的贡献率、最大化公司利润创造条件。

(1) 蚁群聚类算法进行处理

取cycle_number=2000，antnumber=20，R=3.5，a=1，ß=1，进行聚类。最终生成的聚类个数为11个。聚类结果的平均话费直方图如图6-4所示：

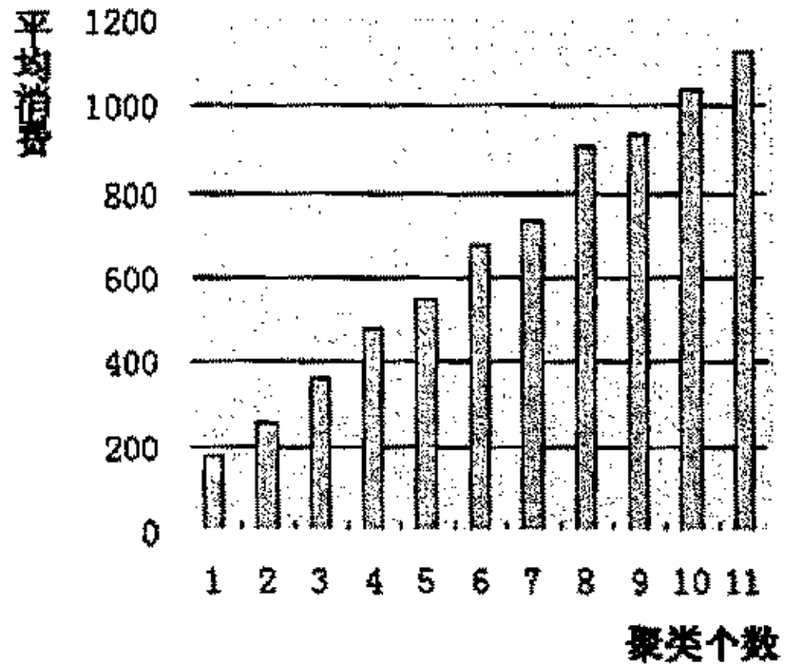

图6-4 蚁群聚类算法的平均话费聚类结果

(2) k-means 算法进行处理的结果：

取聚类个数 k=11，则 k-means 算法进行的聚类结果如图 6-5 所示：

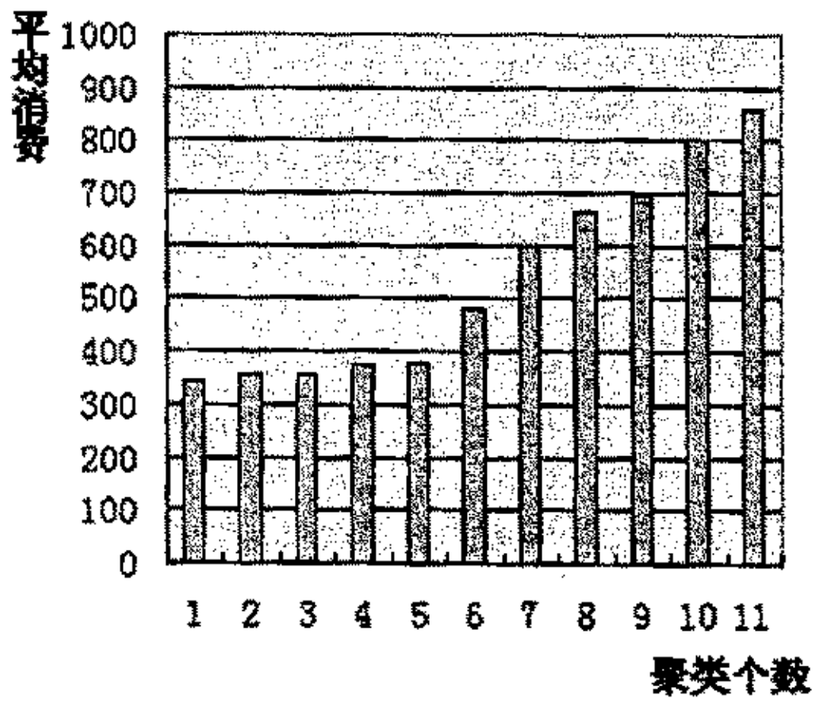

图6-5 k-means算法的平均话费聚类结果

(3) 基于信息素的 k-means 算法进行处理

取聚类个数 k=11，则 k-means 算法进行的聚类结果如图 6-6 所示：

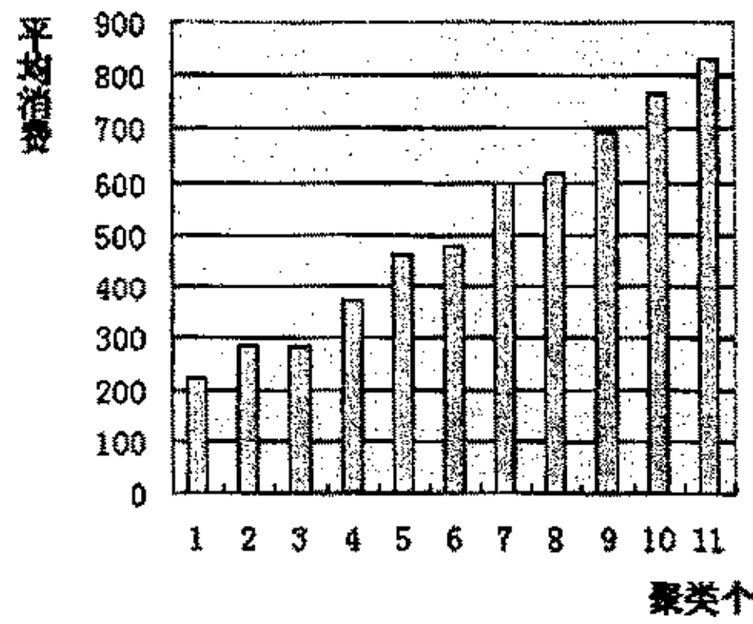

图6-6 基于信息素的k-means算法的平均话费聚类结果

(4) T-Value 聚类组合算法进行处理

