分类号：_____ 密　　级：_____

U D C：_____ 学　　号：_____

# 江西理工大学

# 硕 士 学 位 论 文

## 相关性加权 K-means 算法的改进及其应用

## Improvement and Application of Correlation Weighted K-means Algorithm

学 位 类 别：理学硕士

作 者 姓 名：吴斌

学 科、专 业：计算机科学与技术

研 究 方 向：智能计算

指 导 教 师：刘建生(副教授)

年　月　日

# 学位论文独创性声明

本人声明所呈交的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。据我所知，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，也不包含已获得江西理工大学或其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中做了明确的说明并表示谢意。

申请学位论文与资料若有不实之处，本人承担一切相关责任。

研究生签名：　　时间：　　年　　月　　日

# 学位论文版权使用授权书

本人完全了解江西理工大学关于收集、保存、使用学位论文的规定：即学校有权保存按要求提交的学位论文印刷本和电子版本，学校有权将学位论文的全部或者部分内容编入有关数据库进行检索，并采用影印、缩印或扫描等复制手段保存、汇编以供查阅和借阅；学校有权按有关规定向国家有关部门或者机构送交论文的复印件和电子版。本人允许本学位论文被查阅和借阅，同意学校向国家有关部门或机构送交论文的复印件和电子版，并通过网络向社会公众提供信息服务。

保密的学位论文在解密后适用本授权书

学位论文作者签名（手写）：　　导师签名（手写）：

签字日期：　　年　　月　　日　　签字日期：　　年　　月　　日

# 摘　要

伴随着科学技术的迅速发展，当今社会已演变成为一个信息爆炸的时代，每天大量的数据信息被产生与更新。因此，通过对每天产生的数据进行挖掘，并从中提取出有用信息变得尤为重要。然而数据的规模早已超越了传统方法分析与处理数据的能力，出现了“数据爆炸，却知识贫乏”的现象。快速、精确的提取出海量数据中隐藏的信息成为众多研究人员所研究的热点。而对无标签问题的挖掘与提取，聚类分析扮演着不可或缺的角色。并在众多领域得到有效应用。如病毒入侵检测、统计分析、图像处理等。

K-means 算法作为数据挖掘算法中十大经典算法之一，是采用交替最小化方法求解非凸优化问题的迭代型算法。该算法结果简单易懂、运行效率高，它作为一种无监督的聚类算法，在历史上，有着许多不同领域的研究人员对其进行研究与改进，其中比较知名的有 Forgey，McQueen 等人。该算法已被广发应用在许多不同的领域。但是仍旧有着许多的问题没有得到很好的解决。如初始中心点的选取、确定数据集的类别数、样本对象间相似性等问题。因此，为提高该算法在聚类过程中的稳定性以及对象间的相关性等问题。本文分别以算法中的初始化、聚类中心数目的确定以及距离划分函数三点作为主要研究目标，并提出了相应的改进方法。同时将改进的算法对股价进行分析。具体工作如下：

（1）针对K-means算法随机选取初始中心点过程中所选聚类中心敏感。本文通过将最大最小初始化与密度相结合，提出一种新的初始化方法。文中通过所选初始点密度与类间距离来确定阈值，然后对样本对象进行划分。从而获得稳定、唯一、逼近真实分布的初始聚类中心。实验结果表明，本文能获得更优的初始簇中心。

（2）预先确定类别数作为K-means算法研究中的一大难点。针对此问题，本文在Rt-kmeans算法的基础上进行了改进，从而获得一种自动确定数据集类别数的K-means算法，相比Rt-kmeans算法，本文方法所得到的数据集的类别数目更为准确。

（3）传统K-means算法难以体现对象间相关性的问题。本文使用皮尔相关性系数对欧式距离进行加权，从而增强簇中样本对象间的相关性。相比传统K-means算法以及一些改进的较为新颖的K-means算法，实验数据表明，本文改进的K-means算法所得结果更准确。

（4）股票数据的分析过程中主要分为两部分：一是对多支股票间的组合分析；二是关于单支股票自身股价波动变化可能性的研究。结果表明，C-kmeans算法能将关联性较强的股票划分在相同的类中。

**关键词：**聚类分析；K-means算法；自动划分；皮尔逊相关系数

# Abstract

With the rapid development of science and technology, today's society has evolved into an era of information explosion. A large amount of data information is generated and updated every day. Therefore it is particularly important to extract useful information and knowledge from these vast amounts of data. However, the scale of the data far exceeds the ability of traditional methods to analyze and process the data, resulting in the phenomenon of "data explosion but lack of knowledge". The fast and accurate extraction of the hidden information in massive data has always been a hot research topic of many researchers. Clustering analysis plays an indispensable role in the mining and extraction of unlabeled problems. Cluster analysis has been effectively applied in many fields. Such as virus intrusion detection, statistical analysis, image processing and more.

K-means algorithm as one of the ten classic data mining algorithms, it is an iterative algorithm that uses alternating minimization to solve non-convex optimization problems. The algorithm is simple and easy to understand and has high operation efficiency. As an unsupervised clustering algorithm, it has been researched and improved by many researchers in history,such as Forgey, McQueen and so on. The algorithm has been widely used in many fields. But there are still many problems are not solved. For example, the selection of the initial center point, determining the number of data set categories, the similarity between sample objects and other issues.Therefore, in order to improve the stability of the algorithm in the clustering process and the correlation between objects and other issues. In this paper,the initialization of the algorithm, the determination of the number of cluster centers and the distance function are the main research objects, and the corresponding improvement methods are proposed. At the same time, the improved algorithm will analyze the stock price. The specific work is as follows:

Selecting cluster center sensitivity during the process of randomly selecting the initial center point for the K-means algorithm. In this paper, we combine max-min initialization method with density and propose a new initialization method. In this paper, the initial point density and inter-class distance are used to determine the threshold, and then the sample object is divided. And obtain stable, unique, approximate initial cluster center that is truly distributed. Experiments show that the initial cluster center is better.

Pre-determining the number of clusters as a major difficulty in K-means algorithm. To solve this problem, this paper improves on the Rt-kmeans algorithm to obtain a K-means algorithm that

automatically determines the number of clusters. Compared with Rt-kmeans algorithm, the proposed method can more accurately determine the number of data set .

Traditional K-means algorithm is difficult to reflect the inter-object correlation problem. In this paper, the Pearson correlation coefficient was used to weight the Euclidean distance, thus enhancing the correlation between sample objects in the cluster. Compared with the traditional K-means algorithm and some novel K-means algorithms,the experimental results show that the paper improves K-means algorithm accurate results is better.

The analysis of stock data is mainly divided into two parts: one is the combination analysis of multiple stocks, and the second is the research on the possibility of volatility of stock price of a single stock. The results show that the algorithm can classify the stocks with strong correlation in the same class.

**Key Words**： cluster analysis; K-means algorithm; automatic division; Pearson correlation coefficient

# 目 录

# 第一章 绪 论

## 1.1 研究背景及意义

现今是一个信息爆炸、数据量迅速膨胀、各行各业经济蓬勃发展的时代。无论是从数据的范围还是数据的总量而言，传统的数据挖掘方法已不能很好解决人们对现今数据信息的需求。数据挖掘技术已从最初的搜索阶段逐渐演变到了现今的决策支持阶段。表 1.1 中展现了在各个时期，数据挖掘技术的变化情况及其代表产品[1,2]。

表 1.1 数据挖掘的发展历程

| 进化阶段 | 支持技术 | 产品厂家 | 产品特点 |
|---|---|---|---|
| 数据搜索（20 世纪 60 年代） | 计算机、磁带和磁盘 | IBM,CDC | 提供历史性的、静态的数据信息 |
| 数据访问（20 世纪 80 年代） | 结构化查询语言、数据库 | Oracle、IBM、Microsoft、Sybase | 提供历史性的、动态的数据信息 |
| 数据仓库（20 世纪 90 年代） | 联机分析处理、数据仓库 | Pilot、Arbor、Comshare、Cognos | 在各种层次提供动态的数据信息 |
| 数据挖掘（现在流行） | 高级算法、海量数据库 | Pilot、SCI、IBM | 提供预测性的信息 |

数据挖掘的任务及其方法存在着多种形式，为该领域的研究提出了较多具有建设性的课题，在未来将成为一个巨大的研究热潮。具体可能包含以下几点：研究如何使得数据挖掘语言变得正规化及其规范化，探索在分布式条件下高效的数据挖掘方法，探寻用户易于理解与接受的简易挖掘模型，增强对各种数据类型挖掘的适用性与广泛性，如文本型、图形图像型、多媒体型等各种数据类型，研究出满足各种数据类型、噪声容限的挖掘方法；使该技术的应用范围能够得到较大幅度的增长，如应用在智能电网数据的采集[3]、金融行业的数据分析、提取犯罪信息、药物研发等领域之中[4,5]。

现今，数据挖掘技术已逐渐成熟，其中异常值分析，关联规则分析，聚类分析等在该技术中扮演着重要角色[6,7]。聚类作为数据挖掘非常重要的研究分支，在该领域的研究中具有极高的价值与意义。聚类最早由分类学中所区分出来，并逐渐演变为一门单独的学科。但聚类并不等同于分类，它是对未进行标记的事物进行划分，使相似性较强的样本对象能够聚集在同一类中，而对于差异性较大的样本对象能够被划分在不同的类别之中，是一种无监督学习的划分方法。

“物以类聚，人以群分”，生活中存在着大量类的划分问题[8]，聚类分析作为数据挖掘与模式识别等多门学科中的一大研究热点，在识别无符号标记的数据上有着重要的作

用。研究人员针对不同的应用背景研究出了不同的聚类算法。众多被提出的算法之中，常被划分为以下五类，分别是基于模型的聚类算法、基于密度的聚类算法、基于层次的聚类算法法、基于划分的聚类算法以及基于网格的聚类算法[9]。时至今日，这几类算法经过长时间的研究与发展，都有着各自典型的代表性算法。如以模型为主的聚类算法中主要代表性算法有统计学方法、神经网络方法等；而以密度为原理的代表性算法主要有 DBSCAN 算法和 DENCLUE 算法；采用层次分析的算法中其代表性算法主要有 BIRCH 算法、CURE 算法、OPTICS 算法；基于划分的方法中主要有 K-means 算法、K-中心算法等；在基于网格的方法主要有 STING 算法和 CLIOUE 算法[10-19]。在这些众多的算法中，具体应该选择哪一种算法要结合不同的应用背景进行选取，有些聚类算法适合应用于时间序列数据集，而有的算法则可能适用连续属性的情况，有的算法需要经过严谨的数学推导，而有的则思想直观，简单明了。但无论使用何种聚类算法，都是为了能够高效的对数据信息进行聚类与预测，使得聚类结果能够对决策者做决策时起到帮助。

然而，这些不同的聚类方法都或多或少的存在着各自的不足，没有一种聚类方法可以适用于所有的应用与分析。因数据挖掘对于聚类分析的重要需求性，因此吸引了大量研究人员的目光，并因此涌现出了许多的研究成果，但仍有许多的待解决的问题。

K-means算法作为十大经典数据挖掘算法之一，具有简单、高效等众多优良特性。同样，该算法中依旧存在众多的不足，因此，对于K-means算法的研究与改进是一件非常有意义的事情。

## 1.2 国内外研究现状

20 世纪 80 年代后期，许多学者逐渐开始对数据挖掘的理论进行研究与分析；并且 90 年代期间得到了迅速的发展。同时，数据挖掘又称为数据库中的知识发现，1989 年研究学者第一次正式在第十一届国际联合人工智能交流会议上提出，紧接着在 1995 年又举办了 KDD & Datamining国际学术会议，该会议后来每年都会被举办一次。除KDD & Datamining国际会议外，国际上还有一些关于数据挖掘领域中的顶尖会议。如PODS、DaWak、CCDM等会议。

数据挖掘的本质是对大批量的、存在异常值或是模糊的数据集进行有效分析，然后从中挖掘出隐含在数据集中并且有效的数据信息。并且它作为一门交叉学科，集成了许多学科中成熟的工具和技术，包括仓库技术、模式识别、概率统计、智能计算、神经网络等等[20-22]。数据挖掘技术是现今社会分析大数据时的有效手段，研发人员已开发出了许多的数据挖掘产品，并在很多公司得到了有效的使用。

相比其他国家，国内对于数据挖掘的研究起步较晚，在该领域中所获得的成果也相对较少。在国内，该领域的研究分为两部分，一部分是关于理论知识的研究，另一部分是对挖掘算法进行研究，对于实际的应用开发研究在国内还相对较少。并且无论是对理论研究还是对算法进行研究，研究经费主要来源于国家项目，如自然科学项目基金等，仅有极少部分为一些公司的自主性研究。

虽然与其他国家相比，国内关于数据挖掘领域的研究相对比较落后，但是也已粗略具备了整体研究的实力。在 1993 年国家自然科学基金第一次对数据挖掘领域的研究进行了资助。随后包括中国科学院计算技术研究所、清华大学、中山大学等在内的科研院所和各著名大学相继开展了对数据挖掘领域基本理论及其应用的研究，并且涌现出来许多的研究成果。

## 1.3 研究内容

传统K-means算法作为十大经典数据挖掘算法之一，因其具有较多优点而被研究人员所青睐，但同时该算法中依旧存着在许多的不足。本文主要通过对传统K-means算法中所存在的各不足进行了相应的研究，并提出了相应的改进方法。本文主要研究了K-means算法初始中心点敏感、聚类中心数目难以预先确定以及样本对象间的相关性不能得到有效反映等问题。本论文的组织结构如下。

第二章中主要对数据挖掘技术以及聚类分析研究中所涉及到的相关知识进行了介绍。文中先对数据挖掘技术进行了简短介绍，其中主要包含数据挖掘的概念、流程以及样本数据的类型等；其次对于数据挖掘中的聚类分析方法做了简要回顾，主要包含了聚类分析的原理、过程等，同时并列举了一些聚类分析的代表性方法。最后对传统K-means算法进行简短的介绍，其中主要包括了传统K-means算法的发展、原理、步骤流程以及优缺点等内容。除此之外，本章节中也将传统K-means算法的一些经典的演变算法及最新改进的算法进行了简短介绍。

第三章是本文对传统K-means算法的改进思想及其改进步骤，同时并对本文改进方法的有效性及其合理性进行了验证。本章中对传统K-means算法的研究以及对于该算法的改进主要分为三个部分进行介绍。第一部分是对初始聚类中心的选取；第二部分是对聚类中心数目k值的自适应划分；最后是对欧式距离采用皮尔相关性系数进行加权，从而增强聚类过程中样本对象间的相关性。

第四章中主要是对本文改进算法在股票应用中的介绍。本章主要分为两部分，第一分部是对股票数据的特性进行简短介绍；第二部分是将本文所改进的算法对股票的价格变化

进行分析，主要分析了处在相同类或是不同类中不同股票间的检验值及其相关性系数之间的关系。

第五章是对全文进行总结，同时对未来的研究进行展望。

# 第二章 数据挖掘及聚类分析

## 2.1 数据挖掘

### 2.1.1 数据挖掘概念

数据挖掘技术是采用科技技术手段对大批量的数据进行整合，然后统计分析出数据中所隐含的有效信息的过程。因此数据挖掘技术呈现以下特点：一是数据挖掘技术主要是通过采用其他专业学科知识，如统计学、神经网络等来建立模型，同时根据应用背景来设计出相适用的算法，并从中寻找出隐藏在内部的潜在信息，从而挖掘出事物之间的内在关联性；二该技术主要是使用算法来对所隐藏的信息进行分析与提取，而样本信息的形式具有多样性，因此在数据挖掘过程中需对所使用的样本信息进行预处理；三是通过建立模型的方式来处理实践项目[23]。数据挖掘的最终目的不仅是挖掘出数据的规律，而是希望所挖掘出的信息能够满足用户的需求。

### 2.1.2 数据挖掘流程

数据挖掘是从大批量的信息中挖掘出用户需求的信息知识，主要包括经验知识与理论知识，然后对所获得的知识进行归纳并提取出用户所需求的信息。数据挖掘过程主要经历了三步，首先是数据准备，其次是数据开采，最后是对结果的表达与解释[24]。流程图如图2.1 所示。

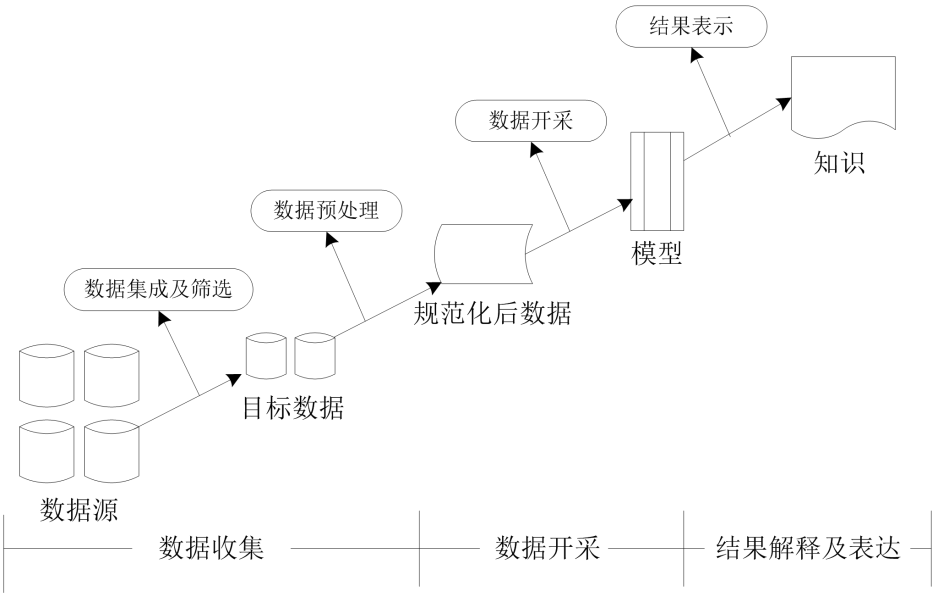



图 2.1　数据挖掘流程图

数据的准备过程将直接影响到最终所获得信息的有效性以及准确性，在数据准备过程中主要包括数据的集成、选择及其预处理三个子步骤。

数据集成过程就是当使用者确定了数据挖掘目标以后，通过把各个数据源进行整理与合并处理的过程。在提取数据源时，应从所有的数据中进行筛选，并且在筛选数据的过程中应遵行相关性、可靠性、有效性几个标准。使用筛选后的数据集进行分析，将能使得数据间的规律性变得更强，所获得的结果更为准确。除此之外，精选的数据样本在分析过程中所需的工作时间将更少，并且能有效的节约更多的系统存储内存空间。

数据选择就是为了能够识别出所需分析的样本数据，对数据进行收缩处理，保证样本数据的有效性，从而为保证模型的有效性打下坚实的基础，确保所挖掘结果的正确性。

数据预处理是为了保证数据的形式在挖掘过程中满足使用者的要求。例如当数据维度过大，如何对高维数据集进行降低维度处理，又或是存在缺失数据时，如何对缺失值进行填补等都是数据预处理所需要解决的基本问题。并且，在数据的采集过程中，所得到的许多数据经常包含噪声或是所得数据不完整等，因此对数据的预处理则变得必不可少。常用的数据预处理方法主要有数据筛选、缺失值填补、数据维度处理、属性选择等。

