**** 京 江 学 院



**JINGJIANG COLLEGE OF J I A N G S U U N I V E R S I T Y**

**读 书 笔 记**

专 业： 计算机科学与技术

班 级： 15计算机2班

学 号： 20150301202

姓 名： 陈 裕

指导教师姓名： 王 海 燕

指导教师职称： 副 教 授

2019年 2 月

1、时间序列分析法

时间序列预测方法是将股票价格或者价格指数看作变化的时间序列，通过建立合理的时间序列模型以预测未来发展变化的规律和趋势，而时间预测方法正迎合股指的变化发展的随机性及其时变性等特点，有较好的短期预测效果。常用的时间序列分析法有自回归模型(AR)、移动平均模型(MA)及自回归移动(ARMA)平均模型等。预测方法又有两种：一种是动态预测：只能进行一期预测，在由实际值预测出第一期的值之后，将第一期预测值带入时间序列，和历史数据一期再进行第二期的预测，以此递推，对于长期预测，可能会产生累计误差；另一种是静态预测：用原序列的实际值来进行预测，只有当真实数据可以获得时才可以使用这种方法。

2、时间模式挖掘

在讨论时间模式挖掘方法之前,需要对其相关的重要概念进行说明。

事件(event)是指时间序列中所发生的重要事情,而事件的重要性依赖于挖掘者的兴趣。时间模式(temporal pattern)则是时间序列中可以表征和预测事件的一系列隐含结构模式。时间模式p可以用重构的相空间中的点来描述,即有p∈Rd。假设有时间序列,可以定义事件特征函数(event characterization function)如下:



由于时间序列中往往存在随机干扰,而包含了许多噪声数据,所以用固定时间模式无法很好的匹配和预测事件。时间模式类(temporal pattern cluster)被定义为重构的相空间中能表征和预测事件的某些区域。假设有时间模式类C,提取预测点的时间模式p,如果p∈C,则认为预测点将发生时间模式类所表征的事件,将预测点作为事件点。

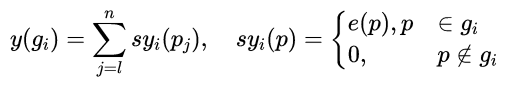
设有事件序列,其序列时间模式挖掘方法如下:

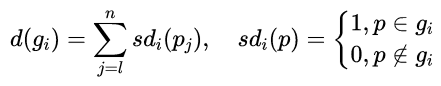
(1)网格划分

首先设定时间模式向量的维度和每一维度上值的变化区间[min,max],构建d维数据空间。每一维度在区间[min,max]上划分为m个间隔。网格的大小划分对于聚类的性能有较大影响,考虑到小波聚类的多分辨率特性,初始可以选取较小的划分间隔。

(2)量化空间

将序列 进行相空间重构,映射到d维数据空间。则原序列用d维时间模式向量描述为P，其中 ，根据向量的各个元素的值分配到对应网格。取时间特征函数为e(p),假设网格对象为gi,i=1,2,…,,定义累计事件特征函数y和累计密度函数d如下:





量化空间之后,网格单元gi可用一个二元组<yi,di>描述。

(3)应用小波变换聚类

原始小波聚类以密度为聚类指标,然而对于时间序列模式挖掘来说,挖掘者更关心的是模式表征事件的强度,然而如果仅仅基于累计事件特征函数作为聚类指标,则有可能导致聚类结果包含的点数过少,属于偶然事件,必然缺乏可信度。因而需要针对事件和密度两个指标聚类。由上小节的计算结果可以构建累计事件函数yi和累计密度di的两个信号空间,分别应用小波变换。

3、金融数据挖掘应用支持向量机研究现状

在金融领域，尤其是证券市场领域，主要有两个发展方向，分别是财务风险预警和金融市场投资分析。其中金融市场投资分析又包含两个方向，第一个方向将支持向量机作为非线性估计来预测市场趋势和波动，第二个方向是使用支持向量机作为中介进行选股，决定交易策略（例如长/短仓策略）。