取 T=57.7，MinPts=4，其聚类结果如图 6-7 所示：

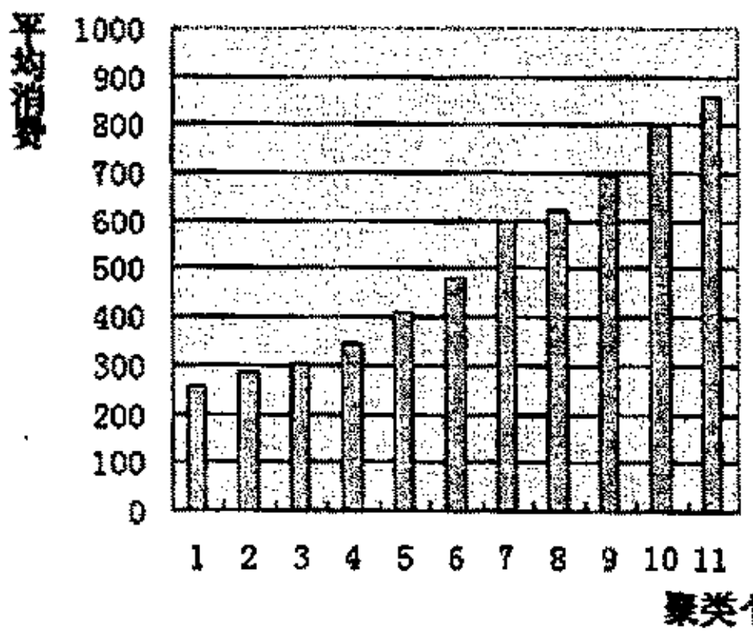

图6-7 T-Value聚类组合算法的平均话费聚类结果

由图6-4分析，根据客户通话的一些典型特征可将客户划分为普通客户、银卡客户、金卡客户、VIP大客户等几个等级。这些客户群体表现的特征也各不相同，如对于在1000元以上的客户，其数量是少的，但其通话频率比较高，一般为一些公司老总或从事高薪高通话率的工作，这些客户可看作为VIP大客户，根据其话费的消费，合理推出相应的移动产品；对于600至1000元之间的客户，其通话频繁，一般具有较高的职务， 从事高薪工作或政府高级官员等，这些客户可认为是金卡客户；对于400至600元之间的客户，认为其通话频率也是比较高，而且漫游和长途话费比较多，这些客户可看作银卡客户，可重点培养和发展，具有较大的升值潜力。大多数为一般普通客户群，其消费水平在100至400之间，该群体的较大的特征是通话费、信息费和其它费用较为平均，因此这类客户提供的效益也具有较大的提升空间，但同时也具有较大的流失性，因此移动公司

