数据挖掘技术在股票预测中的应用探讨

西南大学计算机与信息科学学院　郭宇澄　许思远　魏正亚

【摘要】随着大数据技术的快速发展，数据挖掘技术也在股票行业得到了广泛的应用，本文主要通过对数据挖掘技术在股票市场的研究，来分析和预测股票价格的走势和应用。经过数据挖掘分类和聚类算法的分析，对股票市场大盘走势和个股走势进行分析研究，利用真实的、大量的数据来做测试，最终得到更适合投资者的投资结论。

【关键词】数据挖掘；股票预测；聚类算法

一、数据挖掘技术概述

数据挖掘是对数据库进行研究和应用的新技术。它的主要功能是通过对一些已存在的数据记录进行分析、统计，并得到投资者想要的结果。尤其是在股票行业的应用上，股票交易事务的分析过程中，每天都会将大量的客户交易信息来进行分析，将大量的数据汇入到数据库中，这些分析后的数据无疑是给股民对股市的走势有了更为正确的认识和了解，以便能做出更加正确的投资决策;经济学家将会对各个层次的用户的投资行为和股票之间的内在关系进行分析，及时对股市的一些不正常现象有所防范;也有一些上市公司和相关政府部门最新出台的有效解决方案提供了诸多方面的重要参考价值。股票市场是市场经济最为重要的特征，成立之日起就备受大量投资者的关注。股票市场的特征就是高风险和高回报，所以投资者会时刻关心着股市，并进行不断分析，试图预测出股市的发展趋势。近一百年来，根据股市的需求，一些数据的分析技术也随着也不断完善起来。如：道氏分析法、K线图分析法、点数图分析法、移动平均法、还有形态分析法、神秘级数与黄金分割比螺旋历法、四度空间法等[1]，随着数据挖掘技术在证券行业的普及和应用，将不断地推出新的分析方法。然而，从严格意义上说，数据挖掘技术仅仅只是分析手段的区别，还不能直接预测股市的发展动态。除此之外，人民也在不断地尝试预测股市，将利用回归分析的统计算法来进行分析和总结。然而，传统的预测技术对股市进行预测，将会是一个很艰苦的问题，那是因为有很庞大的数据源需要处理。然而目前由于股市的行情受到政治和经济等多方面因素的影响，使得内部走势规律变得很复杂，甚至有一些变化规律的周期有可能是一年或者是几年不等，所以我们必须要分析大量的数据才能得到预期。传统的预测技术并不能准确的判断出股市的走势，再近10年的发展中，数据挖掘的研发技术有了很大的进步，将各类数据挖掘技术都运用。极大地促进了人们对大量数据的分析、处理的能力，并未给人们带来很好的经济效益。总得来说数据挖掘技术在股市预测中将会有更大的发展潜力。

二、数据挖掘预测模型的建立原理

在数据挖掘的模型的建立，主要分为以下几个方面来建立，首先是确定问题，确定自己需要研究的方向;其次是准备阶段，在研究问题需要的数据;然后建立数据挖掘模型，来解决问题，最后对模型进行评价和评估。数据挖掘是一个不断递归的过程[2]，它内部有很多事物之间的相互联系，例如数据定义、数据分析、数据的预处理、算法的选择、提取规则和构建模型需要的知识。

（一）问题定义

在使用数据挖掘处理数据的时候，我们需要调查和分析股票行业，并进一步了解股票领域的内部运作情况。在确定相关人员对股票的走势上[3]，对目前所有已存在的资源和历史记录来进行评估，并确定利用数据挖掘技术处理后的数据是否满足股票投资者的需求，然后再利用数据挖掘技术制定出合适的数据挖掘计划。

（二）数据准备

数据挖掘技术一般是用来处理过量的、冗余、不完全的数据的手段。所以数据的准备过程中，主要包括对数据的抽取、清洗、转换和加载等一系列的过程，在将合适的数据准备妥当，以便测试使用。

（三）建立合适的模型

建立模型其实就是将目前已经存在和分析清楚的数据中的公有部分整合出来一种能够通用的模型，这种模型既可以将已知的数据进行合理的描述，又可以有效的将未知的数据和情况进行模型化，以便更好的预测未知的数据的走势。在数据挖掘的技术中，能够进一步的利用以下几种数据模型:例如:关联规则模型、决策树模型、神经网络模型、粗糙集模型、数据统计模型、时间序列分析模型[4]。

（四）模型的评价和评估

数据挖掘的有些模型是没有实际意义的，还有可能并不能清晰的呈现出股票下一步的走势，甚至还可能存在与实际数据相反的情况。所以，我们在建立模型的时候，首先要结合实际数据进行分析，并对数据结果进行正确评估，确定用模型得到的结果与我们想要的结果是否有偏差，是否正确。这样就能够更快的找到合适的模型，是否有存在的价值，是否能够解决用户的实际需求。

评估的方法是利用原有的数据建立的挖掘数据库中的数据来进行检验，同时利用另外新的测试数据来进行检验。最后在实际的运行环境上进行测试和审核校验[5]。

三、利用算法对数据处理的结论

数据挖掘对股票全部数据进行分析和处理，主要是利用分类算法不断地对数据进行处理，将属性从20维降低到8维，但是对于现在如此多的数据组成的数据集而言，没有特别显著的效果。这和我们所掌握的知识想吻合，用简单的规律是不足以描述股票的涨跌趋势的。用聚类Apriori算法挖掘个股000005世纪星源[6]，个股虽然只是全部区域的有限代表，但是我们还是有了新的发现。成交的金额和成交的数量基本上是以相同的趋势进行变化的。大部分的股票的收盘价都不是当前的最高价和最低价；成交量下降的股票，将不会再比最低收盘价低，对于成交额的变化趋势也是如此；使得股票的开盘价军高于之后五天的价格，然而，股票开盘前一天的价格不会比最低收盘价的价格低。这些预测出的股票数据走势就能做为一个参考，不能够起决定因素。但是利用数据挖掘技术建立模型来预测的股票数据就不同了，这种股票预测方式将是未来的发展趋势。

参考文献

[1]王冀宁,孔庆燕.信念、偏好及策略:基于股价波动的机构与散户的博弈研究[J].财政研究,2004(6).

[2]甘霖敏,杨忻.用人工神经网络方法对股票收益率影响因素的实证分析[J].清华大学学报(哲学社会科学版),2004(2).

[3]李兴绪.证券市场中的机构操纵行为研究—基于中国股市中机构与散户的博弈分析[J].数量经济技术经济研究,2003(8).

[4]吴德胜,梁裸殷尹.不同模型在财务预警实证中的比较研究[J].管理工程学报,2004(2).

[5]沈睿芳,郭立甫,时希杰.数据挖掘中的数据预处理模型与算法研究[J].计算机系统应用,2005(07).

[6]陈晓辉.基于数据挖掘与神经网络技术的股票预测模型的研究[J].科技风,2008(20).