数据开采是根据所研究数据样本的背景选择对应的模型与算法，如线性回归算法、聚类算法、分类算法等，并将其用于挖掘数据中所存在的潜在规律。该步骤是挖掘数据潜在规律最为关键的一步，同时也是主要技术难点之一。

结果的表达和解释是使用者对所挖掘出的结果进行统计与分析，并将有用的信息提取出来，从中探寻出数据中所存在的潜在规律。如果最终所获得的结果并不理想，那么将要重新重复以上挖掘过程，获得新的结果。

### 2.1.3 数据挖掘主要任务

数据挖掘技术是通过对已有的数据信息进行分析，进而发现数据中的内部规律，以此对未来数据的发展及其变化趋势进行预测。从大量无规则的数据中发现出数据集中发现隐藏的、潜在的、有价值的信息是数据挖掘的主要目标，常见的数据挖掘任务主要包含七种，分别是分类分析、聚类分析、回归分析、关联规则、特征分析、偏差分析以及Web挖掘技术，不同的挖掘任务所应用的场景可能不同。例如聚类分析是一种无标签的学习方法，可以用来对新生事物进行研究。图 2.2 展现了各挖掘任务及所应用到的部分场景。

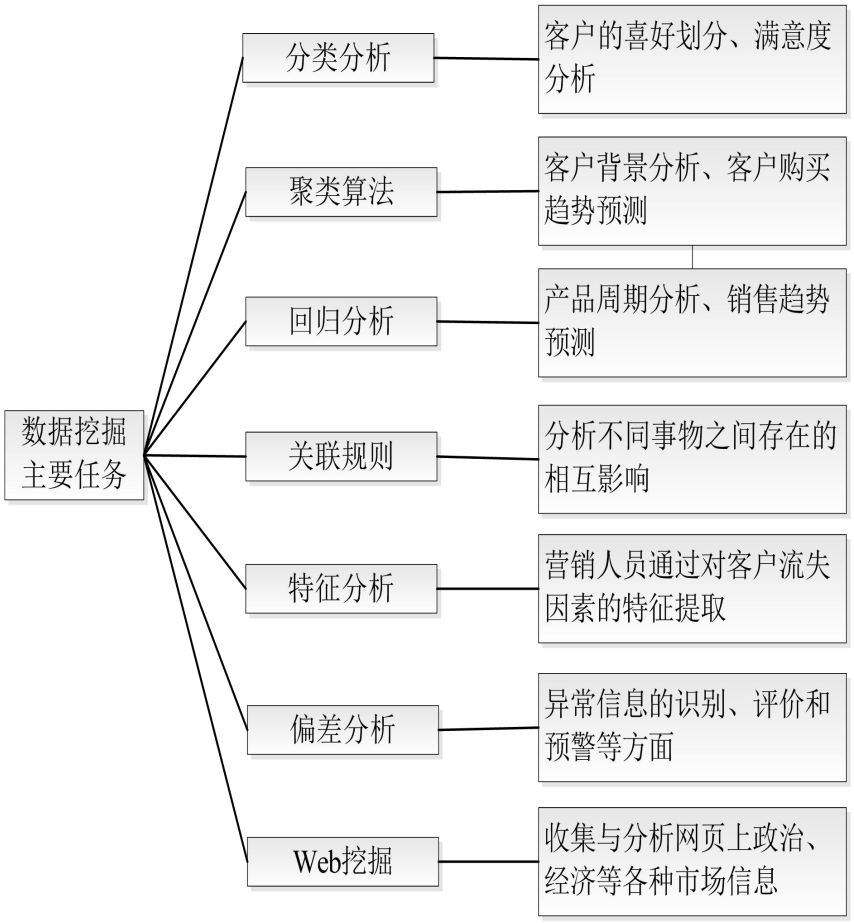



图 2.2　数据挖掘主要任务及应用

数据挖掘是通过借助科学工具分析市场上已有的数据信息，从中挖掘出潜在的、具有较强规律性的有效信息，帮助用户能够做出正确的决策。

21世纪，作为信息技术高速发展的时代。越来越多的人迫切的希望从海量的数据信息中提取出有用的信息，因此数据挖掘技术有着广阔的应用前景，并且已经有越来越多的学者将投入该领域的研究。伴随着数据挖掘技术研究的不断进步，必将为用户带来更多的利益。

## 2.2 聚类分析

### 2.2.1 什么是聚类分析

聚类分析是人们研究事物的本质规律，发现事物之间相互依存关系的一种技术手段。“物以类聚，人以群分”就是这种现象在社会生活中的直观体现。聚类的原理就是将一大

整体划分成若干个小团体，使得每一个团体之间有着较高的相似性，而对于不同的团体之间，差异度较大。

聚类分析最早是从分类分析中所区分出来的，随着人们研究的深入，逐渐演变为了一门独立的学科，但它与分类是不相同的，聚类是识别无标记事物的主要技术手段。以前人们在对无标记事物的辨别中主要依靠历史经验和专业知识来对事物实现分类，但是伴随着科学技术的快速发展，人类知识的不断前行，人们对于事物的分类变得更细，所需的要求也逐渐变得更高。此时，仅凭借历史经验来分类已达不到了人们的要求。并且，对于新出现的事物而言，没有任何的借鉴经验，此时只能凭借人们的主观感受来对事物进行划分，因此划分错误的概率将非常的大。聚类问题逐渐的从分类中区分出来成为了一门相对独立的科学分支。例如在机器学习中，聚类与分类相比，分类若不能提前知道训练样本的属性，则在分类过程中所得结果将不准确。而聚类分析一种无监督的学习方法，则可通过聚类的结果来获取训练样本的不同属性。

### 2.2.2 聚类的原理及算法

聚类方法主要包括两方面的内容，分别是聚类的模型与聚类的算法。其中聚类算法如果按照原理进行划分主要可以划分为五类，并且不同形式划分的聚类算法有着各自典型的代表性算法[25,26]，如图 2.3 所示。研究聚类方法的一般原理，可以有效的指导聚类方法在工作中的实际应用，聚类模型原理如图 2.4 所示。其中各聚类方法各自的优劣及其适用范围如下所示。

基于层次的聚类方法主要有两种划分。一种以合并为聚类的主要原理；另一种是通过逐渐分裂来进行聚类。基于层次的聚类算法中，不需要提前知道所聚类对象的聚类中心的个数，并且对于问题的初始化也不用被考虑。这属于一种静态的聚类方法，所以导致该方法只考虑了由一个质心来作为一个类的代表，使得在样本对象较大的类中，某些子簇就可能远离其父簇的中心。

基于模型的聚类方法是通过给每个聚类构建出一个数学模型，然后使用该模型去对数据信息进行挖掘，发现出符合该模型的样本数据，使得样本数据与模型能够达到高度的拟合状态。期望最大化(ML)算法是基于模型的一种典型代表算法。它是一种广泛应用于极大似然(ML)估计的迭代型计算方法。但是该算法本身不能自动的生成参数估计值，同时算法本身的收敛速度是比较缓慢的，模型的设计也是较为复杂。

基于密度的方法是利用密度作为分割标准，密度相同或相近则归为一类，密度差别较大的则划分为不同的类。但是该方法对于密度参数的选择较为敏感，若不能选取适当的密度参数，所得的聚类结果将不准确。代表算法有 DBICAN、DENCLUE 算法。

基于划分的方法是一种动态的聚类方法，K-means 是其中最具有典型代表性的一个聚类分析算法，该算法在各领域之中都得到了广泛的应用。K-means 算法是一种简单、高效并且有着较高准确度的聚类方法。然而，K-means 算法在聚类的过程中其自身也有着较大的局限性。首先 K-means 算法本身对于初始聚类中心点的选取比较敏感，其次 K-means 算法主要采用欧式距离来度量各个簇间的相似性，然而对于数据间的相关性并没有被很好的展现出来。

基于网格的方法是通过把样本对象进行空间量化转换为有限个单元，形成网络结构。该方法处理速度快，聚类的全部过程都是由网络结构来完成，所需的处理时间与样本对象的数据量无关，只与网络结构中每一维的单元数量有关。

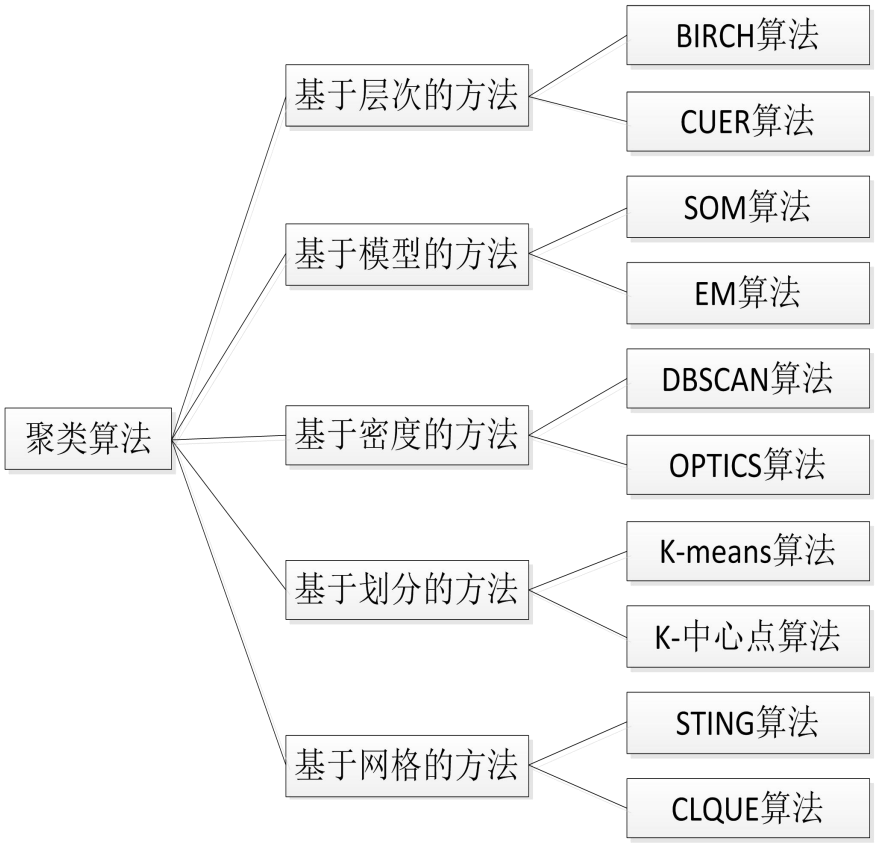



图 2.3　常用的聚类算法

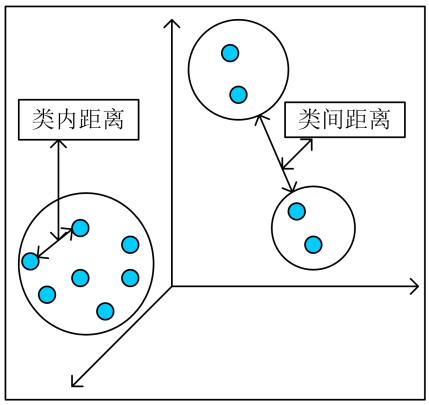



图 2.4　聚类模型原理

### 2.2.3 聚类分析的过程

聚类分析是以数据对象为基础来建立模型的过程，数据集是该过程中的主要处理对象。主要目标是提取数据集中各样本对象所属的类别。该过程中主要包含了数据准备、特征生成、聚类分析三个阶段。

数据准备主要包括了对于数据的获取、属性的选择、数据的清洗等，其中数据的获取就是采用所需要进行挖掘的原始数据，而对于属性的选择则是剔除数据属性中对挖掘信息无影响或是影响较小的属性。数据清洗则是剔除数据中的存在错误数据源或是填补数据源中的存在缺失数据。数据的准备其实就是对数据集进行规范化，数据的规范化对于挖掘数据信息而言非常重要，是数据挖掘中的最为基础的工作。

数据挖掘过程中不同评价指标往往具有不同的量纲，导致数值间具有较大的差别，不预先对样本数据进行规范化处理可能使得最终所分析的聚类结果不准确。为消除不同数据源间取值范围的差异，需对数据源进行规范化处理，即将不同数据源按照比例进行缩放，使之落入一个特定的区域，便于进行综合分析。如将工资收入属性映射到[1,1]或者[0,1]内。数据的标准化处理在基于距离的挖掘算法中体现的尤为重要。具体数据标准化方法主要有最大-最小规范化、零-均值规范化、小数点标规范化等。以下是几种最常用的数据标准化方法[27]。

最大-最小规范化：

$$x' = \frac{x - \min}{\max - \min} \tag{2.1}$$

最大-最小规范化也叫离差标准化，是通过对原始样本数据进行线性变换，使样本数据的值能够映射到[0,1]的区间内，最大-最小规范化保留了原始数据间的相互关系，是消除数据量纲的有效方法之一。

零-均值规范化：

$$x' = \frac{x - \bar{x}}{\sigma} \tag{2.2}$$

零-均值规范化也叫标准差化，采用该方法处理后的样本数据标准差为 1，均值为 0。

除了这几种方法以外，还有着总和标准化，小数点标准化等等。

### 2.2.4 聚类性能的评估方法

评价和比较聚类方法的性能在一定程度上能有效帮助选取合适的聚类算法。聚类只根据数据集自身的属性来进行分类。因聚类在划分的过程中没有一个绝对的划分指标，因此常以距离来对其进行相似性的划分。划分的标准是相同类中的样本对象相互之间具有着较高的相似性，而不同类中的样本对象相互之间具有的较大的差异性。如果相同类间的相似性越大，且不同类间的相似性越小，则认为聚类效果越好。其中主要有 Purity 评价法，RI 评价法，F 值评价法等等。本文研究中，将采用 Purity 评价法来对本课题的聚类效果进行综合性的评价。

Purity 评价法：

Purity 评价法[25]是一种评价聚类效果简单且有效的方法，评价过程只用计算正确聚类数目在总聚类数目中的比重，公式(2.3)所示。

$$purity = \frac{X}{Y} \tag{2.3}$$

其中，X 代表正确聚类对象数目，Y 代表样本对象的总数目。

## 2.3 K-means 算法

### 2.3.1 K-means 算法流程

K-means 算法作为十大经典数据挖掘算法之一，是一种采用交替最小化方法求解非凸优化问题的算法[28]。K-means 算法是一种基于划分的聚类算法，同时也是最具有代表性的一个算法。K-means 算法因其简单、高效并且精确的聚类特性，在许多领域之中已得到了广泛的应用。K-means 算法是一种最常用的传统聚类算法，该算法将一个给定的样本数据集分为用户指定的 k 个类，实现和运行该算法都非常简单，且该算法的时间复杂度仅为(nkt)，其中 n 表示样本数据集的个数，k 表示聚类中心数目，t 表示该算法需要迭代的次数。该算法是划分聚类算法中最流行的算法之一，同时该算法易于修改，因其高效性的原因，这种方法常被广泛的应用于处理大数据分析。

该算法作为一种无监督的迭代型聚类算法，因其算法中所展现的良好特性以及不足之处，从而吸引许多研究人员不断对其进行研究与改进，在历史上，有着许多不同领域的研究人员都对基础的K-means算法进行了研究，其中比较知名的有Forgey（1965），McQueen（1967）等人。Jain与Dubes 详细的描述了K-means 算法的发展历史和多种变体，希望能克服该算法中所存在的不足，获得性能更加完善的K-means算法[23]。

K-means 算法在 n 个样本数据中随机选取 k 个样本数据作为初始聚类中心点，而对于剩余的其它样本数据，根据与所选的各聚类中心点的相似度或者距离，将它们分别分配给相似度最高或是距离最近的类中；然后再计算每一类中样本数据的平均值，更新聚类中心点。不断重复这个过程。直到准则函数 J 开始收敛。

$$J=\sum_{i=1}^{k}\sum_{j=1}^{n_k}(c_i-x_j)^2 \tag{2.4}$$

其中 J 表示所有类中样本对象的平均误差的总和，$c_i$ 表示第 i 类中的聚类中心点，$x_j$ 表示第 i 类中的样本对象。

K-means 算法优点：

(1)该算法简单易懂，运算速度快，且不需要对数据进行范围约束；

(2)对于相互独立的样本数据，能获得较为精确的聚类结果。

K-means 算法缺点：

(1) k 值得选取是困难的，同一簇中的对象不能很好的体现其相关性；

(2)选取不同的初始簇中心最终所得的聚类结果可能不同；

(3)K-means 算法的本质上是一种面向非凸数优化的贪婪下降求解算法，不能得到全局最优解。

K-means 算法步骤：

算法输入：数据集 $X$ ， $X=\{x_m\}_{m=1}^{n}$, 聚类数目k

算法输出：聚类代表集合 $C, C=\{c_i\}_{i=1}^{k}$

Step1：从数据集 $X$ 中，任意选择 k 个数据对象作为初始簇中心；

Step2：利用公式 $dis(x_m,c_i)=\sqrt{(x_m-c_i)^2}$ 计算数据集中每个样本 $x_m$ 到聚类中心点 $c_i$ 的距离；

Step3：找到每个数据对象 $x_m$ 到聚类中心 $c_i$ 的最小距离 $\min\_dis(x_m,c_i)$，并将数据对象 $x_m$ 归为与 $c_i$ 相同的类中，即为 $C_i=\{x_m : dis(x_m-c_i)<dis(x_m-c_j), 1\le j\le k\}$；