应当多为其提供较好的产品和服务，重点维持、防止客户流失。

比较图6-4、6-5、6-6、6-7可以看出，在聚类中心数目相近的情况下，基于蚁群的聚类算法取得的聚类中心层次与其它几种聚类算法相比更加清晰、聚类效果更好。对照详细的聚类结果，蚁群聚类组合算法不仅保证了聚内距离小和聚间距离大的良好聚类特性，而且形成了其它几种算法没有找到的特征突出的聚类中心。如在聚类中心个数相近的条件下，图6-4中话费平均值在800元以上的模式类别有第8、9、10、11类（共4 个），1000元以上的有两个类（第10、11类），图6-5、6-6、6-7中话费平均值在800元以上的模式类别分别只有第11类（共1个），1000元上的一个也没有。可见蚁群聚类组合算法对消费金额大的大客户分析得更细，这将有利于一对一营销中特别客户的分析，在此基础上再结合客户的其它自然属性如年龄、性别及收入水平等可以从多角度对聚类得出的客户群体进行分析。

6.4 本章小结

本章选用UCI机器学习库中的数据集进行几种算法的性能分析。与k-means相比较，其它几种算法均优于k-means，具有相对较好的效果和质量，而其中蚁群聚类算法具有最优的聚类性能，但所用的时间比较长。采用移动话费数据集进行试验，实验表明它能找出有特点的客户群体，特别是在大客户分析方面有一定的优势。通过群体相似系数可以调整算法收敛速度，但这可能导致聚类效果较差。在群体相似系数确定后，算法的结束条件是由最大循环次数决定的，这个值一般是通过观察大致确定的，因而如果能更准确地确定的算法收敛条件，算法运行时间将会减少。

结论

本文对作为数据挖掘技术之一的聚类分析作了较为深入的研究，结合蚁群算法和聚类的思想，提出一些自己的想法和改进：

1. 提出一种基于信息素的 k-means 改进算法。由蚁群觅食的习性联想到数据聚类，把数据样本作为蚁蚁，聚类中心作为食物源，通过转移概率来决定数据的归属。
2. 受 DBSCAN 算法的启发提出了基于阈值的 T-Value 算法，并结合基于信息素的 k-means 算法提出了一种基于阈值的 T-Value 聚类组合算法，该算法使用 UCI 机器学习库中的数据集进行测试和分析，分析表明，该算法具有较好的聚类效果。
3. 对由蚂蚁堆积幼卵和死尸的习性提出的 LF 算法和后来者的改进算法 CSI 和 SACA 进行研究，在总结前人的研究成果的基础上，结合受蚂蚁觅食原理的启发而改进的基于信息素的 k-means 算法，提出一种基于蚁群算法的聚类组合算法，并把算法用于 UCI 机器学习库的数据集和移动话费数据集进行分析，实验结果表明，该算法能够得到较好的聚类质量。

今后工作的展望：

1. 各种聚类算法的应用研究。聚类算法可用于很多领域，由于知识面和获取数据值的渠道有限，因此应用研究没有完全展开。
2. 蚁群算法用于路由算法的研究。由于时间仓促，并没有对蚁群路由算法进行深入的研究，在以后的工作中，将结合实际工作展开路由算法的研究。
3. 银行 CRM 中的聚类分析。由于未能获得银行方面的数据集，而未能进行银行 CRM 中的聚类分析研究。

致谢

首先要感谢我的导师杨燕副教授。杨老师对我一丝不苟的指导以及无微不至的关怀。像一盏明灯，在迷惘的时候，给我指明了前进的方向。她严谨的治学态度永远值得我学习。对我的悉心指点与鼎立相助。她对于科学执着追求的坚韧精神以及无私、热情的气质深深地感动着我，堪称楷模。在撰写硕士论文期间，杨老师更是以她新颖独特的思维方式、开拓创新的独到见解、温文尔雅的学者气质、诲人不倦的长者风范为我论文的开题、改进、完善和定稿付出了大量的精力。本论文的完成凝聚了杨老师的无数心血，在此向杨老师致以诚挚的谢意！并祝您及您的家人幸福健康！工作顺利！