文献中，Kao L J 等提出了一种集成小波变换，多元自适应回归样条（MARS）和支持向量回归（SVR）即称为小波-MARS-SVR 的股票价格预测模型，不仅解决了小波变换、系列选择的问题，同时也提高了预测精度。张建宽等人采用支持向量机 (SVM) 和最小二乘支持向量机(LSSVM)预测股票价格的涨跌，实验结果表明 SVM 和 LSSVM 在股票价格涨跌预测方面都有很好的效果，且 SVM 预测效果比 LSSVM 稳定。基于多项式拟合的离散数据趋势关键拐点识别方法识别出的关键拐点在其引领的上涨趋势的涨幅以及持续时间上，有更好的应用效果。毕军龙考虑到股票市场波动的非线性特征以及我国的 T+1 交易制度、序列最小优化（SMO）算法，将 PCA 降维与 SVM 相结合建立 PCA-SVM 模型，很好的应用到预测沪深300 指数每周涨跌的预测。谭梦羽则更多的与人工智能领域的信息处理方法相结合，建立了一个基于核主成分分析和稀疏的最小二乘支持向量机的分类模型。在金融大数据的背景下，第一次将深度信任网格用于金融数据的分类直接使用深度信任网格对原始数据进行重建和分类。朱磊使用统计学习中支持向量机中的 epsilon-SVR 支持向量回归机方法，对上证 50 所有成分股第二天收盘价和开盘价的预测。其中的核函数采用径向基核函数，并与 BP 神经网格预测模型对比了预测效果。涨跌预测命中率的均值达到 75.58%。

4、股票预测方法出现的问题与分析

市场投资者能够通过买卖价格不断变化的股票，从而赚取其中的溢价收益。然而，若不是资深投资者，不具备敏锐的市场观察力和扎实的金融理论和实操基础，想从股票交易市场中获取利润是存在很大风险的。因此，为了对股市变化规律有更精准的预测，尽可能把投资人会面临损失的投资风险降到最低，前人研究了人工经验以及时间序列预测法，目前兴起了机器学习预测法，并处在快速发展阶段，但都有各自的技术缺点。

人工经验预测法仅凭借人力来分析股票各相关因素之间的影响，进而用来对股票未来价格趋势做出一定判断。而当今社会是个数据爆炸的时代，这种人工预测的方法不但在节约时间和人力成本方面不占优势，还不易控制因人的情绪波动而直接影响到的趋势判断，从而对投资造成亏损。时间序列分析法依据时间和时间序列变量间的关联度，通过量化方法对历史时间序列对象建立相应的数学模型，对股票未来价值进行预测。然而股票价格是非线性的，因此时间序列分析方法不适合对股票市场进行相关的模型预测。

把机器学习的方法运用到国内股票预测市场时，目前有两种主流的算法，即 SVM 支持向量机和 ANN（ArtificialNeuralNetwork）人工神经网格算法。

依据模拟人类大脑中神经元结构，从而形成复杂神经元网格为基础的ANN 算法，不能避免因股票价格不稳定的波动而可能会给学习模型带来过度拟合的情况，会造成股票预测结果偏离实际结果，严重时完全失真。比如常用的 BP 神经网络这类 ANN 算法，通常是在修改设定的全值时采用一阶随机差分方程，容易让模型陷入局部最优解，另一方面，在经验风险最小化原则的基础上，也容易让模型出现过度拟合。

相比之下，在结构风险最小化原则基础上，SVM 预测算法泛化能力强，且非常适合用于处理非线性特征的股票市场数据，其原理是通过把具有非线性特征的数据有效映射到另一个高维空间，以此把问题转化成线性关系来处理。总的来说，SVM 模型对经验的依赖程度较小，并且可以得到模型的整体最优解，进而很好地克服如神经网格等股票预测算法所导致的局部极值相关问题，且具有较好的策略泛化能力。另外，SVM 算法下的学习机模型可以针对有限样本进行专门的算法设计，对策略模型所产生的经验风险和学习机的复杂度会依照结构风险最小化理论来进行调控，从而可以有效避免过学习现象出现。

