**湖 南 信 息 学 院**

**数据挖掘课程论文**

|  |  |
| --- | --- |
| **论文题目** | **数据挖掘技术在股价分析中的应用** |
| **学生姓名** | **张金平** |
| **学号** | **201501420309** |
| **二级学院** | **电子信息学院** |
| **专业名称** | **软件工程** |
| **班级** | **15软件工程3班** |
| **指导教师** | **张钰莎** |

二〇一八 年 十 月 十三 日

目录

[摘 要](#_Toc16068_WPSOffice_Level1) [3](#_Toc16068_WPSOffice_Level1)

[ABSTRACT](#_Toc11051_WPSOffice_Level2) [3](#_Toc11051_WPSOffice_Level2)

[第一章 前 言](#_Toc11051_WPSOffice_Level1) [3](#_Toc11051_WPSOffice_Level1)

[第二章 数据挖掘技术与股市知识概述](#_Toc20029_WPSOffice_Level1) [4](#_Toc20029_WPSOffice_Level1)

[2.1数据挖掘概述](#_Toc20029_WPSOffice_Level2) [4](#_Toc20029_WPSOffice_Level2)

[（1）问题定义](#_Toc11051_WPSOffice_Level3) [4](#_Toc11051_WPSOffice_Level3)

[（2）数据的收集及其数据的预处理](#_Toc20029_WPSOffice_Level3) [4](#_Toc20029_WPSOffice_Level3)

[（3）数据挖掘](#_Toc29602_WPSOffice_Level3) [4](#_Toc29602_WPSOffice_Level3)

[2.3 股票分析概述](#_Toc29602_WPSOffice_Level2) [4](#_Toc29602_WPSOffice_Level2)

[第三章 数据挖掘算法在股票中的应用](#_Toc29602_WPSOffice_Level1) [5](#_Toc29602_WPSOffice_Level1)

[3.1决策树算法核心思想](#_Toc18322_WPSOffice_Level2) [5](#_Toc18322_WPSOffice_Level2)

[3.2决策树算法的优点](#_Toc28569_WPSOffice_Level2) [5](#_Toc28569_WPSOffice_Level2)

[3.3改进Apriori算法在股票板块分析中的应用](#_Toc31604_WPSOffice_Level2) [5](#_Toc31604_WPSOffice_Level2)

[3.4关联规则在股票板块指数中应用示例](#_Toc15808_WPSOffice_Level2) [6](#_Toc15808_WPSOffice_Level2)

[第四章 结论](#_Toc18322_WPSOffice_Level1) [7](#_Toc18322_WPSOffice_Level1)

[参考文献](#_Toc28569_WPSOffice_Level1) [7](#_Toc28569_WPSOffice_Level1)

# 摘 要

股票作为一种最普遍的风险投资理财产品，正在悄悄地进入人们的日常生活。股票是一种最具代表性的有价证券，股票数据分析及预测是当今世界的热柄课题。股票数据分析往往是依赖于对大量历史的交易数进行预测，如何应用好这个与日俱増的庞大的历史数据，从中挖掘出在价值的规律，是每位投资者比较感兴趣的话题，也是股票预测面临的最大难题。目前比较成熟的基本面分析法和技术面分析法对变化多端股市，经常还是无能为力。

本文从股市的实际特点出发，应用已经成熟的数据挖掘技术，把上市公司的主要财务数据指标作为属性，应用ID3决策树算法进行分类，并针对经典Apriori算法存在的一些缺陷提出了改进改进后的Apriori算法更加灵活，更适合挖掘不同板块之间的关系。

**关键词**：股票；数据挖掘；ID3算法；Apriori算法

**ABSTRACT**

As one of the most common venture capital products, stocks are quietly entering people's daily life. Stock is one of the most representative securities. Stock data analysis and prediction is a hot topic in the world today. Stock data analysis is often dependent on a large number of historical transactions to predict, how to apply this ever-increasing huge historical data, mining out the law of value, is a topic of interest to each investor, but also the biggest problem facing stock forecasting. At present, the more mature basic analysis method and technical analysis method are often powerless to the volatile stock market.

Starting from the actual characteristics of the stock market, this paper applies the mature data mining technology to classify the main financial data indicators of listed companies by using ID3 decision tree algorithm, and proposes an improved Apriori algorithm to overcome the shortcomings of the classical Apriori algorithm. The method is more flexible and more suitable for mining the relationship between different plates.