由衷感谢戴齐副教授，戴老师不仅学业上给予了细心指导和谆谆教诲，而且在生活上、工作上都给了无微不至的关怀。戴老师以他博大的胸怀、严谨的治学态度和科研精神让我由衷的佩服。在此向戴老师致以衷心的感谢和敬意！祝戴老师工作顺利、全家幸福。

同时向帮助过我的所有老师和同学们表示衷心地感谢，祝你们工作顺利、生活顺心！

最后向我的朋友、我的父母以及我的所有亲人们表示由衷地感谢，感谢你们对我学业上、生活上的关心和支持！

参考文献

[1] 黄振华，吴诚一. 模式识别[M].浙江大学出版社, 1991：40-62

[2] 蔡元龙. 模式识别[M]. 西安：西北电讯工业出版社, 1986；17-32

[3] 张纪会，徐心和. 变异特征的蚁群算法[J]. 算机研究与发展, 1999；36(10): 1240-1245

[4] 马良，项培军. 蚂蚁算法在组合优化中的应用[J]. 管理科学学报, 2001；4(2): 32-37

[5] 谢维信. 工程模糊数学方法[M]. 西安电子科技大学出版社, 1991

[6] J.Han and M.Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann Publisher, 2001

[8] Michael J.Corey, Michael Abbey, Lan Abramson, and Ben Taub. Oracle 8 数据仓库分析、构建实用指南. 陈越,郭渊博等. 机械工业出版社, 2000

[9] Efrem G.Mallach. 决策支持与数据仓库系统. 李昭智，李昭勇等. 电子工业出版社, 2001

[10] P.Arabie, L.J.Hubert. An overview of Combinatorial Data Analysis (Chapter 1), in: Clustering and Classification, P.Arabie, L.J.Hubert and G.Desoete, eds, World Scientific, 1999.

[11] G.H.Bal and D.I.Hall. Some fundamental concepts and synthesis procedures for pattern recognition preprocessors. In Proc. of Int. Conf. Microwaves, Circuit theory and Information Theory, Tokyo, Japan, pp281-297, Sep., 1964.

[12] J.C.Lin. Multi-class clustering by analytical two-class formulas. Pattern recognition and artificial intelligence.Vol.10, No.4, pp307-323, 1996.

[13] W.C.Chang. On using principal components before separating a mixture of two multivariate normal distribution.Applied Statistics, Vol.32, pp267-275, 1983.

[14] 陈文伟. 智能决策技术[M].北京：电子工业出版社.1998

[15] 沈倩，汤霖. 模式识别导论. 长沙：国防科学技术大学出版社. 1991；30-154.

[16] 范九伦，裴继红，谢维信. 基于可能性分布的聚类有效性[J]. 电子学报. 1998；26(4):113-115

[17] Mika Sato-Eic. On Evaluation of clustering using homogeneity analysis.

IEEE, 3588-3593, 2000

[18] A.K.Jain andR.C.Dubes, Algorithms for Clustering Data. Englewood Cliffs, NJ: Prentice-Hall, 1988.

[19] 张莉，周伟达，焦李成. 核聚类算法[J]. 计算机学报，2002；25(6): 587-590

[20] Colomi A, Dorigo M, Maniezzo V. Distributed optimization by ant colonies[C]. In: Proc of 1st European conf Artificial Life

[21] Bilchev G, Parmee I C. Searching heavily contrained design spaces[C]. In: Proc of 22nd Int. Conf Computer Aided Design 95, Yelta:Ukraine, 1995: 230-235