Step4：计算同一类中对象的均值，更新聚类中心，

Step5：重复步骤 Step2~Step4，直到所得簇中心不再发生变化或达到最大运行次数。

为了更直观地表示 K-means 算法的步骤，图 2.5 给出了其操作流程图：

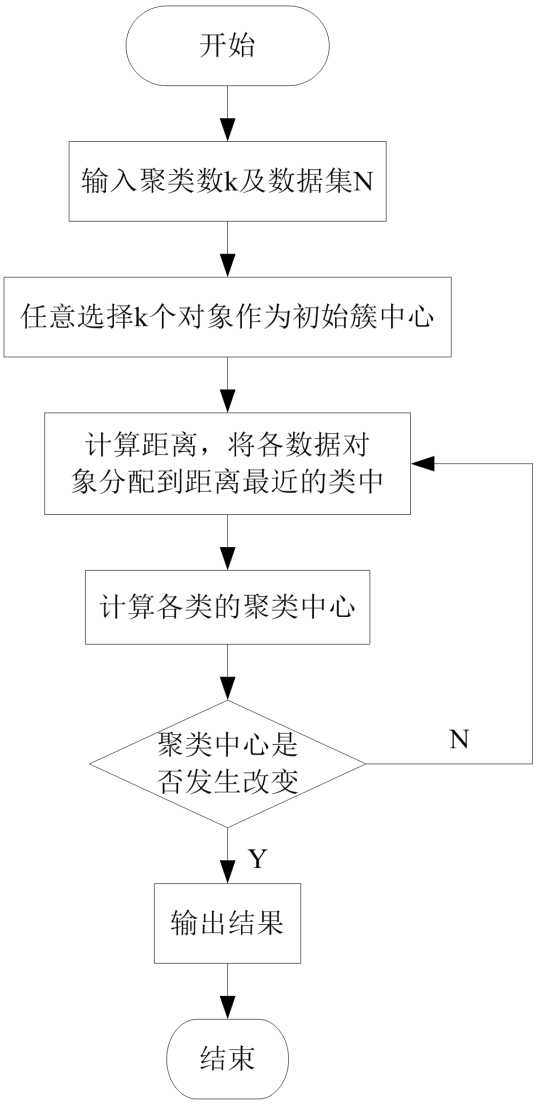



图 2.5　K-means 算法流程图

### 2.3.2 K-means 演变算法

K-means 算法作为基于划分的聚类方法中最为典型的一个聚类算法，在该算法的发展历程中，为满足各式各样的需求分析，从而人们提出了各式各样的改进方法。并且已在许多的领域得到了有效的应用，如图形图像分析、病毒数据的检测等。虽然 K-means 算法得到了广泛的应用，但是仍存在很多不足，如 K-means 算法中初始聚类中心选取敏感、难以预先确定聚类中心的数目等。因 K-means 算法中所存在的各优点及其不足，从而吸引了许多的研究人员对其进行研究。并且根据实际的应用背景，研究人员从不同的方面对其进行改进，衍生出了许多改进的 K-means 算法。

如文献[29,30]改进的 K 中心算法是对每个类随机选取一个初始样本对象，并将剩余样本对象归为与代表对象距离最近的类中，然后不停的通过代表对象与非代表对象之间的相互替换，从而改进聚类的效果。文献[31,32]中的 HK-means 算法，这个算法的思想是先从样本中进行采样，对采样的样本数据采用 K-means 算法进行聚类，将得到的聚类中心再作为整个样本数据集的初始聚类中心，从而改进 K-means 算法随机选取初始聚类中心易陷入局部最优的缺点。文献[33]提出的 GWO-KM 算法，该算法是一种较为新颖的算法，该算法通过将将灰狼优化算法与 K-means 算法相结合来 K-means 算法进行改进，通过使用 GWO 算法全局搜索能力较强的特性来改善 K-means 算法中搜索能力不足的缺点。文献[34]提出的采用 BWP 指标评价聚类效果，从而确定最优的聚类中心的数目，该方法通过对聚类中心数目取不同值时，比较 BWP 指标的数值，当 BWP 值最小时，则认为所得聚类效果最优，从而认为此时的聚类中心数目最佳。文献[35]提出了 BWACR 指标来搜索聚类中心的数目，该算法是通过对样本数据进行阈值分层来确定聚类中心数目的搜索范围上限，在通过聚类的有效性评价指标 BWACR 对所得到的每个类之间的相似度进行综合性评价，从而在搜索范围内获得最佳的聚类数目。文献[36]提出的 Rt-kmeans 算法，Rt-kmeans 算法是在原始 K-means 算法基础上改进得新型的算法，该算法最大的特点是通过假设在极端情况下，最小类间距离等于最大类内来计算阈值 d，当最大类内距离大于阈值 d 时，则将该簇进行划分；当最小类间距离小于阈值 d，则进行合并，从而自动的确定簇的个数。如针对网络入侵中所存在的问题，文献[37]中将随机森林算法与权重 K-means 算法相结合使用来解决误用检测和异常检测中存在的缺点，首先通过将随机森林分类算法用在误用检测中，从训练数据集自动构建入侵模式，然后将网络连接与这些入侵模式匹配，以检测网络入侵。即通过训练集将网络连接分类为入侵和正常数据。而在异常检测则使用 K-means 对异常数据进行聚类来识别新型入侵方式，从而加入训练集中。如文献[38]提出使用 K-means 的改进算法 K-means++来对的网页的缺陷进行预测，K-means++算法中采用概率来计算聚类中心,且在大多数情况下 K-means++相比 K-means 算法而言性能增强了 20%~70%之间。文献[39]中提出了一种优化的 k-均值重心的初始化方法，该算法通过使用分而治之的方法来发现初始中心和分配数据到合适的簇中。文献[40]改进 K-means 算法来对股价进行预测，主要通过将传统 K-means 算法与层次聚类算法相结合来对股票的价格变化趋势进行预测。除了以上几种改进的算法以外，K-means 算法也与其他的很多智能算法相结合使用，如文献[41]中一种改进的 K-means 蚁群聚类算法、文献[42]提出对的蜂群 K-means 聚类算法的改进以及研究、文献[43]中的基于差分演化的 K-均值聚类算法等等。

但这些被提出的算法中，绝大部分算法都是从如何优化初始簇中心去考虑，只有很少一部分考虑样本数据集间的相关性。如文献[44]中基于马氏距离的 K 均值聚类算法，该算

法虽然在一定程度上考虑的对象间的相关性，但它只从两两对象来考虑，依旧没有从整体数据集上考虑对象间的相关性。

## 2.4 本章小结

本章中的内容主要分为三部分来进行叙述。第一部分主要对数据挖掘的发展历程进行了简短介绍，其中包含数据挖掘概念、流程以及数据挖掘的任务。紧接着在第二部分中谈论了数据挖掘中的常用挖掘技术——聚类分析方法，该部分首先介绍了什么是聚类分析，紧接着介绍聚类分析的原理，然后是对聚类分析的过程进行了简短回顾，在聚类分析的过程中，提及了几种对于数据集规范化的标准，其中着重介绍了离差标准化方法与标准差标准化方法。然后介绍了聚类方法的分类，并对各类方法所涉及的主要原理进行了简短介绍。在本章的最后介绍了几种对于聚类效果的评价方法，其中对 Purity 评价法进行了详细的介绍，同时该评价方法也是本文中对于聚类效果的评价标准。

本章节的第三部分中，主要介绍了传统 K-means 算法的基本思想，并对该算法的流程图以及算法步骤进行了简要的说明，同时对该算法中所存在的优缺点进行了分析。最后在本章节中对一些改进的经典的聚类算法及最新改进 K-means 算法进行了简短介绍。

# 第三章 K-means 算法的研究与改进

## 3.1 初始化

### 3.1.1 问题的提出

由于 K-means 算法是一个迭代型的聚类算法，而迭代型方法易收敛到众多的局部最优解。如图 3.1 所示，图 3.1 中是对于一个数据集进行聚类，因陷入局部最优从而导致产生不同的聚类效果。并且，当所获得的初始聚类中心如果与最终的聚类中心相差较大，那么该算法在收敛的过程中，将会较大程度的增加算法的迭代次数。因此，对于迭代型聚类算法的初始启动条件是特别的敏感，一个精炼的初始聚类中心的设置将使得算法能达到更好的聚类效果。并且对于大规模数据集而言，如果所选取的初始聚类中心的位置与最终真实位置比较接近，那么该算法在运行的过程中能够较大程度的节约程序的运行次数[45,46]。现今，虽然已有较多改进的 K-means 算法被提出，并且在这些被提出的算法中，绝大部分都是对传统 K-means 算法的初始聚类中心的选取问题进行改进，但在改进的过程中，许多改进的算法依旧没有很好的解决传统 K-means 算法初始聚类中心选取敏感的问题。很多算法在对初始聚类中心选取进行改进的过程中，经常忽略对第一个初始聚类中心点选择的问题，常以平均值或是距平均值最近样本对象作为第一个初始聚类中心，然而有时这将与聚类中心真实位置相差较大。例如最大最小距离初始化方法，该初始化方法往往选择边界的样本对象或异常样本对象作为初始聚类中心。但是，在数据对象的真实位置分布中，实际上仅除了一些离群点外，数据对象的聚类中心基本上是很少分布在边界上，绝大多数的聚类中心都是分布在数据对象密度较高的范围内。因此，边界上的初始聚类中心不能很好的代表簇的分布情况，并且可能导致聚类产生不好的结果。

### 3.1.2 方法的改进

为解决传统 K-means 算法因采用随机生成初始聚类中心而导致聚类结果不稳定的问题。本文通过研究以后将最大最小距离初始化方法与各样本对象间的密度相结合来对 K-means 算法进行初始聚类中心的选取。本文在使用样本对象间的欧式距离与样本对象所处位置的密度作为双重标准来对初始聚类中心位置进行选取，其中对于第一个初始聚类中心，不再随机生成或是选择距平均值距离最近的对象作为第一个初始聚类中心，本文通过搜索一个密度较高的样本对象来代替距平均值最近的样本作为第一个初始聚类中心，从而解决第一个初始样本选取局限性的问题。然后对于剩余的样本对象，将其划分到已选择聚

类中心所属类别中，搜索距所属聚类中心距离最远的样本对象作为下一个临时初始聚类中心，计算该临时初始聚类中心的密度，并选择临时初始聚类中心密度范围内的所有样本对象的距平均值最近的样本对象作为更新的临时初始聚类中心，若临时初始聚类中心不再发生改变，则不再重复该过程，直到确定所有初始聚类中心。从而使得对于所研究的样本对象中能够搜索出分布尽量均匀，各初始聚类中心密度相对较高且是唯一的初始样本集。使所得的初始聚类中心尽量的逼近真实聚类中心，从而为最终的聚类效果提供有效的保证，并且减少算法在聚类过程中的迭代次数。

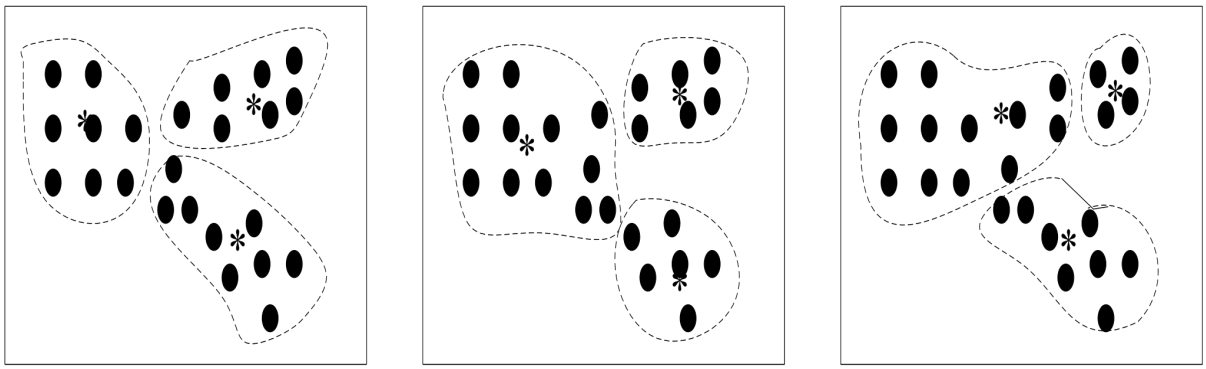



图 3.1　局部最优解

距离定义：

$$dis(x_i, x_j) = (x_i - x_j)^2 \tag{3.1}$$

其中 $dis(x_i, x_j)$ 表示 $x_i$ 与 $x_j$ 之间的欧式距离。

密度定义：

$$N_\lambda(x_m, c_i) = \{x_m \mid dis(x_m - c_i) \le \lambda\} \tag{3.2}$$

$N_\lambda(x_m, c_i)$ 表示距对象 $c_i$ 距离小于 $\lambda$ 的所有样本对象。

平均密度：

$$\overline{N_\lambda(x_m, c_i)} = \frac{1}{n} N_\lambda(x_m, c_i) \tag{3.3}$$

其中 $\overline{N_\lambda(x_m, c_i)}$ 表示平均密度，$n$ 表示 $N_\lambda(x_m, c_i)$ 中的样本对象数。

改进初始化方法步骤如下：

Step1：建立一个包含 n 个数据对象的数据集 $X$，$X = \{x_m\}_{m=1}^n$，并计算整个数据集的平均值 $\bar{x}$；计算距平均值距离最远点为 $c_1$；

Step2：令 $\lambda = \frac{1}{2} * dis(c_1, \bar{x})$，搜索距 $c_1$ 距离不大于 $\lambda$ 的所有数据对象 $N_\lambda(x_m, c_1)$；

Step3：计算 $N_\lambda(x_m,c_1)$ 的平均值，选择距平均值最近的对象作为更新的聚类中心。若 $c_1$ 不发生变化，则 $c_1$ 作为第一个初始聚类中心，否则返回 Step2；

Step4：对于剩余对象，将其划分到已选择聚类中心所属类别中，搜索距所属聚类中心距离最远的对象作为下一个临时聚类中心 $c_i$；

Step5：令 $\varepsilon$ 为所选对象到所属聚类中心距离的一半。搜索距 $c_i$ 距离不大于 $\varepsilon$ 的所有数据对象 $N_\varepsilon(x_m,c_i)$；

Step6：计算 $N_\varepsilon(x_m,c_i)$ 的平均值，选择距平均值最近的对象作为更新的聚类中心。若 $c_i$ 不发生变化，则 $c_i$ 作为下一个初始聚类中心，否则返回 Step5；

Step7：重复 Step4~Step6 步，直到找到 k 个聚类中心。

### 3.1.3 实验结果的分析

本文中，为了改进方法的合理性及其有效性，文中所进行的研究均在仿真硬件环境为 AMD phenom CPU 2.6GHz，内存为 4GB；软件环境为 MATLAB 2014a 的环境下进行实验分析。

在本章节中，为了验证本文初始化方法的准确性，主要采用了两组二维合成数据集进行实验分析，其中每组合成的数据集均包含 3 个类别。合成数据集 1 中包含 60 个样本数据，该样本数据集是通过固定范围以后进行随机生成；合成数据集 2 中包含 150 个样本数据，该样本数据集是通过对 iris 数据集采用主成分分析的方法进行降维以后所获得。两组合成数据集分布情况如图 3.2 所示。

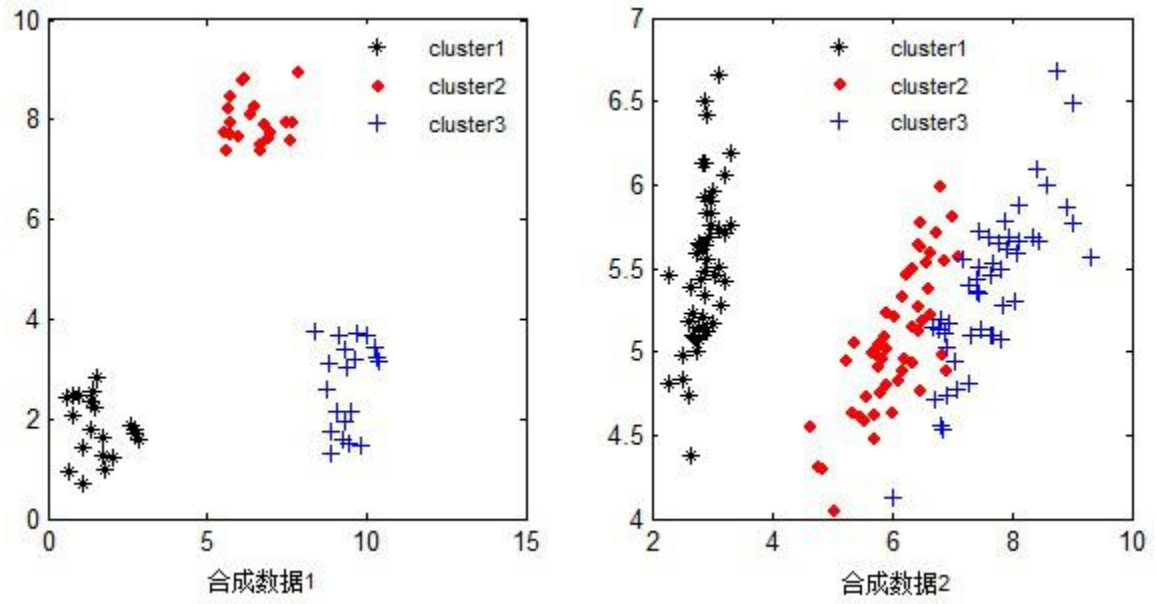



图 3.2　两组合成数据集

在对改进初始化方法进行实验效果比较的过程中，通过采用随机初始化方法、最大最小距离初始化方法、文献[36]中 Rt-kmeans 算法的初始化方法以及本文所改进的初始化方法，这 4 种初始化方法分别对两组合成数据集进行初始化分析，观察各初始化方法在两组合成数据集上所得的初始聚类中心的位置。两组合成数据集所得初始聚类中心位置分别如图 3.3 与图 3.4 所示。然后，本节中将这四种初始化方法应用于 K-means 算法分别对两组合成数据集进行聚类分析，为保证实验的有效性及消除实验的随机性，本文进行 10 次独立实验并记录实验的聚类准确率。聚类准确率如表 2 所示。除此以外，本文还对合成数据集中每次实验结果进行分析，记录每次聚类迭代次数与聚类准确率。迭代次数如图 3.5 所示，聚类准确率如图 3.6 所示。