**Keywords:** Stock; Data Mining; ID3 Algorithm;Apriori Algorithm

# 

# 第一章 前 言

随着我国经济体制改革的不断深化，资产证券化的步伐不断的加快，股票市场在资本市场中的重要性日益的突出，社会投资的模式也日益的増多，例如基金投资等等，人们的投资理念日益增强。从证券业的开端，人们就没有停止过对股票市场的探究，而探巧的最终目的都是在降低风险的同时保证可观的收益。股票数据的分析预测具有很大的现实意义，然而，股票市场是一个极其复杂的系统，它贾穿着经济学、金融学、也理学、投资学等众多学科领域，投资者的盲目性、政局的不稳定性、经济的周期性、银行利率的浮动性等诸多因素决定了股价走势预测的复杂性，但是，归根结底，金融交易过程有一定的规律及联系。

# 第二章 数据挖掘技术与股市知识概述

我国的股票市场起步较晚，但是发展迅速，电子信息化程度与国际接轨，积累了大量的金敲数据。目前，我国股票市场发展良好，在兰十多年的探索中不断地成熟起来，不断地改革，不断地创新。为了解决高度发达的金融市场和滞后的数据处理能力，有必要把数据挖掘技术引入到证券市场上来，把证券市场做大做强，使金融结构更加合理，市场更加智能化。

## 2.1数据挖掘概述

（1）问题定义

数据挖掘的最终目的是为了从所要挖掘的领域数据中挖掘出有用的不为人们直觉所发现的知识。在问题定义阶段中，我们往往要先知道什么样的数据可以用来我们的挖掘工程中来。在实际应用中，问题的详细定义是要整个领域的相关人都要参与，使问题更加明朗，要解决的问题什么样的总量，要怎么样解决，要做到屯、中有数，对需要解决的问题进行多次确认，进而在后续的数据挖掘过程中对数据进行进一步的预选和处理。

（2）数据的收集及其数据的预处理

按照定义阶段中的要求，当采集到数据后，下一步必须要做的事情是对数据进行预处理，尽量消除数据中存在的错误及缺失信息ｕｗ，有时候还要将数据转换为数据挖掘工具所需的格式，数据变换是为了数据挖掘的具体操作而对数据进行的特殊处理，通过变换去除数据中的干扰信息，这可Ｗ使得结果更加理想化。

（3）数据挖掘

依据在问题定义阶段提出的挖掘要求，明确挖掘任务，找到适合的算法来对数据进行挖掘。送么阶段的任务是选择和应用不同的模型及工具，将模型参数调整到最佳的数值。一般而言，有些技术可Ｗ解决一类相同的数据挖掘问题，有些技术在数据形成上有特殊要求，因此需要经常跳回到数据准备阶段。

## 2.3 股票分析概述

传统的股票分析方法，主要是基本面分析和技术面分析。基础面分析又叫基本面分析，所谓基本面，就是指证券的内在价值＂。基本面分析是指对影响股票市场走势的一些影响因素状况进行基本分析，这样可把握决定股价变动的基本因素，是证券投资分析的基础。影响公司股价的因素很多，比如宏观经济状况、利率水平、通化膨胀、企业综合素质等。从长期和根本上看，股票市场的走势变化是由该股票的公司基本面决定的，股价的波动很大程度上是股票基本面的状况的变化的反应。技术分析主要依据是各种图形，来判断价格的走势。核也的是价格、成交量、时间、价格的波动幅度这四个要素。最基本的就是要懂得先学习怎么利用技术分析去投资，还有就是在投资的时候，要同时考虑很多方方面面的问题，投资股票就是投资好公司。不是每个人都能成为己菲特，但是每个人都有兴趣学习如何在股市中赚到钱。中国股民中，散户偏多，散户在股市上失败的多成功的少。

# 数据挖掘算法在股票中的应用

## 3.1决策树算法核心思想

算法是ID3决策树的核也算法，为了便于理解，在此简单介绍一下算法的

大体流程如下：

步骤１：初始化决策树，对数据进行规范化处理：

步骤２：抽样选取实例集作为训练集，统计并记录数据；

步骤３：应用上个步骤中统计到的数据，计算各个属性的信息增益度，选取最能分类训练集的属性：

步骤４：把上个步骤中得到的那个最能分类训练集的属性作为根结点，创

建一棵树，并根据该属性的值进行分类，创建该结点的边，递归以上步骤，当

出现一个结点的数据都是属于同一类或者属性不能再分割，停止分类算法：

步骤5：任意抽取实例来对所生成的决策树进行验证，并记录决策树的准确度。

## 3.2决策树算法的优点

决策树算法是一种贪也算法，得到的结果不一定是最优解。决策树的每个节点可以有多个子节点，也可以没有节点，没有节点的为叶子节点，其他的有节点的称为内部节点。叶节点是代表类或类分布，每个内部节点表示在一个分割属性上的测试，每个分枝代表一个输出。决策树分类法的优点很多，主要包括以下点：