[22] 杨欣斌，孙京诰，黄道.一种进化聚类学习新方法[J]. 计算机工程与应用, 2003；15:60-62

[23] 吴斌，傅伟鹏，郑毅，刘少辉，史忠植. 一种基于群体智能的Web文档聚类算法[J]. 计算机研究与发展, 2002；39(11):1429-1435

[24] 张宗永，孙静，谭家华. 蚁群算法的改进及其应用[J]. 上海交通大学学报[J]. 2002；36(11): 1564-1567

[25] 杨燕，靳蕃, M.Kamel. 一种基于蚁群算法的聚类组合方法[J]. 铁道学报[J]. 2004；26(4): 64-69

[26] 吴斌. 群体智能的研究及其在知识发现中的应用. 博士论文，中国科学院计算技术研究所，2002

[27] http://www.ics.uci.edu/~mleam/MLRepository.html

[28] http://www.ibm.com.cn/

[29] Bonabeau, Dorigo M, Theraulaz G. Inspiration for optimization from social insect behavior. Nature, 2000, 406(6):39-42

[30] Dorigo M, Bonabeau E, Theralulaz G. Ant algorithms and stigmergy. Future Generation Computer Systems, 2000, 16(8):851-871

[31] Stutzle T, Hoos H.MAX-MIN Ant systems. Future Generation Computer Systems, 2000, 16(8):889-914

[32] Bonabeau E, Dorigo M, Theraulaz G. Swarm Intelligence: From Natural to Artificial Systems. New York:Oxford University Press, 1999

[33] Gianni Di Caro, Marco Dorigo. AntNet: Distributed stigmergetic control for communications networks. Journal of Artificial Intelligence Research,1998；9:317-355

[34] Deneubourg J L, Goss S, Frank N, Sendova-hanks A, Detrain C, Chrerien

L. The dynamics of collective sorting: robot-like ants and ant-like robots. In: Proceedings of the 1st International Conference on Simulation of adaptive behavior: From Animals to Animals, MIT Press/ Bradford Books, Cambridge, MA, 1991；356-363

[35] Holland O E, Melhuish C. Stigmergy, self-organization , and sorting in collective robotics. Artificial Life 1999；5(2):173-202

[36] Lumer E, Faieta B. Diversity and adaptation in populations of clustering ants. In: Proceedings of the 3rd International Conference on Simulation of Adaptive Behavior: From Animals to Animals, 3, MIT Press/ Bradford Books, Cambridge, MA, 1994；501-508

[37] Yang Y, Kamel M. Clustering ensemble using swarm intelligence[A]. In: IEEE swarm intelligence symposium[C]. Piscataway, NJ: IEEE service center, 2003；65-71

[38] 行小帅，焦礼成.数据挖掘的聚类算法[J]. 电路与系统学.2003；8(1): 59-67

[39] 忻斌健，汪镭，吴启迪. 蚁群算法的研究现状和应用及蚂蚁智能体的硬件实现[J]. 同济大学学报. 2002；30(1):82-87

[40] 周勇，陈洪亮. 蚁群算法的研究现状及其展望[J]. 微型电脑应用. 2002；18(2):5-9

[41] M.Dorigo, "Optimization learning and natural algorithms", Ph.D. Thesis, Dip. Electronic Information, PolitecnicodiMilano, Italy,1992.

[42] 李生红,刘泽民,周正. ATM网上基于蚂蚁算法的VC路由选择方法[J]. 通信学报. 2000；21(1): 32-37

[43] 林锦，朱文兴. 凸整数规划问题的混合蚁群算法[J]. 福州大学学报(自然科学版). 1999；27(6): 123-127

[44] 段云峰，吴唯宁等. 数据仓库及其在电信领域中的应用[M]. 电子工业出版社.2003

[45] 万小军，杨建武，陈晓鸥.文档聚类中k-means算法的一种改进算法[J]. 计算机工程. 2003；29(2): 102-104

[46] 严蔚敏，吴伟民. 数据结构. 清华大学同版社. 1995

[47] 蔡自兴，徐光佑. 人工智能及其应用(第二版). 清华大学出版社.1999

[48] 何新贵. 知识处理与专家系统. 国防工业出版社. 1990

[49] 同济大学教研室. 线性代数(第二版). 高等教育出版社. 1996

[50] 赵选民，徐伟，师义民，秦超英. 数理统计. 科学出版社. 2002

[51] 何平. 数理统计与多元统计. 西南交通大学出版社. 2004

[52] Ishioka T. Evaluation of criteria for information retrieval. 2002. http://www.rd.dnc.ac.jp/~tunenori/doc/ishiokat_criteria.ps

[53] 杨冬青. 把握数据挖掘新动向. 中国计算机报. 1998.