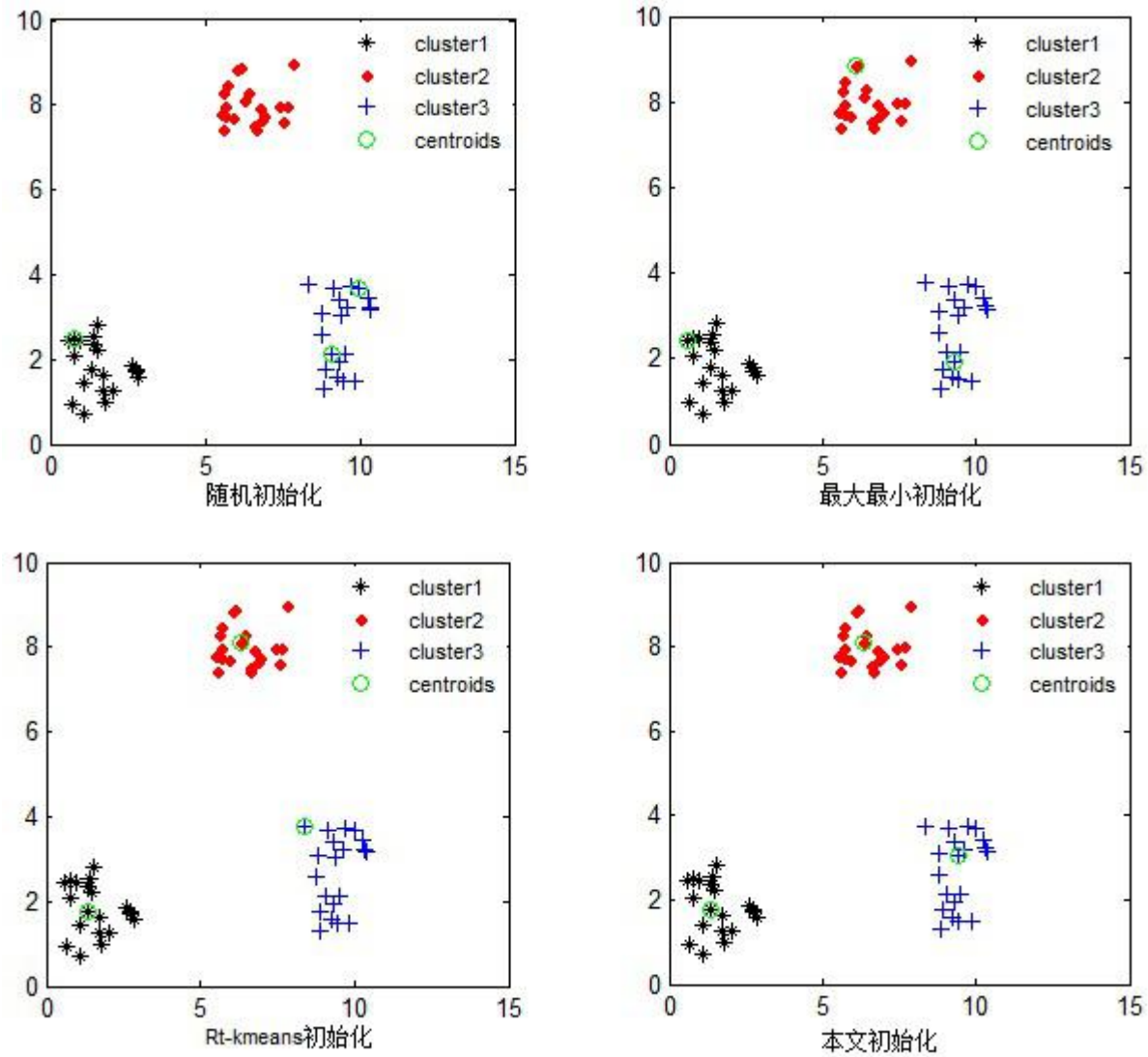



图 3.3　各初始化方法对合成数据集 1

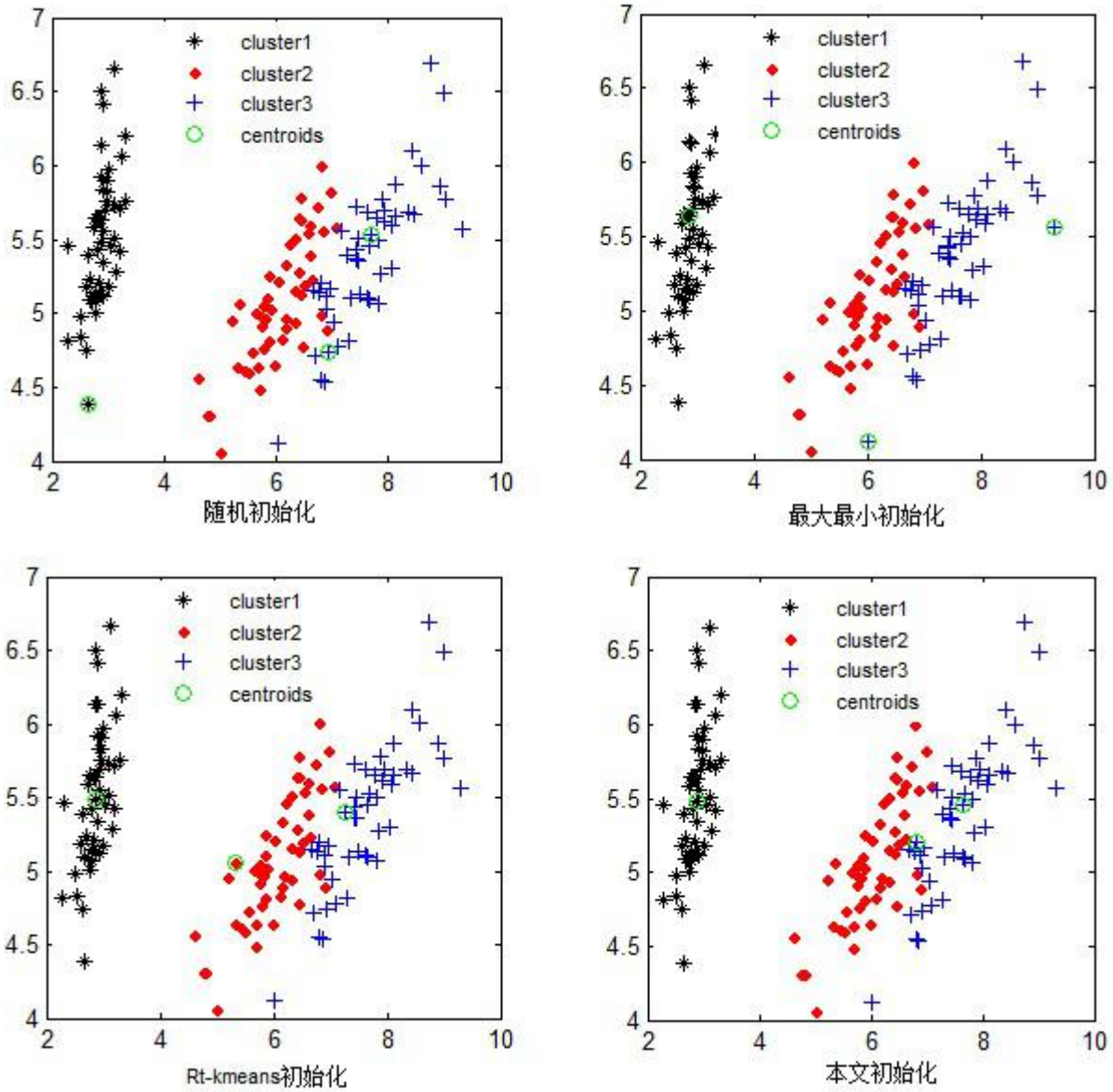



图 3.4　各初始化方法对合成数据集 2

从对两组合成数据集的初始化结果可知。随机初始化所得到的初始聚类中心位置分布最差，最大最小初始化方法相比随机初始化方法所得到的初始聚类中心位置稍好，但是通过图 3.3 与图 3.4 对比可知，对于两组合成数据集，最大最小初始化方法所得到的初始聚类中心大部分都是分布在样本数据集的边界位置，只是很少一部分分布在各簇的中心位置，并且所得的各初始聚类中心与最终的真实聚类中心位置差距都比较的大。而对于文献[36]中 Rt-kmeans 算法的初始方法，从图 3.3 与图 3.4 中可知，虽然所得的聚类结果相比前面两种方法所得的结果都更好，但是无论是对于合成数据 1 还是合成数据 2，都仍然存在处于样本边界的初始点，而该点并不能很好作为聚类的代表簇中心。而对于本文的初始法

可以看到，在对合成数据 1 的初始化中，所得的初始中心基本在真实聚类中心所在位置，并且对于数据集 2 的初始化结果同样可以看出，各初始点所在位置同样非常逼近真实位置。所以可以得出，本文的初始点在对数据集进行初始化时是非常有效的，同时也证明了本文方法选值的合理性。

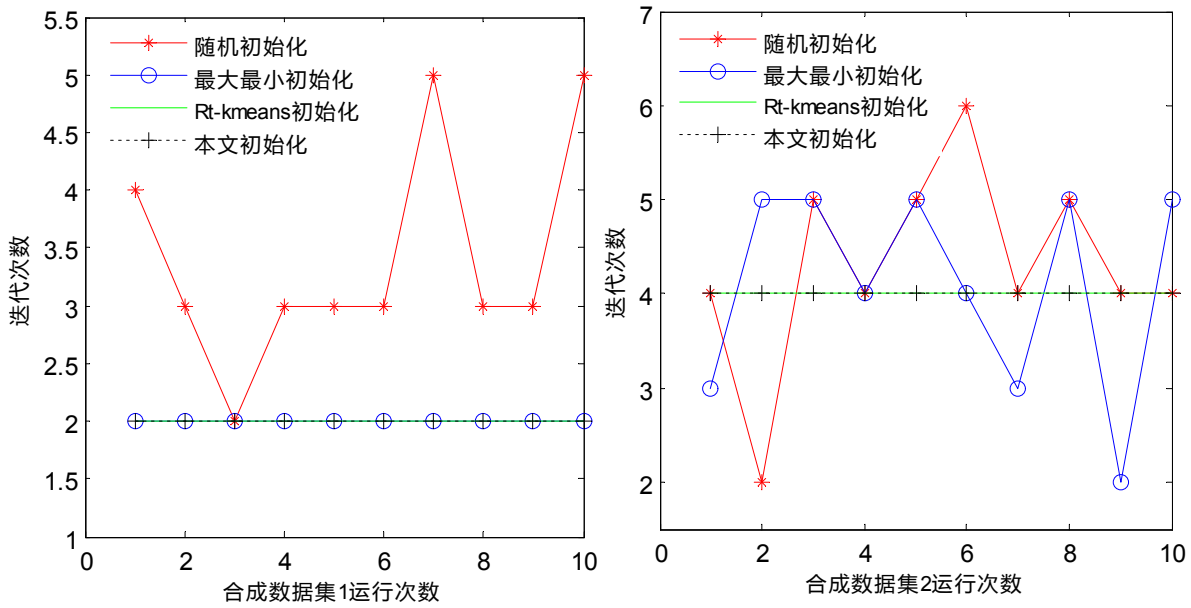



图 3.5 各初始化法下合成数据集的迭代次数

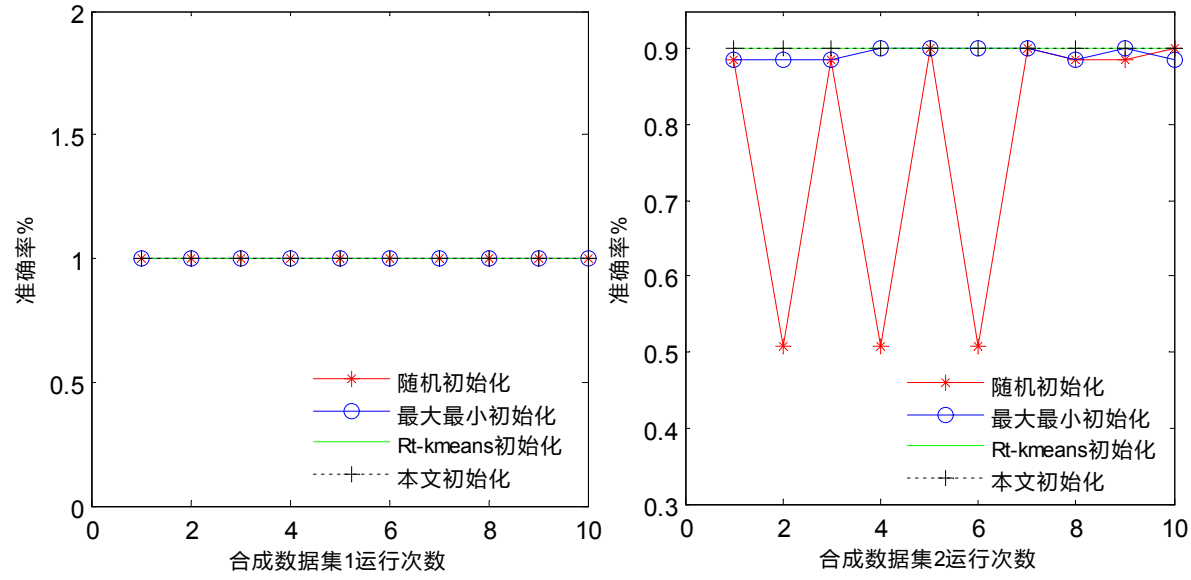



图 3.6 各初始化法下合成数据集的准确率%

表 3.1 各初始化方法下各数据集的准确率（%）

| 算法 | 合成数据 1 | | 合成数据 2 | |
|---|---|---|---|---|
| | 准确率% | 迭代次数 | 准确率% | 迭代次数 |
| 随机初始化 | 100.00 | 3.3 | 77.67 | 4.3 |
| 最大最小初始化 | 100.00 | 2.0 | 89.07 | 4.1 |
| Rt-kmeans 初始化 | 100.00 | 2.0 | 90.00 | 4.0 |
| 本文初始化 | 100.00 | 2.0 | 90.00 | 4.0 |

从图 3.5 中可以看出，应用本文的初始化方法对合成数据集聚类时，相比传统的 K-means 方法，本文只需要较少的迭代次数就能使数据集达到收敛状态，在合成数据集 1 中，本文初始法方法仅需要迭代 2 次就能使其收敛，对于数据集 2，本文也仅仅只需要 4 次就能得出聚类结果，但是对于随机初始化方法可以看出，对于合成数据 1，在所做的 10 次独立试验中，最小需要迭代 2 次，最大迭代了 5 次；而对于合成数据 2，最少需要 2 次，最大需要迭代 6 次，通过对两组数据集的聚类结果可以看出，随机初始化方法每次实验所需迭代次数不稳定，通过分析我们可以知道，这很大程度上是因为每次试验中，每次实验中所得初始化点不一致所造成的，然而，对于本文的初始化方法，却不存在这种情况，本文方法每次所得的结果都比较的稳定。并且，在本文的实验中，因所使用的数据集数据量还较少，所以使用随机初始化方法对算法的运行效率影响还不是非常的大，但是当所分析的数据容量较大时，采用随机初始化方法将大大的增加算法的迭代次数。而本文的初始法方法将会更大程度的减少其聚类的迭代次数，达到更高的效率。从图 3.6 中可以看出，采用本文的初始化方法对合成数据集聚类的结果相比随机初始化下聚类的结果，本文的初始化方法能够得到更好的准确率。

## 3.2 k 值自适应划分

### 3.2.1 问题的提出

聚类是将样本对象根据各自的属性分配到不同的类别中。然而，在对新兴事物的研究过程中，将无法提前预知所研究对象的类别数目，而传统 K-means 算法在对数据集聚类的过程中却需要提前确定聚类中心的数目 k。图 3.7 是当输入不同聚类中心数目 k 时，传统 K-means 算法对合成数据集 1 的聚类结果，从图结果中可知，只有当输入的聚类中心数目 k 的值为 3 时，所得聚类结果与数据集的原始分布情况一致，结果将是有效的；而当输入 k 值为 4、5 时，所得到的聚类结果将无效。对此，确定适合的聚类中心数目 k 将非常的有必要。针对此问题，本文是在不提前预知聚类中心数目 k 的情况下，通过对样本对象间相

似程度的研究确定聚类过程内类距离与类间距离之间的关系，从而自动的确定聚类中心数目 k。

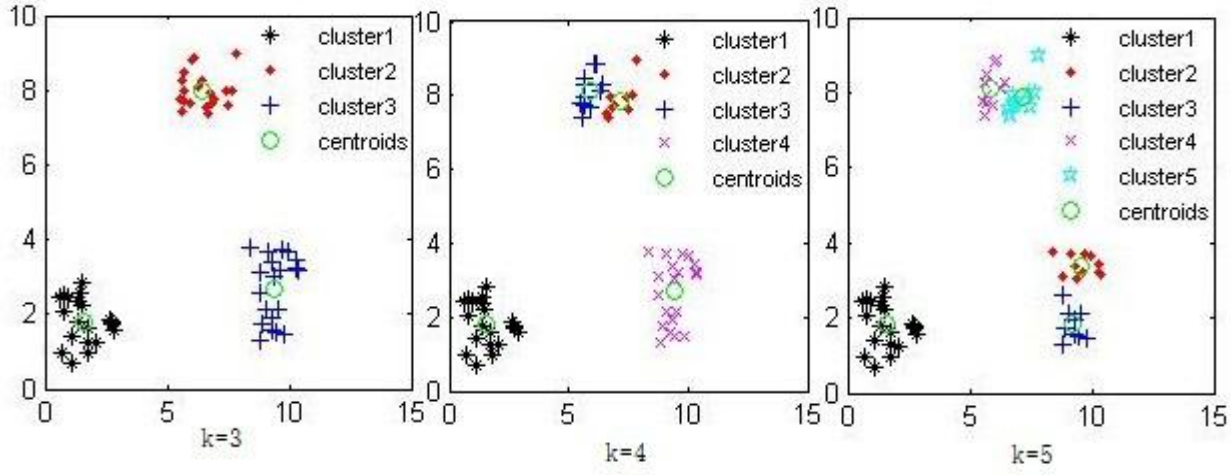



图 3.7　输入不同 k 的聚类结果

### 3.2.2 方法的改进

从图 3.7 以及上文的叙述中可知，准确的确定聚类中心数目非常的重要，在聚类的过程中，如果所选取的聚类中心数目与真实聚类中心数目差距较大，那么聚类的结果将会造成较大的偏差甚至是错误。因此如何的去增强聚类中心数目的准确性将变得非常的重要。

由于在聚类的过程中，聚类的效果主要是通过目标函数来进行衡量，而现今对于聚类的划分过程中并没有一个固定标准的划分准则。因此研究人员常通过对样本对象间的相似度来对其进行类别的划分。也就是使得对于相同类中的样本对象能够尽可能的相似，而对于不同类之间的样本对象间差异度尽可能的大，因此在这种情况下，要使得划分效果能够达到尽可能的好，则最小的类间距离应该不小于最大的类内距离。然而在 k-means 算法中，如果划分的类别数 k 的取值偏小时，那么对于整体模型也就越复杂，越容易发生过拟合现象；但如果 k 的取值偏大，那么又将不能很好的体现出样本对象之间的相似性。因此，合理的聚类中心数目 k 的取值对于整个模型的聚类效果影响非常大。考虑这些问题，本文通过对 Rt-kmeans 算法的研究，在其自动划分与合并的基础上通过改变对划分阈值的定义，来进一步的提高算法自动确定聚类中心数目的准确度。

最小类间距离：

$$Inter = \min_{i \neq j} \{dis(c_i, c_j)\} \tag{3.4}$$

最大类内距离：

$$Intra = \max_{i} \{\max_{x_i \in c_j} \{dis(x_i, c_j)\}\} \tag{3.5}$$

其中 $x_i$ 是以 $c_j$ 为聚类中心的对象。

平均类间距离：

$$\overline{d} = \frac{1}{k^2}\sum_{j=1`}^{k}\sum_{i=1`}^{k} dis(c_i, c_j) \tag{3.6}$$

其中 k 表示聚类中心的数目，$c_i$ 和 $c_j$ 表示所选择的聚类中心点。

因为 $\overline{d}$ 为平均类间距离，所以存在 $a_1, a_2$ 分别使得 $Inter = a_1 \times \overline{d}$ ，$Intra = a_2 \times \overline{d}$ 。聚类过程中应使得最小类间距离及最大类内距离与理想最小类间距离与理想最大类内距离尽可能接近。因此存在如下关系式。

$$y = \min\{(a_1 \times \overline{d} - d_1^{*})^2 + (a_2 \times \overline{d} - d_2^{*})^2\} \tag{3.7}$$

其中 $d_1^{*}$ 为理想划分的最小类间距离，$d_2^{*}$ 理想划分的最大类内距离。

因为最小类间距离小于等于平均类间距离，最小类间距离越大，类与类之间的差异性越明显，则聚类效果越好，所以选用当前划分的平均类间距离来近似理想划分的最小类间距离，即 $d_1^{*} = \overline{d}$ 。最大类内距离越小，类中对象相似性越强，聚类效果越好，因此理想划分下的最大类内距离不妨为 0，即 $d_2^{*} = 0$ 。