5、数据挖掘的方法

数据挖掘的对象种类的不同决定着数据挖掘使用方法的不同，如对多媒体的数据挖掘、文本类挖掘、Web 数据挖掘和空间数据挖掘，依赖于可以使用的其他学科的技术，如以信息论或集合论的理论与方法为基础的挖掘、神经网络和遗传算法、模糊或粗糙集理论、知识表示、公式挖掘、统计分析或高性能计算；依赖于所挖掘的数据对象或类型等。数据挖掘的分类方法有很多:从挖掘技术的角度，可分为机器学习、统计学、可视化、模式识别、神经网络以及面向数据库或数据仓库的技术等；从数据类型的角度考虑，针对复杂的数据挖掘系统可分为多种数据挖掘结合使用技术、有效集成等技术；从应用的角度分析，针对不同的应用领域需要采用特定有效的方法，对于特定领域的任务则不能用普通的挖掘技术。利用这些模型来描述数据的相互关系及模式，建造仿真世界的模型，为生活提供更多有必要的决策信息，帮助我们对未来做出预测并指导投资。数据挖掘方法有多种，一般有五种比较典型的模型方法：关联分析、序列模式分析、分类分析、聚类分析和异常检测。

（1）关联分析

关联分析主要能针对海量数据进行分析，并利用挖掘算法发现隐藏在数据间的相互关系与规则，其规则形如“90%的股票 A 在股票 B 上涨后上涨”之类的知识，挖掘的过程中一般都要设定支持度-置信度阈值来评价规则的有效性。近年来许多研究者提出了多种关联规则的挖掘算法，如 Apriori、AIS、DHP 等算法。目前对关联规则的分析研究已经比较深入，除了对自身算法效率的改进优化之外，还常跟实际结合起来，得到广泛的应用。

（2）序列模式分析

序列模式分析与关联分析类似，即都是在数据中利用算法挖掘出数据背后的关系与规则，不同之处在于序列模式中的数据与规则具有时间先后顺序性，如“某一只股票在第一天上涨了 10%，而另一只股票在一天后上涨的概率是 70%”的规则，形成一个客户行为的‘A→B’模式。序列模式分析描述的思想是：针对含有交易时间先后顺序序列数据，采用序列函数其数据进行挖掘，得到的高频序列模式即为序列规则。在进行序列模式挖掘时，同样也需要通过用户输入的支持度与置信度阈值来实现对规则的筛选与评价。

（3）分类分析

分类分析的目的是确定对象属于哪个目标类，其挖掘任务是利用不同的分类技术对数据进行分析并分类。分类是先通过对训练集数据进行分析和计算，提前准确描述每个类别及建立模型以挖出分类规则，然后利用已知的目标类对数据库中剩下的数据进行分析记录并分类和预测。作为数据挖掘中最广泛使用的技术，目前已有多种分类分析模型得到应用，主要的分类模型有线性回归模型、支持向量机(Support Vector Machine,SVM)、贝叶斯分类(NB)、决策树模型、基本规则模型和神经网络模型。

（4）聚类分析

如果说分类分析是有指导的学习，那么聚类分析就是无指导的学习，这就是聚类分析与分类分析的根本差别。聚类分析的目的是通过挖掘技术能使得数按照各自相似的特点进行组合，形成的组内的数据相似性越大、组间的差别性越大，那么聚类挖掘的效果达到最好。目前主要是基于距离的大小来进行聚类，如欧式距离、海明距离等。通过聚类分析，可以帮助我们识别整个数据的分布模式，从不同的区域中得到想要的信息，分析数据集合之间的关系。聚类的应用十分广泛，常应用于商业、证券、生物技术、互联网服务等，聚类的方法主要包括神经网络法、机器学习法等等。

（5）异常检测

异常检测的目的是用于发现与大部分数据集不同的对象。由于来源不同、自然变异和测量误差等原因，数据常出现异常的状况。因此，为了找出与大部分数据集不同的对象，常采用异常检测的挖掘技术。异常检测是数据预处理的一部分，旨在改进常见数据对象分析的一种，其方法主要是发现参照值和观察结果之间的差别，主要基于模型的技术、基于邻近度的技术和基于密度的技术等。