[54] 胡雪梅. 浅议数据仓库技术在中国电信的应用前景. 四川绵阳分公司. 2001.

[55] 潘维民. CRM中的数据仓库. 计算机世界网. 2004.

[56] http://www.dwway.com

[57] 邵峰晶, 孙仁诚, 于忠清. 基于单元的孤立点发现改进算法. 青岛大学学报, 2003(1)

[58] 李建中, 高宏. 一种数据仓库的多维模型. 软件学报,2000; 11(7):908-917.

[59] 刘明吉, 王秀峰, 黄亚楼. 数据挖掘中的数据预处理. 计算机科学, 2000; 27(4):54-57

[60] 邵峰晶, 于忠清. 数据挖掘原理与算法. 中国水利水电出版社.2003

[61] 王珊. 数据仓库技术与联机分析处理. 科学出版社. 1999

[62] 王颖, 谢剑英. 一种基于蚁群算法的多媒体网络多播路由算法. 上海交通大学学报. 2002; 36(4):526-528

[63] 熊伟清, 余舜洁. 具有分工的蚁群算法及应用. 模式识别与人工智能. 2003, 16(3):328-333

[64] 杨勇, 宋晓峰. 蚁群算法求解连续空间优化问题. 控制与决策. 2003; 18(5):573-576

[65] 张素兵, 刘泽民. ATM业务控制中的一种新的神经网络方法. 北京邮电大学学报. 2001; 24(2):15-19

[66] 张宗永, 孙静, 谭家华. 蚁群算法的改进及其应用. 上海交通大学学报. 2002; 36(11): 1564-1567

[67] 周伟, 刘粉林. 简单蚁群算法的仿真分析. 控制与决策. 2003; 18(3):317-319

[68] 王笑蓉, 吴铁军. Flowshop问题的蚁群优化调度方法. 系统工程理论与实践. 2003; 23(5):65-71

[69] 汪镭, 吴启迪. 蚁群算法在系统辨识中的应用. 自动化学报. 2003; 29(1):102-109

[70] 李勇, 段正澄. 动态蚁群算法求解TSP问题. 计算机工程与应用. 2003; 39(17):103-106

[71] 洪炳熔，金飞虎. 基于蚁群算法的多层前馈神经网络.赠哈尔滨工业大学学报. 2003; 35(7):823-825

[72] Bland J A. 1999. Layout of facilities using an ant system approach. Engineering Optimization, 32(1):101-115

[73] Boryczka M. 1998. Some aspects of ant systems for the TSP. Fundamenta Information, Aug., 35(1-4):197-209

[74] Dorigo M, Maniezzo V, Coloni A. 1994. Introduction to natural algorithms. Rivista-di-Infomatic, 24(3):179-197

[75] Dorigo M, Gambardella L M.1996. A study of some properties of Ant-Q International Conference on Evolutionary Computation, Sept.:656-665

[76] Dorigo M, Maniezzo V. 1996. Ant system:optimization by a colony of cooperating agents. IEEE Transactions on System, Man, and Cybernetics, Part B:Cybernetics, Feg., 26(1):29-41

[77] Bonabeau,E.,Dorigo,M.& Theraulaz,G. Swarm Intelligence: From Natural to Artificial Systems ,Oxford Univ. Press, New York, 1999.

[78] 周水庚，周傲英，曹晶，胡运发. 一种基于密度的快速聚类算法. 计算机研究与发展. 2000; 37(11): 1287-1292

[79] 温文波，杜维. 蚁群算法概述. 石油化工自动化. 2002;1(19): 18-22

攻读硕士学位期间所发表的论文

[1] 张昭涛，武勇，杨燕．银行应用系统集成研究．山西财经大学学报．2004；26(2):144-146

数据挖掘聚类算法研究

作者：张昭涛
学位授予单位：西南交通大学
被引用次数：3次

参考文献(79条)

1. 参考文献

2. 黄振华.吴诚一 模式识别 1991

3. 蔡元龙 模式识别 1986

4. 吴庆洪.张纪会.徐心和 具有变异特征的蚁群算法[期刊论文]-计算机研究与发展 1999(10)