另一方面，在聚类过程中，随着类别数的增多，类与类之间将变得越来越密集，假设此时对于相邻的两类之间边界样本间距趋近于 0 时。此时存在 $a_3 + a_4 \geq a_5$，其中 $a_3$ 、$a_4$ 为内类距离，$a_5$ 为类间距离。而聚类过程中，最大的内类距离应尽可能小，既相邻两类间内类距离之差应尽可能小，不妨建立 $a_3 = a_4 = 0.5 \times a_5$，可得 $Inter = 2 \times Intra$，既 $a_1 = 2 \times a_2$ 。关系示意图如图 3.8 所示。

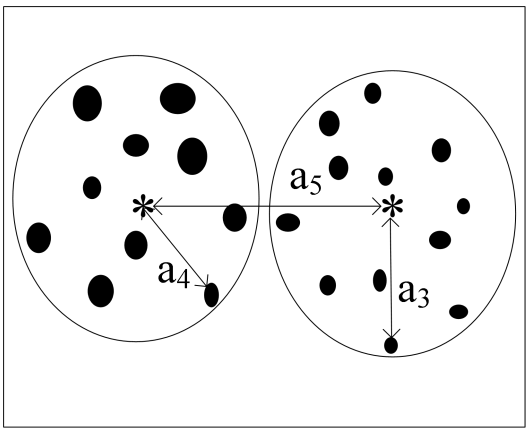



图 3.8　类内与类间距离示意图

将$a_1 = 2 \times a_2$代入公式（3.7），可得：

$$y = \min\{\frac{5}{4}d^2[(a_1 - \frac{4}{5})^2 + \frac{4}{25}]\} \tag{3.8}$$

因此当$y$取得最小时，得$a_1 = \frac{4}{5}$，$a_2 = \frac{2}{5}$。由此可构造如下启发式规则，用于样本划分：（1）如果一个簇中，存在类内距离大于$\frac{4}{5}\overline{d}$时，认为该簇相似度较低，则将该簇划分为不同的两个簇；（2）如果两个簇间，当类间距离小于$\frac{2}{5}\overline{d}$时，则认为这个两个簇之间相似度较高，应对其进行合并。应用以上规则可构造算法(A-kmeans)，具体步骤描述如下。

算法输入：包含 n 对对象的数据集$X$，$X = \{x_m \mid m = 1,2,\cdots n\}$，初始聚类中心数目$k_0$，点密度领域半径$\lambda$；

算法输出：聚类代表集合$C, C = \{c_i \mid 1,2,\cdots k\}$；

Step1：初始化$k = k_0$，按本文 3.1 节改进初始化法选取密度高的$k_0$个样本为初始聚类中心$c_i$；

Step2：按传统 K-means 算法流程聚类，计算平均类间距离$\overline{d}$；

Step3：如果存在$Intra > \frac{4\overline{d}}{5}$,再将该类进行划分，更新$k = k + 1$，返回 Step1；

Step4：如果存在$Inter < \frac{2}{5}\overline{d}$，在将该类进行合并，更新$k = k - 1$，返回 Step1，但不再执行 Step3；

Step5：如果$k$不变，停止。

### 3.2.3 实验结果的分析

为了验证提出算法的有效性及合理性。本文选用 UCI 数据库中的六组数据集做测试数据，数据集如表 3.2 所示。在相同软硬件环境下与传统 K-menas 算法、文献[34]提出的算法、文献[36]提出的 Rt-kmeans 算法进行实验，对算法的聚类准确率、标准差、收敛时间、k 值搜索上界、最终确定的聚类中心数目五方面进行比较。

表 3.2 UCI 实验数据集

| 数据名称 | 类别数 | 属性个数 | 样本数 | 类别分布 |
|---|---|---|---|---|
| (Statlog)heart | 2 | 13 | 270 | 120、150 |
| Haberman | 2 | 3 | 306 | 225、81 |
| Wine | 3 | 13 | 178 | 59、71、48 |
| Hayes-Roth | 3 | 5 | 132 | 51、51、30 |

| | | | | |
|---|---|---|---|---|
| Seeds | 3 | 7 | 210 | 70、70、70 |
| Yeast | 10 | 8 | 1484 | 463、429、244、163、51、44、35、30、20、5 |

本文的验证过程分为三步。首先是将本文 A-kmeans 算法与传统 K-means 算法对合成数据集 1 进行分析，当输入不同的聚类中心数目时聚类结果如图 3.8 与图 3.10 所示。其次是对 A-kmeans 算法对于确定聚类中心数目的鲁棒性及其精确性的分析。将 A-kmeans 算法与文献[36]算法(Rt-kmeans)算法，当输入不同的初始聚类中心数目 $k_0$ 时，比较 A-kmeans 算法与文献[36]算法(Rt-kmeans)算法对于确定聚类中心数目的稳定性与精确性。本文实验中初始聚类中心数目分别取 2、5、8 进行对比，结果如表 3.3 所示。最后是对本文所选的五个评价指标进行实验分析，实验中的聚类准确率，收敛时间是通过 10 次单独实验取平均值进行记录。

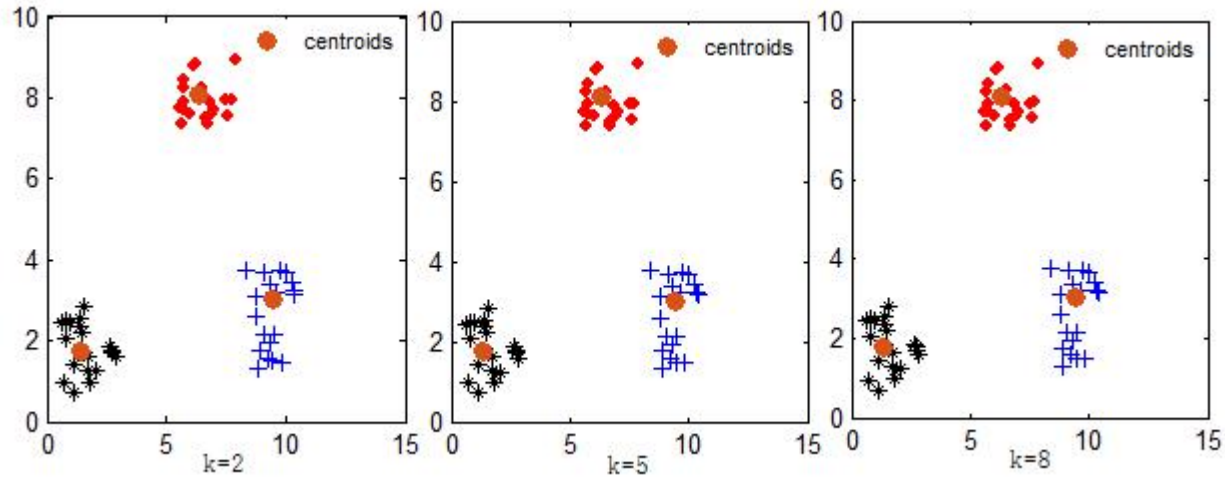



图 3.9　A-kmeans 算法对不同 k 值时的聚类结果

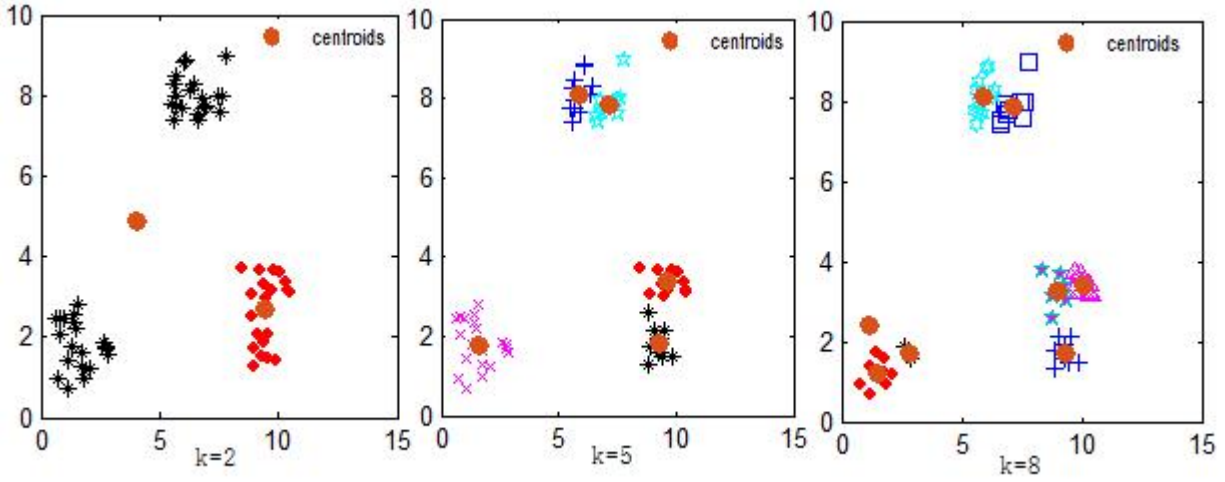



图 3.10　传统 K-means 算法对不同 k 值时的聚类结果

表 3.3 不同 k 值时各数据集在 Rt-kmeans 与 A-kmeans 算法下的聚类中心数

| 数据集 | 标准聚类中心数 | 文献[36]算法(Rt-kmeans) | | | A-kmeans 算法 | | |
|---|---|---|---|---|---|---|---|
| | | k=2 | k=5 | k=8 | k=2 | k=5 | k=8 |
| (Statlog)heart | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| Haberman | 2 | 2 | 2 | 2 | 3 | 3 | 3 |
| Wine | 3 | 2 | 2 | 2 | 3 | 3 | 3 |
| Hayes-Roth | 3 | 2 | 3 | 3 | 3 | 3 | 3 |
| Seeds | 3 | 2 | 2 | 2 | 3 | 3 | 3 |
| Yeast | 10 | 2 | 2 | 2 | 5 | 5 | 7 |

表 3.4 四种算法聚类准确率 %

| 数据集 | 标准聚类中心数目 | 传统 K-means | 文献[34]算法(BWP) | | 文献[36]算法(Rt-kmeans) | | A-kmeans 算法 | |
|---|---|---|---|---|---|---|---|---|
| | | | 最终聚类中心数目 | 准确率 | 最终聚类中心数目 | 准确率 | 最终聚类中心数目 | 准确率 |
| (Statlog)heart | 2 | 59.25 | 3 | 45.48 | 2 | 60.74 | 2 | 61.11 |
| Haberman | 2 | 51.11 | 3 | 41.96 | 2 | 74.51 | 3 | 45.42 |
| Wine | 3 | 67.52 | 3 | 67.52 | 2 | 65.73 | 3 | 67.98 |
| Hayes-Roth | 3 | 44.62 | 3 | 44.62 | 2 | 41.66 | 3 | 42.42 |
| | | | | | 3 | 43.18 | | |
| Seeds | 3 | 89.14 | 3 | 88.90 | 2 | 66.19 | 3 | 90.48 |
| Yeast | 10 | 38.61 | 4 | 41.86 | 2 | 37.67 | 5 | 42.16 |
| | | | | | | | 7 | 43.06 |

表 3.5 四种算法标准差和收敛时间/s

| 数据集 | 评价指标 | 传统 K-means | 文献[34]算法(BWP) | 文献[36]算法(Rt-kmeans) | A-kmeans 算法 |
|---|---|---|---|---|---|
| (Statlog)heart | 标准差 | 0.0251 | 0.0342 | 0 | 0 |
| | 收敛时间 | 0.0933 | 1.0564 | 0.3168 | 0.7571 |
| | k 搜索上界 | -- | 16 | 3 | 3 |
| Haberman | 标准差 | 0.0174 | 0.0499 | 0 | 0 |
| | 收敛时间 | 0.0785 | 1.3022 | 1.0309 | 1.1425 |
| | k 搜索上界 | -- | 17 | 5 | 7 |
| Wine | 标准差 | 0.0651 | 0.0568 | 0 | 0 |
| | 收敛时间 | 0.0095 | 0.4719 | 0.4502 | 0.5216 |
| | k 搜索上界 | -- | 13 | 3 | 3 |
| Hayes-Roth | 标准差 | 0 | 0 | 0 | 0 |
| | 收敛时间 | 0.0094 | 0.2680 | 0.1274 | 0.1137 |
| | k 搜索上界 | -- | 11 | 3 | 9 |
| Seeds | 标准差 | 0.0021 | 0.0024 | 0 | 0 |
| | 收敛时间 | 0.0086 | 0.6357 | 0.2544 | 0.7004 |
| | k 搜索上界 | -- | 14 | 3 | 4 |
| Yeast | 标准差 | 0.0273 | 0.0437 | 0 | 0 |
| | 收敛时间 | 0.2459 | 45.2719 | 14.6657 | 7.3690 |
| | k 搜索上界 | -- | 38 | 10 | 5 |

由图 3.9 与图 3.10 对比可知，对于合成数据集 1，当输入不同的聚类中心数目时，采用传统的 K-means 算法所得到的聚类结果不一致，从图 3.10 中的聚类结果可明显看出，三次所输入的聚类中心数目都不准确；所得的聚类结果都不是有效的；但是由图 3.9 聚类结果可知，虽然每次所输入的聚类中心数目不一致，但是最终所得聚类结果都是一样的。

由实验表 3.3 结果可知，A-kmeans 算法与文献[36]算法(Rt-kmeans)算法相比，当输入初始聚类中心数目 k 的值不同时，对于文中所采用的大多部分数据集，A-kmeans 算法最终所得的聚类中心数目都比较的稳定，并且与采用实验数据集的标准类别数都比较的接近。而文献[36]的 Rt-kmeans 算法虽然对于大部分数据集所得结果也比较接近标准类别数，但是与本文的方法进行对比，最终所得结果还是稍显不足。

通过表 3.4 与表 3.5 进行分析可知，传统 K-means 算法虽然收敛速度较快，但是从六组数据集的聚类结果可知，采用传统 K-means 算法聚类所得到的结果不稳定，并且聚类的准确率也较差。除此以外，传统 K-means 算法聚类过程中需要提前输入准确的聚类数目。文献[34]算法在一定程度虽然解决了预先确定聚类中心数目的问题，但是该算法需要较大范围的对不同的 k 值进行搜索与比较，从中寻找出最优的聚类中心数目 k，因此该算法相比其它几个算法，收敛速度明显耗时过长。Rt-kmeans 算法虽然收敛速度也较快，但该算法对于确定聚类数目的准确度以及聚类效果的准确率都有着明显的不足。并且从 Yeast 数据集的聚类结果看出，当数据集类别数目较多时，文献[34]与文献[36]中的算法所得的最终聚类中心数目的结果与标准聚类中心数目结果差别较大，虽 A-kmeans 算法也没有的到标准聚类数，但分析 Yeast 数据集发现，Yeast 数据集的样本数据分布并不均匀，由表 3.2 可知 Yeast 数据集后四类样本数仅为总样本数的 7%，其余六类样本占了总样本数的 93%，所以 A-kmeans 算法所得聚类结果还是比较可靠的。相比其他三个算法，A-kmeans 算法无论是从聚类准确率、确定聚类中心数目以及收敛时间等方面分析都能得到较为优越的聚类结果。通过上述实验结果表明，A-kmeans 算法对于确定聚类中心数目、提高聚类准确率、提高收敛速度都更为的有效。

## 3.3 距离相关性加权

### 3.3.1 问题的提出

除此之外，为增强同一簇中数据间的相关性，本文通过相关性系数来对距离进行加权，使得距离稍远但相关性较高的样本对象能够被聚类在同一簇中；但对于距离稍近，但是相关性很低的点，使其能够聚类在不同的类中，通过相关性加权，保证在同一类中的样本对象相似性更高，而对于不同类中的样本对象，差异性更大。为了有效的说明本文所使

用方法的合理性，主要通过一个假设的示例对来本文改进的思想进行论证。假设对于一组已知类别的样本数据 A、B、C，其中 A、B 处在相同的类中，而样本对象 C 则单独属于一类，A=[3 2.3 1], B=[3 1 -2],C=[3 2 4]。通过计算可知该组样本中实际上 A、B 的间距与 A、C 的间距相差不大，但 A、B 间的相关程度却较高。若采用传统 K-means 算法来对改组样本进行分析，即度量量纲采用欧式距离，那么最终所得结果将是 A、C 处在相同类中，B 独自成为一类。这将与实际结果有较大偏差。然而，如果通过相关性对距离加权以后，那么最终到的结果将是 A、B 间距离最短，即可以得到更为真实的聚类结果。由此论证了本文提出方法的有效性，本文算法(C-kmeans)具体描述与论证过程将展示在下文。

### 3.3.2 方法的改进

为了使同一簇中的对象具有更强的相关性，本文通过采用 Pearson 相关系数来对数据对象间的距离进行加权。Pearson 相关系数计算公式如下[27,47]：

$$r=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=i}^{n}(y_i-\bar{y})^2}} \tag{3.9}$$

相关系数的取值范围：$-1\le r\le 1$

(1)Pearson 系数距离加权如下：

Step1：通过公式（3.10）计算数据对象间的相关性系数 $r=(x_m,x_i)$；

Step2：寻找数据集中，当 $m\neq i$ 时相关数系数绝对值最大值 Max_r 与绝对值最小值 Min_r;

Step3：通过离差标准化对数据对象间的相关性进行规范化，计算公式如下：

$$new\_r=\frac{|r|-Min\_r}{Max\_r-Min\_r+0.001} \tag{3.10}$$

其中本文 0.001 是为防止 $Max\_r=Min\_r$ 时分母为 0，式(3.10)分母为 0。

Step4：距离加权系数计算公式为：

$$cor=Q-new\_r+0.001 \tag{3.11}$$

其中 $Q$ 值取 1，0.001 是为防止 $Q=new\_r$ 式（3.11）为 0，使得加权距离为 0。

（2）相关相 K-means 算法

算法输入：数据集 $X, X=\{x_m\}_{m=1}^{n}$, 聚类数目 k

算法输出：聚类代表集合 $C, C=\{c_i\}_{i=1}^{k}$

Step1：从整个数据集 X 中，按照 3.1 章节的方法选取初始聚类中心；

Step2：利用公式 $d(x_m,c_i)=\sqrt{(x_m-c_i)^2}$ 计算数据集中每个数据对象 $x_m$ 到聚类中心 $c_i$ 的欧式距离；

Step3：按照公式（3.12）的方法计算数据对象 $x_m$ 与聚类中心 $c_i$ 的加权系数 $cor$；

Step4：利用 Step2 中算出的欧式距离 $\min\_d(x_m,c_i)$ 与步骤中的加权系数 $cor(x_m,c_i)$ 相乘作为数据数据对象 $x_m$ 到聚类中心 $c_i$ 的真实距离，即为