第一：计算的速度快，计算量比较小，容易得到分类规则。只要沿着树根向下一直走到树叶，就能唯一确定一棵决策树。

第二：准确性相对来说还是比较髙的，挖掘出的分类规则便于理解，直观视觉上也比较便于分析。

第三：决策树的分类虽然不是最优解，但是足以应付一般的决策问题，能够指导现实生活中很多问题的决策。

## 3.3改进Apriori算法在股票板块分析中的应用

＂股票板块＂指的是因为有某一共同特征而被人为地归类到一起的一类股票群体，这种特征的题材股，所谓股市庄家及主力经常用来炒作＂。板块有按地理划分的，例如＂福建自贸区板块珠江三角洲板块＂；有的是业绩上的，如＂绩优板块＂；有的是上市公司重组方面的，如＂资产重组板块＂；还有的是所属行业方面的，如＂证券板块＂、＂银行板块＂、＂互联网＋板块＂、＂工业板块＂等等。总之，只要是新兴的题材名称可用来炒作的，都可適之板块的名称。通常情况下，股票市场具有连动效应，如果股票市场大盘走势强势，每个股票个体上攻走强的可能性就较大；如果股票市场大盘走势疲软，则每个股票疲软的概率也大。所以，对于广大股民来说，不仅要看清个股的价格变化趋势，更要了解大盘指数的变化情况及变化庭势。股价指数是衡量股票价格水平及其变动趋势的尺度，也是反映板块发展走势状况的有效信号。投资者可根据此检验自己投资的效果，预测股票市场的动向。当某个板块有明显的一两只股票领先大盘涨跌时，同一板块的其他股票一般随后也会跟随涨跌，这就是所谓的＂板块联动＂效应。股市中有一些股票会共同具备某种共同的特殊性质，这种共同性质被市场发现，就会形成股市中的板块结构。

## 3.4关联规则在股票板块指数中应用示例

这里从同花顺炒股软件中提取出猪肉、云计算、物联网、大数据、二胎概念、环保工程六大板块从2016年02月29日至2016年03月04日送5个交易日的板块收盘指数，如下表所示：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 日期 | 猪肉 | 云计算 | 物联网 | 大数据 | 二胎概念 | 环保工程 |
| 2016-02-09 | 1923.9 | 2339.2 | 1561.6 | 2897.8 | 1678.8 | 3982.3 |
| 2016-03-01 | 1958.3 | 2388.6 | 1606.2 | 2963.2 | 1726.3 | 4065.1 |
| 2016-03-02 | 2029.1 | 2518.4 | 1695.9 | 3117.6 | 1820.1 | 4288.0 |
| 2016-03-03 | 2013.2 | 2526.4 | 1689.3 | 3123.6 | 1847.6 | 4307.0 |
| 2016-03-04 | 2023.4 | 2376.2 | 1611.9 | 2950.5 | 1767.3 | 4069.3 |
| 2016-03-07 | 2075.5 | 2454.2 | 1665.4 | 3039.7 | 1816.1 | 4185.8 |
| 2016-03-08 | 2028.2 | 2515.7 | 1699.0 | 3100.9 | 1831.9 | 4220.5 |

设Xi为第（N-1）个交易日的股价指数，也为第Ｎ个交易日的股价指数。则第N个交易日的板块指数涨跌为。分别以字母A、B、C、D、E、F表示以上六个板块。这里我们做如下规定：如果在大盘指数涨的情况下，板块涨幅不小于大盘，则这些板块归属同一个项目集；反之，如果在大盘指数跌的情况下，板块跌幅不小于大盘，则这些板块巧属同一个项目集。则可整理得到下表事务数据。令最小支持度为50%，最小置信度为90%，求事务数据库的频繁关联规则。假设投资者只对B、C、D三者的内在关系感兴趣。

|  |  |
| --- | --- |
| TID | 项目集 |
| 1 | B、C、D、E、F |
| 2 | B、C、D、F |
| 3 | A、B、C、D |
| 4 | B、C、D、E |
| 5 | A、B、C、D、F |
| 6 | B、C、D |

算法执行过程如下：

第一步：求频繁项集：

频繁1-项集{{B},{C},{D},{F}}。

频繁2-项集{{BC},{BD},{CD},{CF},{DF}}。

频繁3-项集{{BCD}}.

综上一共有{{B},{C},{D},{F},{BC},{BD},{CD},{CF},{DF},{BCD}}

第二步：求投资者感兴趣的{BCD}关联关系，{BCD}支持度为100%。

# 结论

随着经济的发展和人们投资意识的转变，资产证券化将是未来投资的趋势，现如今股票投资己成为大部分人投资的一种重要手段，因此，股价走势预测方法的研巧具有重要价值和理论意义。证券市场是一个非常复杂的系统，股票价格趋势的分析是一个极具难度和挑战性的课题。随着信息技术和数据挖掘技术的发展，越来越多的新算法及模型被应用于股市数据的分析，这些方法对降低股票投资分析的难度和増加股票投资的收益具有一定的参考价值。

# 参考文献

[1]孙文高，刘剑桥．证券投资组合的风险与收益权衔[J]．中国经贸，2014(14)：131-131．

[2]王静红，李笔．基于决策树的一种改进算法[J]，电讯技术，2004，44(5)：175-178．

[3]张璐璐．基于关联规则的离群挖掘的研究[D]．安徽大学，2007．

[4]张胜权．基于数据按掘的股价走势预测巧[D]．华中科技大学，2009．

[5]黄智翔.数据挖掘技术在股价分析中应用与研究.云南大学，2016.