5. 马良.项培军 蚂蚁算法在组合优化中的应用[期刊论文]-管理科学学报 2001(2)

6. 谢维信 工程模糊数学方法 1991

7. J Han.M Kamber Data Mining:Concepts and Techniques 2001

8. Michael J Corey.Michael Abbey.Lan Abramson.Ben Taub.陈越,郭渊博 Oracle 8数据仓库分析、构建实用指南 2000

9. Efrem G Mallach.李昭智.李昭勇 决策支持与数据仓库系统 2001

10. P Arabie.L J Hubert An overview of Combinatorial Data Analysis 1999

11. G H Bal.D I Hall Some fundamental concepts and synthesis procedures for pattern recognition preprocessors 1964

12. J C Lin Multi-class clustering by analytical two-class formulas 1996(04)

13. W C Chang On using principal components before separating a mixture of two multivariate normal distribution 1983

14. 陈文伟 智能决策技术 1998

15. 沈倩.汤霖 模式识别导论 1991

16. 范九伦.裴继红.谢维信 基于可能性分布的聚类有效性[期刊论文]-电子学报 1998(4)

17. Mika Sato-Eic On Evaluation of clustering using homogeneity analysis 2000

18. A K Jain.R C Dubes Algorithms for Clustering Data 1988

19. 张莉.周伟达.焦李成 核聚类算法[期刊论文]-计算机学报 2002(6)

20. Colomi A.Dorigo M.Maniezzo V Distributed optimization by ant colonies

21. Bilchev G.Parmee I C Searching heavily contrained design spaces 1995

22. 杨欣斌.孙京诰.黄道 一种进化聚类学习新方法[期刊论文]-计算机工程与应用 2003(15)

23. 吴斌.傅伟鹏.郑毅.刘少辉.史忠植 一种基于群体智能的Web文档聚类算法[期刊论文]-计算机研究与发展 2002(11)

24. 张宗永.孙静.谭家华 蚁群算法的改进及其应用[期刊论文]-上海交通大学学报 2002(11)

25. 杨燕.靳蕃.Mohamed Kamel 一种基于蚁群算法的聚类组合方法[期刊论文]-铁道学报 2004(4)

26. 吴斌 群体智能的研究及其在知识发现中的应用[学位论文]博士 2002

27. 查看详情

28. 查看详情

29. Bonabeau.Dorigo M.Theraulaz G Inspiration for optimization from social insect behavior 2000(06)

30. Dorigo M.Bonabeau E.Theralulaz G Ant algorithms and stigmergy[外文期刊] 2000(08)

31. Stutzle T.Hoos H MAX-MIN Ant systems 2000(08)

32. Bonabeau E.Dorigo M.Theraulaz G Swarm Intelligence:From Natural to Artificial Systems 1999

33. Gianni Di Caro.Marco Dorigo AntNet:Distributed stigmergetic control for communications networks 1998

34. Deneubourg J L.Goss S.Frank N.Sendova-hanks A,Detrain C,Chrerien L The dynamics of collective sorting:robot-like ants and ant-like robots 1991

35. Holland O E.Melhuish C Stigmergy,self-organization,and sorting in collective robotics 1999(02)

36. Lumer E.Faieta B Diversity and adaptation in populations of clustering ants 1994

37. Yang Y.Kamel M Clustering ensemble using swarm intelligence 2003

38. 行小帅.焦李成 数据挖掘的聚类方法[期刊论文]-电路与系统学报 2003(1)

39. 忻斌健.汪镭.吴启迪 蚁群算法的研究现状和应用及蚂蚁智能体的硬件实现[期刊论文]-同济大学学报(自然科学版) 2002(1)

40. 周勇.陈洪亮 蚁群算法的研究现状及其展望[期刊论文]-微型电脑应用 2002(2)

41. M Dorigo Optimization learning and natural algorithms 1992

42. 李生红.刘泽民.周正 ATM网上基于蚂蚁算法的VC路由选择方法[期刊论文]-通信学报 2000(1)

43. 林锦.朱文兴 凸整数规划问题的混合蚁群算法[期刊论文]-福州大学学报（自然科学版） 1999(6)