$real\_d(x_m,c_i)=d(x_m,c_i)*cor(x_m,c_i)$；

Step5：找到每个数据对象 $x_m$ 到聚类中心 $c_i$ 的最小距离 $\min\_real\_d(x_m,c_i)$，并将数据对象 $x_m$ 归为与 $c_i$ 相同的类中，即为

$$C_i=\{x_m:real\_d(x_m-c_i)<real\_d(x_m-c_j),1\le j\le k\}$$

Step6：计算同一簇中对象的均值，更新聚类中心；

Step7：重复 Step2~Step6，直到聚类中心不再发生改变或达到最大迭代数。

### 3.3.3 实验结果的分析

本文的实验分析过程分主要为三步。第一是对公式 3.12 中 $Q$ 取值的有效性进行实验验证，文中选取不同 $Q$ 值对四组数据集的聚类结果进行比较。第二是对文中选取的六个评价指标进行分析，其中设置最大的运行次数 T=300。实验中的聚类准确率、收敛时间及其平均值均通过 20 次单独实验结果后，取其平均值，结果见表 3.7 至表 3.10。第三是使用相关性加权改进后的算法对于聚类结果的影响。

在初始化聚类中心验证中将本文方法与随机初始化法、最大最小初始化法进行了比较；验证相关性加权对于聚类效果的影响时，选取 iris 数据集对 C-kmeans 算法、传统 K-means 算法、GWO-KM 算法、Rt-kmeans 算法进行实验验证比较。图 3.11 是 iris 数据集经主成分分析（PCA）方法降维处理后的实验结果。

表 3.6 四数据集在不同 Q 值时的准确率（%）

| $Q$ 值 | iris | (Statlog)heart | wine | sonar |
|---|---|---|---|---|
| 1 | 98.00 | 64.07 | 70.22 | 55.77 |
| 2 | 90.00 | 59.63 | 70.22 | 56.25 |
| 3 | 89.33 | 59.26 | 70.22 | 55.56 |
| 4 | 89.33 | 59.26 | 70.22 | 55.56 |
| 5 | 89.33 | 59.26 | 70.22 | 55.56 |

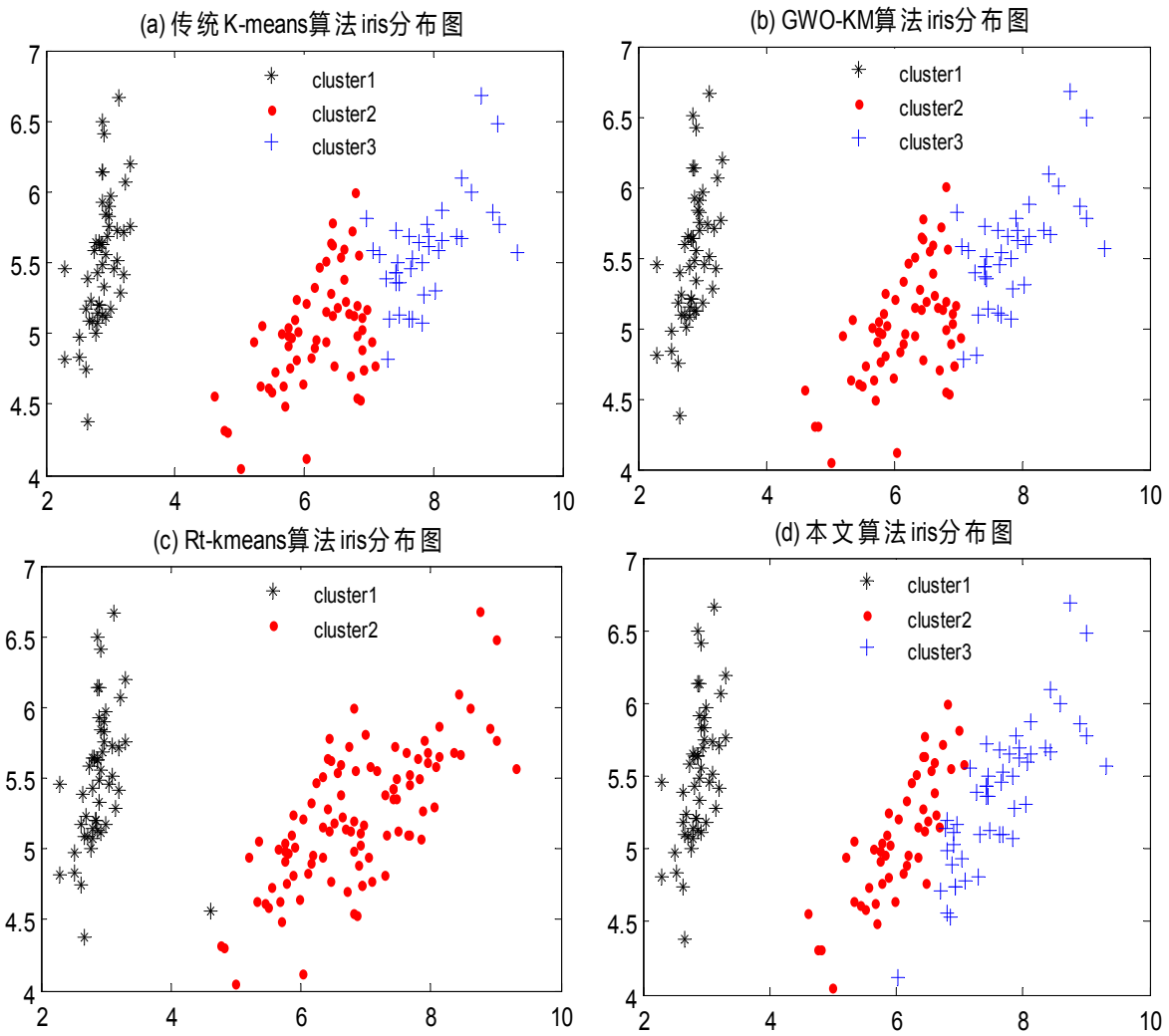



图 3.11 Iris 的聚类效果图

表 3.7 iris 数据集实验结果

| 算法 | 准确率% | 最小最 | 最大值 | 平均值 | 标准差 | 收敛时间/s |
|---|---|---|---|---|---|---|
| 传统 K-means 算法 | 73.93 | 97.32 | 123.96 | 99.91 | 11.83 | 0.49 |
| GWO-KM 算法 | 82.53 | 96.66 | 121.63 | 100.44 | 8.39 | 2.89 |
| Rt-kmeans 算法 | 66.67 | 129.41 | 129.41 | 129.41 | 0 | 0.58 |
| C-kmeans 算法 | 98.00 | 100.46 | 100.46 | 100.46 | 0 | 0.56 |

表 3.8 wine 数据集实验结果

| 算法 | 准确率% | 最小最 | 最大值 | 平均值 | 标准差 | 收敛时间/s |
|---|---|---|---|---|---|---|
| 传统 K-means 算法 | 63.76 | 16555.67 | 18294.85 | 17068.83 | 805.03 | 0.52 |
| GWO-KM 算法 | 71.68 | 16310.55 | 16368.07 | 16329.88 | 14.08 | 3.59 |
| Rt-kmeans 算法 | 65.73 | 23819.11 | 23819.11 | 23819.11 | 0 | 0.86 |
| C-kmeans 算法 | 70.22 | 17924.94 | 17924.94 | 17924.94 | 0 | 1.11 |

表 3.9 (Statlog)heart 数据集实验结果

| 算法 | 准确率% | 最小最 | 最大值 | 平均值 | 标准差 | 收敛时间/s |
|---|---|---|---|---|---|---|
| 传统 K-means 算法 | 52.41 | 10697.12 | 10700.83 | 10698.42 | 1.81 | 0.53 |
| GWO-KM 算法 | 61.11 | 10634.00 | 10676.72 | 10653.59 | 11.74 | 3.09 |
| Rt-kmeans 算法 | 60.74 | 10929.31 | 10929.31 | 10929.31 | 0 | 0.82 |
| C-kmeans 算法 | 64.07 | 10904.69 | 10904.69 | 10904.69 | 0 | 1.251 |

表 3.10 sonar 数据集实验结果

| 算法 | 准确率% | 最小最 | 最大值 | 平均值 | 标准差 | 收敛时间/s |
|---|---|---|---|---|---|---|
| 传统 K-means 算法 | 55.29 | 235.06 | 235.06 | 235.06 | 0 | 0.52 |
| GWO-KM 算法 | 53.46 | 261.99 | 278.48 | 273.97 | 3.94 | 5.07 |
| Rt-kmeans 算法 | 37.98 | 236.53 | 236.53 | 236.53 | 0 | 5.31 |
| C-kmeans 算法 | 55.77 | 235.14 | 235.14 | 235.14 | 0 | 2.86 |

通过表 3.6 结果可知，当 Q 取 1 时，对于实验中所采用的四组数据集，均能获得比较好的聚类结果，但是当随着 Q 取值的增加。文中对于四组数据集的聚类结果没有得到有效的改善，并且对于部分数据集，所得的聚类准确率逐渐降低。其中对于 Iris 数据集而言，当 Q 值逐渐增加时，聚类的准确率非但没有得到提升，反而大幅度的下降，并最终逐渐趋于稳定，由此也证明本文对于 Q 值取值的合理性及其有效性。

由图 3.11 中的四幅子图聚类结果可知，本文所改进后的距离度量方法，其聚类的效果对于 Iris 数据集，相比其他几个算发能够获得更高的准确率。同时通过子图 d 与子图 b、子图 c 对比可以看出，本文所改进以后的算法，对于边界上一些距离较远，但实际相似性较强的对象，通过相关性加权以后能使其聚在同一簇中；而对于一些距离相对较近，但实际差异性较大的对象，通过采用加权以后，能使其归为不同的簇中。并且有表 3.7 至表 3.10 可知，C-kmeans 算法在聚类的稳定性、收敛速度及其准确度等方面相比其它几种算法均能获得更优的聚类结果。

## 3.4 本章小结

针对传统 K-means 算法在聚类求解问题时出现初始聚类中心选取敏感、聚类中心数目难以预先确定以及样本对象间相关性不能得到有效体现的缺点。本章节中分别从初始聚类中心选取、聚类中心数目自适应划分以及对距离采用皮尔相关性系数进行加权三方面对传统 K-means 算法进行了改进。初始聚类中心选取中将最大最小距离初始化方法与样本对象间的密度相结合，使得所选取的初始聚类中心是唯一的并且与最终的聚类中心比较的逼近，从而降低聚类过程中的迭代次数，保证结果的有效性；为了使得聚类中心数目能够自适应划分，本章通过分析最小类间距离及最大类内距离与理想最小类间距离与理想最大类

内距离之间的关系，选用当前划分的平均类间距离来近似理想划分的最小类间距离来实现对聚类中心数目的自适应确定；针对样本间的相关性问题，本文在欧式距离的基础上采用皮尔逊相关性系数进行加权，使得数据样本在聚类的过程中不仅只考虑样本对象之间的欧式距离，同时也能考虑两两样本对象之间的相关性。通过实验分析可知，本文的改进算法相比传统 K-means 算法、GWO-KM 算法、Rt-kmeans 算法具有更高的准确度。

# 第四章 改进算法在股票数据中的应用

伴随着国民经济的迅速发展，金融业在中国市场经济中起着举足轻重的作用，为促进中国市场经济的发展具有着不可磨灭的光辉。证券行业作为金融行业中的一大市场，它与银行体系、保险体系、债券市场等一起构成国内金融基础设施的“组合体”,对经济的建设与发展着至关重要的作用。证券行业已逐渐演变为经济发展中不可或缺的一大经济产业。股票市场作为证券行业中的一个主要组成成分，因该行业的蓬勃发展，越来越多的人投身于股票市场中。投资股票已演变为人们日常生活中的一个主要投资方式，然而近年来，随着金融市场不稳定性的加剧，股市变得更加凶险莫测，因此理性的投资变得尤为重要。投资股票作为投资者的一个主要投资项目，对其进行研究与分析则变得必不可少[48]。数据挖掘对于研究股票波动情况主要体现在两个方面。一个是在股票信息方面的应用；另一个是在股票数据方面的应用[49, 50]。而股价作为公开的、易获取的股票数据信息，因为数据量大且是离散的、数值的、有规律的等原因，因此很适合数据挖掘的应用[51]。股价的研究中除了能直接观察到价格波动以外，还有一个隐藏的，不可直接观测到的信息，即“群成员”。精确的挖掘出股价中所隐藏的信息，将对股票的组合投资产生较大的影响。然而，股票数据是一个数据量庞大，更新快的实时数据，为投资者对股价的分析增添了不少难度。因此，有效的挖掘方法对分析与提取股价中所隐藏的信息变得尤为重要。

## 4.1 股票数据的分析

本文将所改进的算法主要应用在股票数据方面，主要是对股票自身股价进行数据挖掘，股价作为股票数据中的重要数据信息，对于股票未来的波动变化趋势有着重要影响。同时股票价格的波动不仅只与自身变化趋势有关，而且与其他股票也有着重要关系。并且股价之间也有着很强的相关性存在。所以，本课题希望通过使用C-kmeans算法在对股票进行聚类中，使得相关性更强的聚为一类，从而为投资者在投资股票选取组合时提供帮助。

股价作为人们选取股票投资的一个主要参考标准，本文通过采用本文所改进的算法来对股价进行聚类分析，从而帮助股民在选取股票投资是降低其盲目性，分散投资中所存在的风险。本文在对股票数据的分析过程中主要分为两部分：一是对多支股票间的组合分析；二是关于单支股票自身股价波动变化可能性的研究。

对于多支股票间的组合研究，为了研究不同股票间的相关性，本文从地产、金融、信息科技三大行业共计 393 支股票中选取其中 100 支股票 2016 年 2 月 5 日至 2017 年 2 月 17

日的每周交易收盘价格作为数据来源，股票价格数据及所属行业类别从东方财富 Choice 金融终端获取，100 家上市公司股票名称如表 4.1 所示(下文中将直接以编号代表该股票的名称)。考虑到各股票之间的价格差异较大，若直接对各支股票价格的原始数据进行聚类分析，结果将会忽略股价波动的趋势而以数据大小进行分类。这将不能有效的体现出不同股票间所存在的内部关系。因此分析股价变化是需对所研究的数据信息进行预处理，使所有的样本数据处于某一的特定的范围内，然后再使用本文改进后的算法对预处理后的样本数据进行聚类分析，聚类结果如表 4.2 所示。

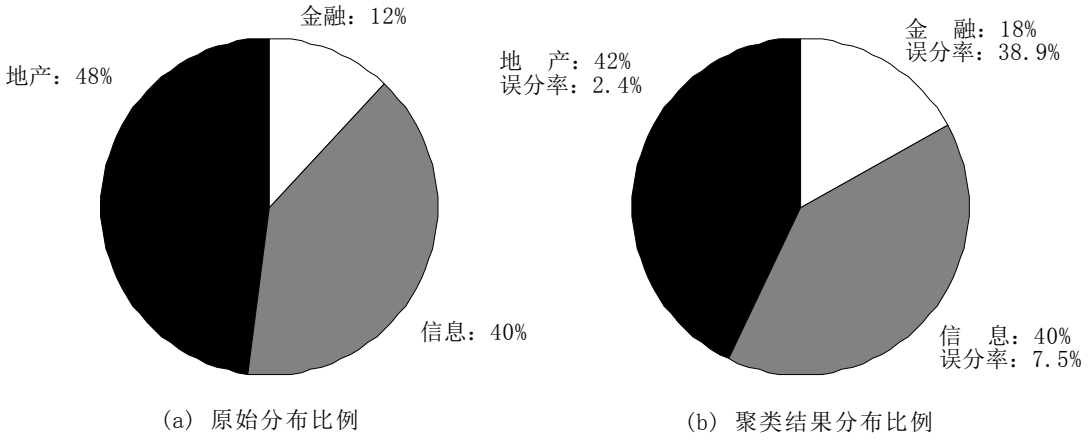
表 4.1　22 家上市公司股票名称

| 股票编号 | 股票名 | 所属行业 | 股票编号 | 股票名 | 所属行业 | 股票编号 | 股票名 | 所属行业 |
|---|---|---|---|---|---|---|---|---|
| 1 | 深振业 A | 地产 | 35 | 光大嘉宝 | 地产 | 69 | 科大讯飞 | 信息 |
| 2 | 深物业 A | 地产 | 36 | 新黄浦 | 地产 | 70 | 启明信息 | 信息 |
| 3 | 中粮地产 | 地产 | 37 | 电子城 | 地产 | 71 | 卫士通 | 信息 |
| 4 | 华联控股 | 地产 | 38 | 陆家嘴 | 地产 | 72 | 中电鑫龙 | 信息 |
| 5 | 中洲控股 | 地产 | 39 | 天地源 | 地产 | 73 | 焦点科技 | 信息 |
| 6 | 广宇发展 | 地产 | 40 | 京投发展 | 地产 | 74 | 天神娱乐 | 信息 |
| 7 | 阳光股份 | 地产 | 41 | 珠江实业 | 地产 | 75 | 太极股份 | 信息 |
| 8 | 金科股份 | 地产 | 42 | 西藏城投 | 地产 | 76 | 中远海科 | 信息 |
| 9 | 荣丰控股 | 地产 | 43 | 京能置业 | 地产 | 77 | 达实智能 | 信息 |
| 10 | 中交地产 | 地产 | 44 | 天业股份 | 地产 | 78 | 凯撒文化 | 信息 |
| 11 | 中国武夷 | 地产 | 45 | *ST 宏盛 | 地产 | 79 | 启明星辰 | 信息 |
| 12 | 嘉凯城 | 地产 | 46 | 上海临港 | 地产 | 80 | 二六三 | 信息 |
| 13 | 福星股份 | 地产 | 47 | 中房股份 | 地产 | 81 | 榕基软件 | 信息 |
| 14 | 天保基建 | 地产 | 48 | 新城控股 | 地产 | 82 | 恺英网络 | 信息 |
| 15 | 世荣兆业 | 地产 | 49 | 平安银行 | 金融 | 83 | 杰赛科技 | 信息 |
| 16 | 广宇集团 | 地产 | 50 | 申万宏源 | 金融 | 84 | 完美世界 | 信息 |
| 17 | 南山控股 | 地产 | 51 | 陕国投 A | 金融 | 85 | 中科金财 | 信息 |
| 18 | 深物业 B | 地产 | 52 | 国海证券 | 金融 | 86 | 真视通 | 信息 |
| 19 | 九鼎投资 | 地产 | 53 | 越秀金控 | 金融 | 87 | 久远银海 | 信息 |
| 20 | 宋都股份 | 地产 | 54 | 浦发银行 | 金融 | 88 | 立思辰 | 信息 |
| 21 | 大龙地产 | 地产 | 55 | 西水股份 | 金融 | 89 | 网宿科技 | 信息 |
| 22 | 卧龙地产 | 地产 | 56 | 中航资本 | 金融 | 90 | 银江股份 | 信息 |
| 23 | 雅戈尔 | 地产 | 57 | 安信信托 | 金融 | 91 | 华星创业 | 信息 |
| 24 | 格力地产 | 地产 | 58 | 东方证券 | 金融 | 92 | 宝通科技 | 信息 |
| 25 | 云南城投 | 地产 | 59 | 南京银行 | 金融 | 93 | 天源迪科 | 信息 |
| 26 | 华业资本 | 地产 | 60 | 北京银行 | 金融 | 94 | 华平股份 | 信息 |
| 27 | 北京城建 | 地产 | 61 | 神州信息 | 信息 | 95 | 数字政通 | 信息 |
| 28 | 天房发展 | 地产 | 62 | 长城动漫 | 信息 | 96 | 银之杰 | 信息 |
| 29 | 华发股份 | 地产 | 63 | 中信国安 | 信息 | 97 | 世纪瑞尔 | 信息 |
| 30 | 粤泰股份 | 地产 | 64 | 电广传媒 | 信息 | 98 | 东方国信 | 信息 |
| 31 | 华丽家族 | 地产 | 65 | 南天信息 | 信息 | 99 | 迪威迅 | 信息 |
| 32 | 黑牡丹 | 地产 | 66 | 远光软件 | 信息 | 100 | 恒生电子 | 信息 |

| | | | | | |
|---|---|---|---|---|---|
| 33 | 海航基础 | 地产 | 67 | 利欧股份 | 信息 |
| 34 | 海航创新 | 地产 | 68 | 麦达数字 | 信息 |