6、序列关联规则的挖掘算法

早期的序列模式挖掘算法都是基于关联规则的经典算法 Apriori 得到的启发，其中的先验原理对序列数据仍然成立，因为包含特定 k-序列的任何数据序列必然包含该 k-序列的所有(k-1)-子序列。序列模式的挖掘过程同样是由候选集的产生，再到候选集的剪枝，最后由支持度的计数得到序列规则。基于这个启发，部分研究者提出了一系列的 Apriori 算法的改进算法：有在经典算法 Apriori 的基础上进行变形的算法，称为类 Apriori 算法，包括 AprioriAll、AprioriSome、DynamicSome 等，其具体思路与 Apriori 算法相似；GSP 算法则是在基于频繁模式的挖掘算法的基础上，在相邻元素之间添加了最大和最小间隔来增加时间约束，并定义滑动窗口来指定事务的范围，还使用了分层技术使得该算法显示出了较好的性能；序列模式中同样有不需要产生候选序列的算法 PrefixSpan 算法，便于处理非常大的序列数据库；仅需扫描三次数据库便可挖掘出所有序列规则的SPADE 算法，是利用格技术和简单的连接方法来提高序列模式挖掘性能；SPIRIT算法则是通过正则表达式来约束模式，避免了用户不感兴趣的模式等。有很多学者研究了序列关联规则在股市中的应用：有从 Apriori 算法里引出的比较经典的跨事务关联规则算法 E-Apriori 和 EH-Apriori，经比较后者的算法更加有效；有利用 hash 技术与 bit-search 算法进行的股票挖掘算法等等，这些算法的核心思想都是采用时间窗口滑动技术。

此外，还有有关序列模式挖掘的延伸的周期模式挖掘，即时间序列数据库上的再生模式，包括全周期模式挖掘、部分模式挖掘和周期关联规则挖掘；序列模式中包括单维序列模式挖掘和多维序列模式挖掘，多维序列模式的挖掘则是发现不同维度属性的信息模式，有 Seq-Dim 以及 UniSeq 算法；当已有的规则随时间变化而不再有效时，增量式序列模式挖掘能进一步对规则进行改善，能发现新的有趣模式，采用弱更新的方式来重新计算与增量有关的数据，以此维护序列模式。

7、随机游走模型

随机游走模型由首先由爱因斯坦在1926年以数学方式描述。由于自然界中的许多实体会以不可预知的方式移动，因此随机游走模型用来描述这种不稳定的移动。在这种移动模型中，移动节点随机选择一个方向和速度来从当前位置移动到新的位置。新的速度和方向分别从预定义的范围【speedmin，speedmax】和【0，2https://images2018.cnblogs.com/blog/1379566/201808/1379566-20180801103145397-1450596404.png】。移动节点的每次移动会以恒定的时间间隔t或恒定的行进距离d进行，结束后会计算新的方向和速度。如果此模型的移动节点到达模拟边界，则它将从模拟边界“弹回”，其角度有入射方向确定，然后沿着这条路径继续移动。

许多随机游走模型已经被研究，包括一维，二维，三维和d-维游走。在1921年，Polya证明在一维或二维的随机游走能够完全确定地返回远点，这一特征确保随机游走模型代表了一种移动模型----可以测试移动节点在其起点附近的移动，不用担心移动节点因游走而永远回不到起点。

二维随机游走模型是热点。下图显示了一个二维随机游走模型的仿真例子。移动节点在300\*600的模拟区域从起点（150,300）移动。在每个拐点，移动节点随机选择【0，2https://images2018.cnblogs.com/blog/1379566/201808/1379566-20180801103145397-1450596404.png】的方向，选择【0，10】m/s的速度。在改变方向与速度之前，移动节点移动60秒。在随机游走模型中，移动节点可以在行进指定距离之后改变方向而不是指定时间。图1的移动节点在改变方向和速度之前总共行进10步（而不是60秒），与图一不同，图二的每次移动都是完全相同的聚类。