44. 段云峰.吴唯宁 数据仓库及其在电信领域中的应用 2003

45. 万小军.杨建武.陈晓鸥 文档聚类中k-means算法的一种改进算法[期刊论文]-计算机工程 2003(2)

46. 严蔚敏.吴伟民 数据结构 1995

47. 蔡自兴.徐光佑 人工智能及其应用 1999

48. 何新贵 知识处理与专家系统 1990

49. 同济大学教研室 线性代数 1996

50. 赵选民.徐伟.师义民.秦超英 数理统计 2002

51. 何平 数理统计与多元统计 2004

52. Ishioka T Evaluation of criteria for information retrieval 2002

53. 杨冬青 把握数据挖掘新动向 1998

54. 胡雪梅 浅议数据仓库技术在中国电信的应用前景 2001

55. 潘维民 CRM中的数据仓库 2004

56. 查看详情

57. 邵峰晶.孙仁诚.于忠清 基于单元的孤立点发现改进算法 2003(01)

58. 李建中.高宏 一种数据仓库的多维数据模型[期刊论文]-软件学报 2000(7)

59. 刘明吉.王秀峰.黄亚楼 数据挖掘中的数据预处理[期刊论文]-计算机科学 2000(4)

60. 邵峰晶.于忠清 数据挖掘原理与算法 2003

61. 王珊 数据仓库技术与联机分析处理 1999

62. 王颖.谢剑英 一种基于蚁群算法的多媒体网络多播路由算法[期刊论文]-上海交通大学学报 2002(4)

63. 熊伟清.余舜浩.赵杰煜 具有分工的蚁群算法及应用[期刊论文]-模式识别与人工智能 2003(3)

64. 杨勇.宋晓峰.王建飞.胡上序 蚁群算法求解连续空间优化问题[期刊论文]-控制与决策 2003(5)

65. 张素兵.刘泽民 ATM业务控制中的一种新的神经网络方法[期刊论文]-北京邮电大学学报 2001(2)

66. 张宗永.孙静.谭家华 蚁群算法的改进及其应用[期刊论文]-上海交通大学学报 2002(11)

67. 周伟.刘粉林.吴灏.王清贤 简单蚁群算法的仿真分析[期刊论文]-控制与决策 2003(3)

68. 王笑蓉.吴铁军 Flowshop问题的蚁群优化调度方法[期刊论文]-系统工程理论与实践 2003(5)

69. 汪镭.吴启迪 蚁群算法在系统辨识中的应用[期刊论文]-自动化学报 2003(1)

70. 李勇.段正澄 动态蚁群算法求解TSP问题[期刊论文]-计算机工程与应用 2003(17)

71. 洪炳熔.金飞虎.高庆吉 基于蚁群算法的多层前馈神经网络[期刊论文]-哈尔滨工业大学学报 2003(7)

72. Bland J A Layout of facilities using an ant system approach[外文期刊] 1999(01)

73. Boryczka M Some aspects of ant systems for the TSP 1998(1-4)

74. Dorigo M.Maniezzo V.Coloni A Introduction to natural algorithms 1994(03)

75. Dorigo M.Gambardella L M A study of some properties of Ant-Q 1996

76. Dorigo M.Maniezzo V Ant system:optimization by a colony of cooperating agents 1996(01)

77. Bonabeau E.Dorigo M.Theraulaz G Swarm Intelligence:From Natural to Artificial Systems 1999

78. 周水庚.周傲英.曹晶.胡运发 一种基于密度的快速聚类算法[期刊论文]-计算机研究与发展 2000(11)

79. 温文波.杜维 蚁群算法概述[期刊论文]-石油化工自动化 2002(1)

本文读者也读过(10条)

8. 白立群 论科学发展观的核心：以人为本[学位论文]2005
9. 万志波 协同设计系统中的版本管理技术研究[学位论文]2005
10. 王永康 供应链信息共享及其技术实现模型研究[学位论文]2005

引证文献(3条)

1. 蒋志为 基于模糊集的蚁群聚类算法研究[学位论文]硕士 2006
2. 翁丽芳 基于银行机构客户账户的可疑洗钱交易行为识别研究[学位论文]硕士 2006
3. 谭华琴 基于蚁群算法的数据挖掘方法研究[学位论文]硕士 2006