表 4.2 聚类结果

| 所属类别 | 股票编号 |
|---|---|
| 第一类 (地产) | 1、2、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、21、22、23、25、26、27、28、29、30、32、34、35、36、37、39、40、41、43、44、45、46、47、57 |
| 第二类 (信息) | 31、33、42、61、62、63、64、65、66、67、68、69、70、71、72、73、74、75、76、77、78、79、80、81、82、83、84、85、86、87、88、89、90、92、93、94、95、98、99、100 |
| 第三类 (金融) | 3、24、38、48、49、50、51、52、53、54、55、56、58、59、60、91、96、97 |




(a) 原始分布比例　　(b) 聚类结果分布比例

图 4.1 原始分布比例与聚类结果分布比例对比

从表 4.1 与表 4.2 中可知，文中所选取 100 支股票最终被自动划分为了三类，与所选股票行业类别数目相同。在第一类中除了编号为 57 的股票属于信息技术行业，其余股票均属地产类，且由图 4.1 中(b)知，其误分率较低，仅为 2.4%。对第二类而言，其误分率同样较低，仅为 7.5%。在第三类中，其误分率较高，为 38.9%，但通过分析该类中股票可知，文中所选 100 支股票中金融行业股票仅占 12%，该类中所选股票数较少，因此在误分股票数较少的情况下依旧会产生较高的误分率。从股票的三类分布来看，其大体上是服从行业分布的，但同时也存在着自身价格波动的特点，能使得一些不同行业间的股票能聚集在同一类中。股票的价格波动情况除了跟随行业整体波动情况以外，也与公司自身经营状况有很大关系，因此在聚类过程中出现误分是正常的。

为检验聚类结果的有效性，本文使用 t 检验的 p 值以及皮尔逊相关系数日来分析样本间的差异性[52-54]。当 p 值越小时，则对象间的差异性越显著，其中统计学跟进显著性检验

方法所得到的 p 值，一般以 $p<0.05$ 为有统计学差异， $p<0.01$ 为有显著统计学差异，$p<0.001$ 为有极其显著的统计学差异。而当 r 越大时，则证明相关程度越高，其中 $r\le 0.3$ 时，为负线性相关或不存在线性相关， $0.3<r\le 0.5$ 为低度线性相关，$0.5<r\le 0.8$ 为显著线性相关，$0.8<r$ 为高度相关[27]。

表 4.3 至表 4.5 为相同类别中部分 t 检验的 p 值以及皮尔逊相关性系数 r；表 4.6 为 p 值与相关性系数出现最值时的情况。

表 4.3 第一类股票统计量分析

| 股票 | 6 | | 股票 | 6 | | 股票 | 6 | |
|---|---|---|---|---|---|---|---|---|
| 编号 | p 值 | r | 编号 | p 值 | r | 编号 | p 值 | r |
| 1 | 0.2425 | 0.6533 | 16 | 0.2135 | 0.8414 | 32 | 0.6897 | 0.7119 |
| 2 | 0.0656 | 0.7627 | 17 | 0.1228 | 0.7744 | 34 | 0.2251 | 0.3290 |
| 4 | 0.0748 | 0.4933 | 18 | 0.9625 | 0.7161 | 35 | 0.1385 | 0.6752 |
| 5 | 0.3367 | 0.8552 | 19 | 0.3819 | 0.4703 | 36 | 0.6441 | 0.7514 |
| 6 | -- | -- | 20 | 0.1216 | 0.5651 | 37 | 0.3544 | 0.8016 |
| 7 | 0.1680 | 0.8143 | 21 | 0.7629 | 0.7540 | 39 | 0.7668 | 0.3874 |
| 8 | 0.0162 | 0.7664 | 22 | 0.7808 | 0.6699 | 40 | 0.4412 | 0.4767 |
| 9 | 0.4561 | 0.6798 | 23 | 3.95E-05 | 0.4018 | 41 | 0.1142 | 0.6138 |
| 10 | 0.0228 | 0.8707 | 25 | 0.2990 | 0.7952 | 43 | 0.3274 | 0.8689 |
| 11 | 0.8971 | 0.8327 | 26 | 0.4077 | 0.4205 | 44 | 0.0051 | 0.6420 |
| 12 | 0.5967 | 0.8029 | 27 | 0.1593 | 0.7698 | 45 | 0.0002 | 0.8001 |
| 13 | 0.8372 | 0.6640 | 28 | 0.4296 | 0.7925 | 46 | 0.0589 | 0.8343 |
| 14 | 0.3673 | 0.8016 | 29 | 0.9655 | 0.5973 | 47 | 0.4037 | 0.6196 |
| 15 | 0.7792 | 0.5718 | 30 | 0.4464 | 0.8334 | 57 | 0.0486 | 0.8699 |

表 4.4 第二类股票统计量分析

| 股票 | 92 | | 股票 | 92 | | 股票 | 92 | |
|---|---|---|---|---|---|---|---|---|
| 编号 | p 值 | 相关性系数 | 编号 | p 值 | 相关性系数 | 编号 | p 值 | 相关性系数 |
| 31 | 0.0237 | 0.4945 | 72 | 0.2261 | 0.7222 | 86 | 0.0081 | 0.6368 |
| 33 | 2.21E-07 | 0.5009 | 73 | 0.0014 | 0.6692 | 87 | 0.1780 | 0.6273 |
| 42 | 0.7545 | 0.5727 | 74 | 0.4536 | 0.4697 | 88 | 3.59E-06 | 0.7076 |
| 61 | 9.81E-04 | 0.5341 | 75 | 3.84E-09 | 0.6341 | 89 | 0.0675 | 0.4835 |
| 62 | 0.8500 | 0.4815 | 76 | 0.0174 | 0.5921 | 90 | 1.36E-05 | 0.5211 |
| 63 | 3.22E-10 | 0.4803 | 77 | 0.0329 | 0.3298 | 92 | -- | -- |
| 64 | 0.3694 | 0.5943 | 78 | 0.1105 | 0.5911 | 93 | 0.5844 | 0.5441 |
| 65 | 0.0135 | 0.2596 | 79 | 0.0011 | 0.6434 | 94 | 0.0192 | 0.6829 |
| 66 | 0.0020 | 0.6063 | 80 | 0.0139 | 0.7294 | 95 | 0.0093 | 0.7268 |
| 67 | 0.0621 | 0.5805 | 81 | 0.0003 | 0.5402 | 98 | 0.8334 | 0.6093 |
| 68 | 0.4962 | 0.6610 | 82 | 0.0504 | 0.5074 | 99 | 0.5242 | 0.5624 |
| 69 | 2.14E-07 | 0.4811 | 83 | 0.0697 | 0.4514 | 100 | 0.5108 | 0.6237 |
| 70 | 0.0091 | 0.6840 | 84 | 0.1256 | 0.6332 | | | |
| 71 | 1.65E-07 | 0.5891 | 85 | 0.0078 | 0.6370 | | | |

表 4.5 第三类股票统计量分析

| 股票编号 | 50 | | 股票编号 | 50 | | 股票编号 | 50 | |
|---|---|---|---|---|---|---|---|---|
| | p 值 | 相关性系数 | | p 值 | 相关性系数 | | p 值 | 相关性系数 |
| 3 | 0.3534 | 0.5302 | 51 | 0.0229 | 0.3197 | 58 | 0.4993 | 0.7776 |
| 24 | 0.1271 | 0.7584 | 52 | 0.4617 | 0.8172 | 59 | 0.9954 | 0.7446 |
| 38 | 0.4159 | 0.8462 | 53 | 0.6263 | 0.7976 | 60 | 0.0005 | 0.5976 |
| 48 | 0.0755 | -0.0676 | 54 | 0.0171 | 0.7479 | 91 | 0.8441 | 0.8030 |
| 49 | 0.2014 | 0.7602 | 55 | 0.0469 | 0.6818 | 96 | 0.6899 | 0.4941 |
| 50 | -- | -- | 56 | 0.9778 | 0.8956 | 97 | 0.0477 | 0.5101 |

表 4.6 统计量 p 值与 r 值分布比重

| 股票类别 | p 值 | | | 相关系数 r | | | |
|---|---|---|---|---|---|---|---|
| | p<0.05 | p<0.01 | p<0.001 | r<0.3 | 0.3<r<0.5 | 0.5<r<0.8 | 0.8<r |
| 第一类 | 0.1429 | 0.0714 | 0.0476 | 0 | 0.1667 | 0.5 | 0.3333 |
| 第二类 | 0.5500 | 0.4000 | 0.2250 | 0.0250 | 0.2000 | 0.7500 | 0.0250 |
| 第三类 | 0.2778 | 0.0556 | 0.0556 | 0.0555 | 0.1111 | 0.5556 | 0.2778 |

表 4.7 股票间最值情况

| 检验值 | 最大 | | | 最小 | | |
|---|---|---|---|---|---|---|
| | 最大值 | 股票编号 | 是否属于相同类 | 最小值 | 股票编号 | 是否属于相同类 |
| p 值 | 0.9989 | 72、87 | 是 | 2.6E-17 | 23、51 | 否 |
| r 值 | 0.9488 | 2、18 | 是 | -0.8934 | 14、38 | 否 |

由表 4.3 至表 4.5 数据可看出，对于被划分在相同类中股票，绝大部分检验值 p 及其相关性系数 r 都比较的大，并且通过表 4.6 中检验值的概率分布可知，对于分布在第一类与第三类中的股票，其 p 值分布在小于 0.001 上的所占比重分别为 4.7%和 5.5%，并且都相对较小，既对于在相同类别中的股票有极其显著的统计学差异是非常低；并且对于 p 值分布在小于 0.05 上所占比重也不高。其中对于第二类，虽然其中对于 p 值分布在小于 0.05 上的概率稍高，但是分析 p 值分布在小于 0.001 上的比重也依旧不高。同样对于相关性系数的分布情况从表 4.6 可看出，文中所分的三类中，对于相同类中的股票，其相关性系数小于 0.3 的比重几乎可以忽略不计，既分布在相同类中的股票几乎都存在一定程度的相关性；其中不同股票间的相关性系数 r 大部分情况都分布在(0.5,0.8]的范围内，既不同股票之间存在着显著线性相关。而对于第一类与第三类的相关性分布，甚至有超多 25%的股票之间存在着高度相关。并且对于各股票之间检验值得最值情况，从表 4.7 可知，t 检验的 p 值与相关性系数 r 的最小值都出现在不同的类中，并且股票之间也分别属于不同的行业；而最大值都是出现在相同类中，并且也都属于同行业中的股票。p 值与相关性系数所得实验结果同聚类过程中相同类中对象比较相似，而不同类间差异性较大的结果是一致的。

股票的价格波动趋势分析中，主要研究单只股票自身价格可能出现波动情况以及对不同股票间价格波动趋势进行对比。对不同股票间价格波动趋势进行对比分析主要从同类同

行业、同类不同行业、不同类不同行业、不同类同行业四种情况进行研究。本文在表 4.2 所得聚类结果基础上对股票数据进行数据随机选取，所得编号为 3、31、72、75 的股票。

对这四只股票取 2016 年 2 月 5 日至 2017 年 8 月 31 日的每日的收盘价进行研究。因股票价格为时间系列，为提高准确度，采用移动选取样本的方式[14]，将该期间内的交易数据每隔 10 天平移选取 20 天的数据作为一个样本，即第 i 个样本中数据为 $x^i = \{x_j^i \mid 10\times i-9 \le j \le 10\times i+10\}$。因此，每只股票都各获得一个拥有 39 个样本 20 维的数据集。分别对所得数据集进行聚类，并对聚类结果的每类求均值得到各类中样本的平均价格波动趋势，结果如图 4.2 至图 4.5 所示。

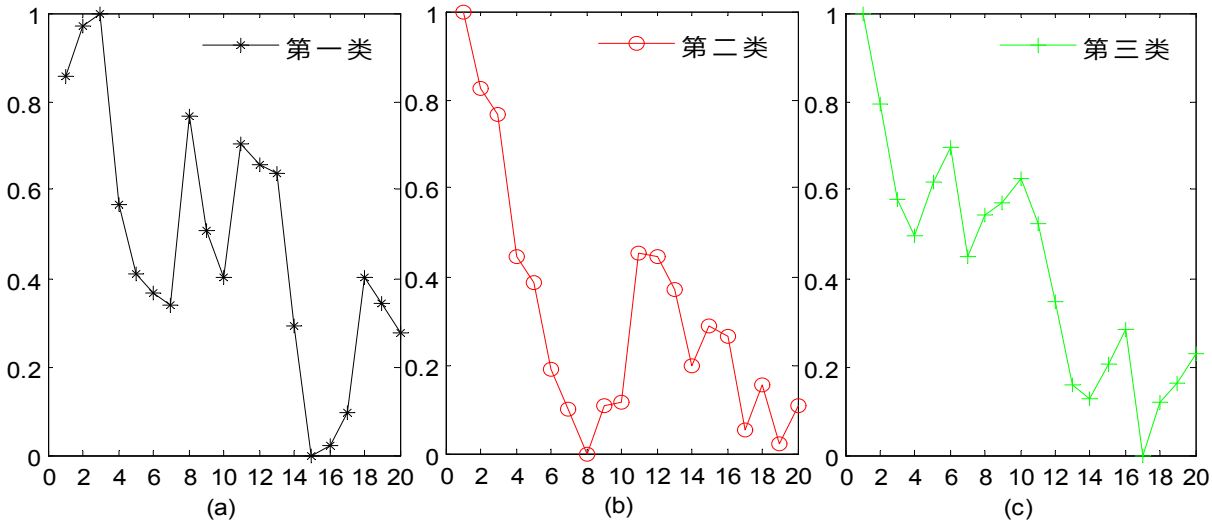



图 4.2　股票 3

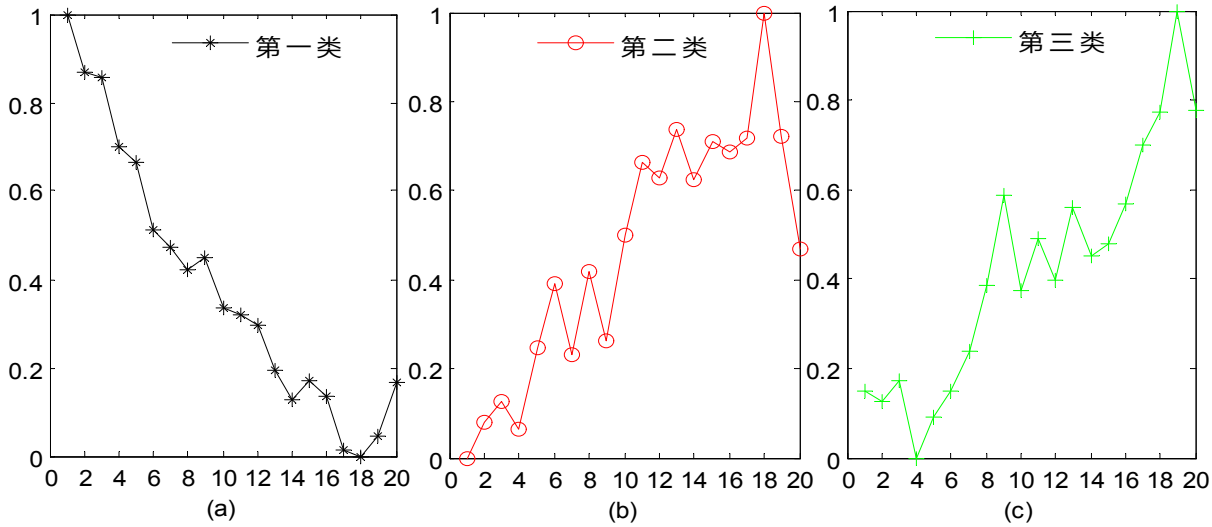



图 4.3　股票 31

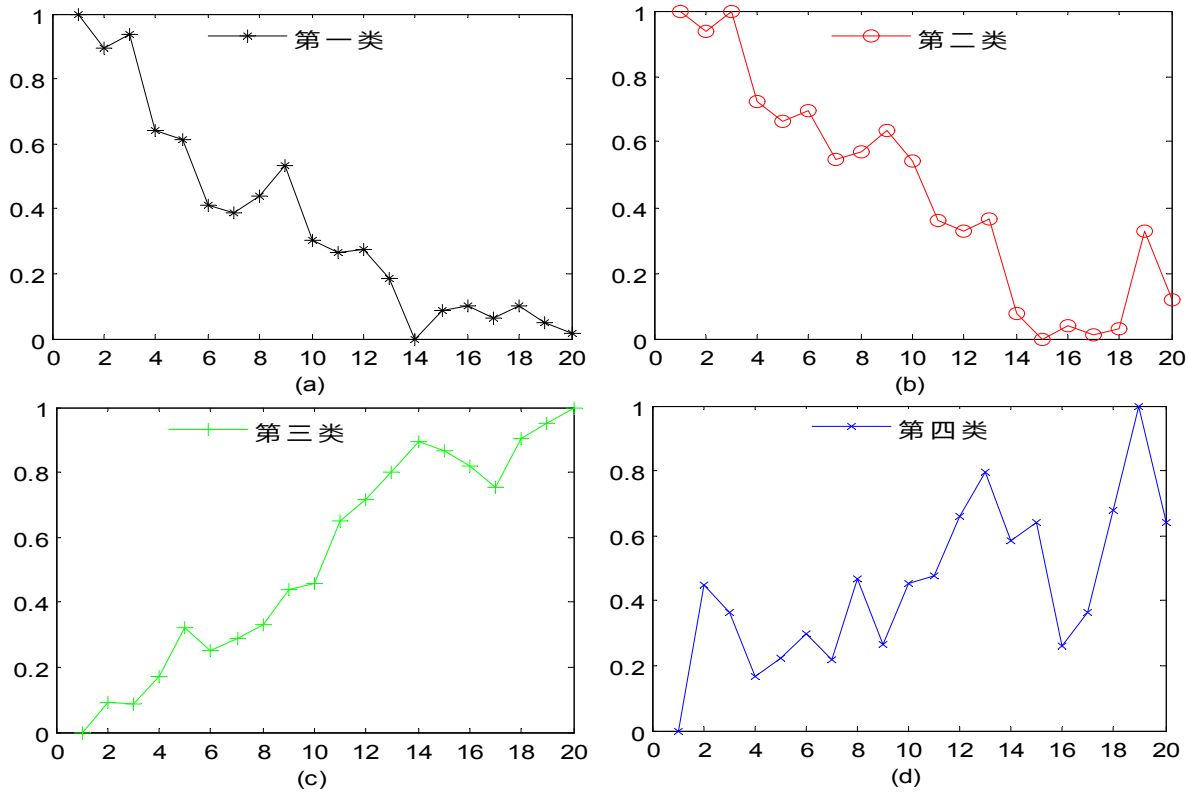



图 4.4 股票 72

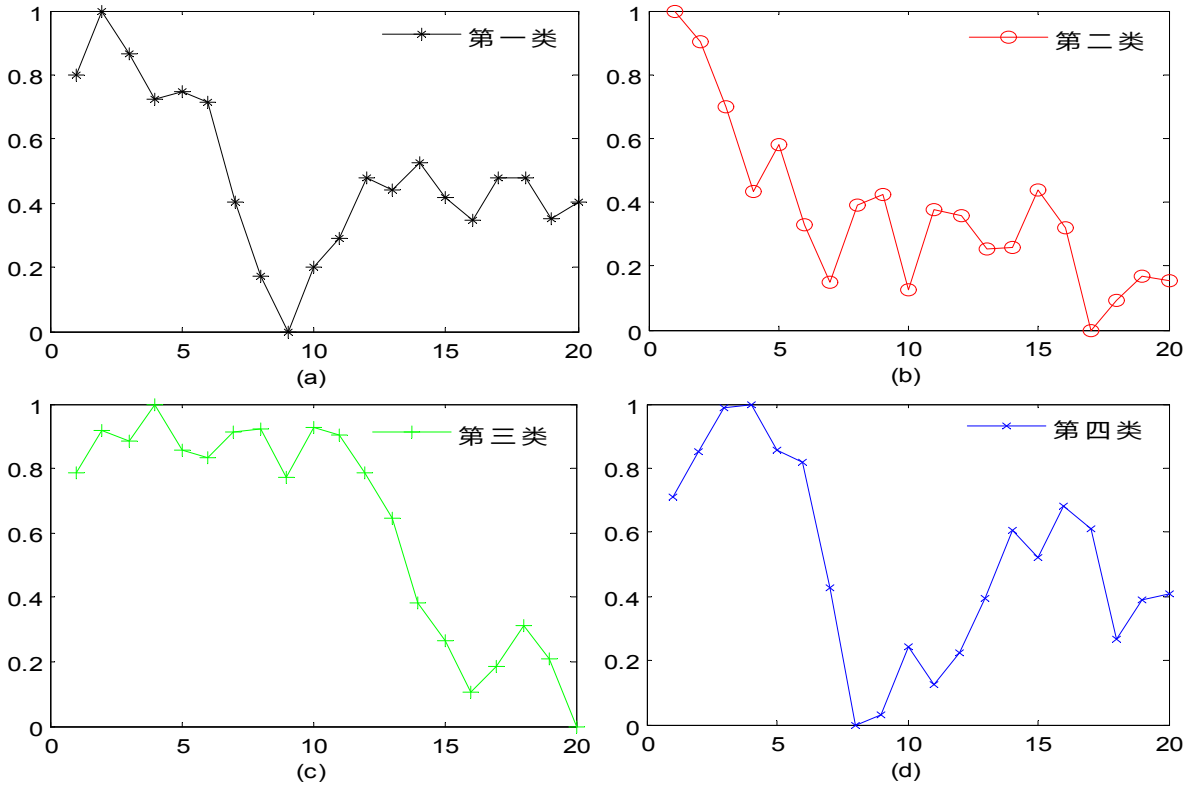



图 4.5 股票 75

针对同类不同行业的股票进行分析，本文主要以股票 31 与股票 72 为例进行对比，其中编号 31 属于的股票地产行业类，编号 72 的股票属于信息行业类，从表 4.2 中的聚类结果可知它们均被划分在第二类中。通过对 4.3 与图 4.4 对比可看出，图 4.3 中(a)与图 4.4 中(a)、(b)的波动趋势比较的相似，并且图 4.3 中(b)、(c)与图 4.4 中(c)、(d)的波动也是较为接近的。

对于不同类不同行业股票的分析，本文主要通过股票 3 与股票 72 进行对比分析，其中股票 3 属于地产行业、被划分在第三类中，股票 72 属于信息行业，划分在第二类中。通过对图 4.2 与图 4.4 对比显示，股票 3 被自动的划分为了三类，股票 72 被划分为了四类，其中股票 3 波动振幅比较大，而股票 72 波动则较为平滑，并且它们在价格波动趋势上差异较大。

不同类同行业股票分析中以股票 3 与股票 31 对比为例，该股票虽均属于地产行业，但在上文聚类过程中分股票 3 被划分在第三类，而股票 31 被划分为第二类，通过图 4.3 与图 4.4 对比可知，股票 3 与股票 31 的波动趋势虽都被自动划分为三个类别，但是其波动趋势存在较大差异。将其划分在不同的类中是较为合理的。

同类同行业的股票数据分析中以股票 72 与股票 75 为例，该股票均属于信息行业且都被划分在第二类中。通过图 4.4 与图 4.5 对比可知，股票 72 与股票 75 的股价波动均被划分为四类。且观察两幅图可看出，除了子图(c)以外，其余子图(a)、(b)、(d)的波动变化情况都比较的相似。

通过上述实验表明，文中对部分同类同行业、同类不同行业、不同类不同行业、不同类同行业股票间的分析所得结果与表 4.2 中所展示聚类结果较为一致。改进算法不仅能将股票大体按行业进行自动划分，同时能满足不同行业中对于价格波动关联性较强的股票，也能使其划分在相同的类中。而对于价格波动差异性较大的股票，能使其分属在不同类中。并且通过对统计值的检验以及股价波动情况的分析文中所得聚类结果是有效的。

## 4.2 结论

股市作为证券与金融行业的中要组成部分，高风险与高收益作为其主要的特性，一直被投资者所关注。本文通过对 K-means 算法的改进，并将改进后的算法应用到对股价的分析中。改进算法能自动的确定划分的聚类数目，使用该方法在对股价的挖掘中，可以得出以下结论。通过将股票短期的价格波动情况与处于相同类中的股票整体波动情况进行对照，从而分析出股票短期内的上涨与下降情况，预测出股票未来短期的变化趋势。如果分析出某一支股票可能上涨，那么对于处于相同类中彼此间的 p 值与相关系数都较大的股

票，则可以继续持有，如果下跌，则同类中的股票应尽快减仓。并且在投资过程中，应使所投资的股票应尽量的分散在不同的类别中，从而可以有效的降低投资的风险。

# 第五章 总结与展望

## 5.1 总结

随着经济社会的迅速发展，数字信息时代的到来，每天大量的数据产生与更新，庞大的数据量为研究人类挖掘出数据中的信息增添了不少的难度。快速、高效的挖掘出数据中所隐藏的信息变得尤为重要。其中 K-means 算法作为十大经典数据挖掘算法之一，因其简单、高效的特点备受许多研究人员的青睐。本文通过研究 K-means 算法，针对 K-means 算法在对数据进行挖掘分析时所存在的各个不足之处进行了改进。本文主要研究工作和创新点如下。

(1) 通过分析传统K-means算法在对数据集初始化的过程中因初始点随机选取的不足，容易出现聚类结果不稳定，且坏初始点的产生易增加算法的迭代次数，从而提出了使用最大最小初始化法与密度相结合的初始化方法。

(2) 分析了传统K-means算法预先确定聚类中心数目的问题，为了使得K-means算法能自动确定聚类中心数目，本文通过研究在聚类过程中，分析类内距离与类间距离之间的关系，从而来确定划分的阈值。

(3) 传统K-means算法主要是通过欧式距离来对样本数据集之间的关系进行划分，欧式距离虽作为最为常用的划分标准，但是欧式距离对于体现对象间的相关性是比较差的，针对此问题，本文采用Pearson相关系数来对距离进行加权，从而获得一种既能体现对象间相关性，又能体现相似性的划分方法。

(4) 股票是一种离散、数据量很大的实时数据，而股票之间又有着较大的相关性，将本文研究的方法用于对股票的股票数据分析可以得到一些有价值的结果。

## 5.2 展望

本文针对传统K-means算法中所存在的不足做了一些研究，提出了一种改进的K-means算法。并通过对算法原理进行理论分析和实验数据验证了改进算法是有效、稳定的，但是一些内容的研究还存在着不足，需要进一步的探索。未来的工作开展主要包含以下三点内容：

(1) 本文所改进的方法主要应用于数值型数据，而对于文本类型数据的划分，本文方法将得不到有效的解决。扩展算法的通用性是今后工作的努力方向。

(2) 对于文中的初始簇中心点的改进，通过采用改进的初始簇中心能够得到高质

量的初始解集。但该改进的方法依旧不能确保绝对获得全局最优解，在未来的工作当中，寻找全局最优解的能力这方面有待进一步提升。

(3) 本文所选用的股票数据样本不够完善，无法包含股票中的所有数据类型，并且在对股票的聚类过程中并没有很好的考虑数据时间特性对聚类结果的影响，即股票数据可以划分为短时和长时数据模式，这两种模式的聚类所得结果可能不相同，未来需要在这方面进行完善。

# 参 考 文 献


[1] 胡文瑜,孙志挥,吴英杰.数据挖掘取样方法研究[J].计算机研究与发展,2011,48(1):45-54.

[2] 任新社,陈静远.关于数据挖掘研究现状及发展趋势的探究[J].信息通信,2016,2:171-172.

[3] Ali Al-Wakeela,Jianzhong,Wu.K-means based cluster analysis of residential smart meter measurements[J]. Energy Procedia,2016,80:754-760.

[4] 李涛.数据挖掘的应用与实践[M].厦门大学出版社,2013,9:3.

[5] 邓宏勇,许吉,张洋.中医药数据挖掘研究现状分析[J],中国中医药信息杂志,2012,19(10):21-23.

[6] 贺瑶,王文庆,薛飞.基于云计算的海量数据挖掘研究[J].计算机技术与发展,2013,23(2):69-72.

[7] 王元卓,贾岩涛,刘大伟.基于开放网络知识的信息检索与数据挖掘[J].计算机研究与发展,2015,52(2):456-474.

[8] 张宇,朱凝秀.数据挖掘中的模糊聚类分析[J].工业设计,2012,3: 1672-7053.

[9] 李翠,冯冬青.基于改进 K-均值聚类的图像分割算法研究[J].郑州大学学报,2011(1):103-113.

[10] 贾瑞玉,李玉功.类簇数目和初始中心点自确定的 K-means 算法[J].计算机工程与应用,2017,03(22):1923-1932.

[11] 车明菊,卢志刚.基于 k-中心聚类与布谷鸟搜索的伙伴选择[J].计算机工程与设计,2017,38(12):3413-3418.

[12] 何童.不确定性目标的 CLARANS 聚类算法[J].计算机工程,2012,38(11):56-58.

[13] 张虎,陈建斌,魏欢.一种改进的 BRICH 算法及其应用[J].软件导刊,2015,14(10):45-46.

[14] 高长元,王海晶.王京.基于改进 CURE 算法的不确定性移动用户数据聚类[J].计算机工程与科学,2016,38(4):678-773.

[15] 杨芳勋.DBSCAN 算法在电子邮件网络社团发现中的应用[J].计算机科学,2017,44(6A):591-593.

[16] 朱亮,李东波,何非.采用改进型 DENCLUE 和 SVM 的电子皮带秤故障诊断[J].哈尔滨工业大学学报,2015,47(7):122-128.

[17] 王红葛,丽娜.基于 OPTICS 聚类的差分隐私保护算法的改进[J].计算机应用,2018,38(1):1-8.

[18] Libao ZHANG, Faming LU.Application of K-Means Clustering Algorithm for Classification of NBA Guards[J].International Journal of Science and Engineering Applications,2016,5(1):1-41.

[19] 任新社,陈静远.关于数据挖掘研究现状及发展趋势的探究[J].信息通信,2016,2:171-172.

[20] 王梦雪,数据挖掘综述[J].软件导刊,2013,10:135-137.

[21] 王惠中,彭安群.数据挖掘研究现状及发展趋势[J],工矿自化,2011,2:29-32.

[22] 陈建伟,李丽坤.数据挖掘技术研究[J],数字技术与应用,2016,1:91-92.

[23] Xindong Wu,Vipin Kumar.数据挖掘十大算法[M].清华大学出版社,2013,19.

[24] 刘越.K-means 聚类算法的改进[D].广西师范大学,2016,8.

[25] 刘金岭.K 中心点聚类算法在层次数据的应用[J].计算机工程与设计,2008,29(24):6418-6419.

[26] 张雪凤,张桂珍,刘鹏.基于聚类准则函数的改进 K-means 算法[J].计算机工程与应用,2011,47(11): 123-127.

[27] 张良均,杨坦,肖刚.MATLAB 数据分析与挖掘实战[M].机械工业出版社,2016,35-46.

[28] Liang Bai,Jiye Liang,Chao Sui.Fast global K-means clustering based on local geometrical information[J]. Information Sciences,2013,245:168-180.
[29] 张晓慧,孙连山.基于改进 K-中心点的电子地图数据质量检查算法[J].软件导刊,2017,16(2):81-84.
[30] Preeti Arora,Dr.Deepali,Shipra Varshney.Analysis of K-means and K-Medoids Alogrithm For Big Data[J].Procedia Computer Science,2016,78:507-512.
[31] 黄韬,刘胜辉,谭艳娜.基于 K-means 聚类算法的研究[J].计算机技术与发展,2011,21(7):54-54.
[32] Kohei Arai,Ali RidhoBarakbah.Hirrarchical K-means:an algorithm for centroids initialization for K-means[J]. Rep.Fac.Sci.Engrg.Saga Univ,2007,36(1):25-31.
[33] 杨红光,刘建生.一种结合灰狼优化和 K-均值的混合聚类算法[J].江西理工大学学报,2015,36(5):85-89.
[34] 周世兵,徐振原,唐旭清.K-means 算法最佳聚类数确定方法[J].计算机应用,2010,30(8):1995-1998.
[35] 王勇,唐靖,饶勤菲.高效的 K-means 最佳聚类数确定算法[J].计算机应用.2014,34(5):1331-05.
[36] Jing Lei,Teng Jiang,Kui Wu.Robust K-means algorithm with automatically splitting and merging cluters and its applications for surveillance data[J].Multimed Tools Appl,2016,75:12043-12059.
[37] Reda M.Elbasiony,Elsayed A.Sallam,Tare E.Eltobely.A hybrid network intrusion detection framework based on random forests and weighted K-means[J].Ain Shams Engineering Journal.2013,4:753-762.
[38] Reda M.Elbasiony,Elsayed A.Sallam,Tare E.Eltobely.A hybrid network intrusion detection framework based on random forests and weighted K-means[J].Ain Shams Engineering Journal.2013,4:753-762.
[39] Muhammed Maruf Öztürk,Unal Cavusoglu,Ahmet Zengin. A novel defect prediction method for web pages using K-means++[J].Expert Systems with Applications, 2015,42:6496-6506.
[40] J.James Manoharan,S.Hari Ganesh.Initialization of optimized K-means centroids using divide-and-conquer method[J].2016,11(2):1076-1081.
[41] 岑晓雪,秦江涛.改进 K-means 聚类在股价波动趋势上的应用[J].科技和产业,2016,16(1):144-148.
[42] 李振,贾瑞玉.一种改进的 K-means 蚁群聚类算法[J].计算机技术与发展,2015,25(12):28-31
[43] 洪月华.蜂群 k-menas 聚类算法改进研究[J].科技通报,2016,32(4):170-173.
[44] 刘凤龙,陈曦,曹敦.基于查分演化的 K-均值聚类算法[J].计算机技术与自动化,2010,29(1):48-50.
[45] 易倩,腾少化,张巍.基于马氏距离的 K 均值聚类算法的入侵检测[J].江西师范大学学报,2012,36(5):284-287.
[46] P. S. Bradley,Usama M. Fayyad.Refining Initial Points for K-means Clustering.Apears in Proceedings of the 15th International Conference on Machine Learning,1998: 91-98
[47] Xiaoyan Wang,Yanping Bai.The global Minmax K-means algorithm[J]. SpringerPlus,2016,5:3329-3333.
[48] 文凤华,肖金利,黄创霞.投资者情绪特这对股票价格行为的影响研究[J].管理科学学报,2014,3(17):60-69.
[49] 张桐,邢东旭,数据挖掘在股票方面的应用[J].数字技术与应用.2013,01:73-73.
[50] 王冬秀,胡迎春,李辉,改进的 Apriori 算法在股票分析中的应用研究[J].科技通报,2013,29(3):125-128.
[51] 玄海燕,孙艳,黄性芳.基于双 AR(p)模型的股价分析及其实证研究[J].数学的实践与认识,2017,13(47).98-104.
[52] 智冬晓,许晓娟,张皓博.z 检验与 t 检验方法的比较[J].统计与决策,2014,20:31-34.
[53] 谢娟英,高红超.基于统计相关性与 K-means 的区分基因子集选择算法[J].软件学报,2014,25(9):2050-2075.

[54] 朱卫东,杜承勇,吴勇.一种基于相关系数矩阵的 TOPSIS 决策方法[J].数学的实践与认识,2014,4(44):33-38.

# 致 谢

本文的研究工作是在导师刘建生副教授的虚心指导下完成的，深深地感谢导师几年来为我付出的巨大心血。刘建生副教授严谨踏实的治学态度，勤奋求实的工作精神，平易近人的品德深深感染着我，使我受益良多，终生难忘，感谢刘老师在我读研期间热情的关怀和严格、耐心的指导，不仅在学业上帮助了我，对我今后的人生也产生了积极而深刻的启迪和影响。

感谢实验室老师以及所有关心我的同学和朋友，从他们身上得到的无私帮助和学习到的宝贵经验让我受益匪浅。

感谢我的家人，感谢他们多年来对我的关心，当我遇到困难时对我的支持和鼓励。

最后，再次衷心感谢在百忙之中评阅论文和参加答辩的各位专家、教授。

# 攻读学位期间的研究成果

## 已发表论文：

[1] 尹宝勇,吴斌,刘建生.一种改进的 K-means 算法[J].江西理工大学学报,2018,05.(已录用)
[2] 刘建生,吴斌,章择煜.基于相关性加权的 K-means 算法[J].江西理工大学学报,2018,01(39),87-92.
[3] 章择煜,刘建生,吴斌.基于嵌入式智能电网电能计量系统的设计与研究[J].工程技术,2017(5):164 -165.