**2017级信息与计算科学专业《数据挖掘原理与实践》课程论文选题**

**根据以下选题，自拟题目**

1. 数据挖掘在疫情控制中的应用
2. C4.5决策树分类算法的实现与性能分析
3. 最近邻分类算法的实现与性能分析
4. CART分类算法的实现与性能分析
5. 神经网络算法的实现与性能分析
6. 贝叶斯分类算法的实现与性能分析
7. 逻辑斯蒂回归算法的实现与性能分析
8. 支持向量机算法的实现与性能分析
9. 基于规则的分类算法的实现与性能分析
10. apriori关联规则挖掘算法的设计与分析
11. FP树关联规则挖掘算法的设计与分析
12. 频繁项集生成算法的实现与性能分析
13. K平均算法的实现技术与性能分析
14. 模糊c均值算法的实现与性能分析
15. DBSCAN算法的实现与性能分析
16. SNN聚类算法的实现与性能分析
17. 最小生成树聚类算法的实现与分析
18. 关联规则中支持度置信度及提升度的实际意义
19. 聚类算法性能比较及应用
20. 数据挖掘算法的评估
21. 基于。。。技术的行业应用（目标客户特征分析、用户路径分析、交叉销售模型、风险评级、推荐模型）
22. 多重共线性的处理方法
23. 数据挖掘与隐私保护
24. 基于数据挖掘的舆情分析
25. 数据挖掘在信用评价/风险控制方面的研究与应用

**下面是知网里查找的数据挖掘相关的论文题目，供拟题参考（知网里有很多）！**

[基于数据挖掘的房地产价格预测](http://kns.cnki.net/kns/detail/detail.aspx?QueryID=1&CurRec=4&recid=&FileName=TJJC201908020&DbName=CJFDPREP&DbCode=CJFQ&yx=A&pr=&URLID=42.1009.C.20190429.1729.021)

[基于数据挖掘的城市旅游舆情分析系统设计与实现](http://kns.cnki.net/kns/detail/detail.aspx?QueryID=1&CurRec=9&recid=&FileName=DZRU201908128&DbName=CJFDTEMP&DbCode=CJFQ&yx=Y&pr=&URLID=10.1108.TP.20190428.1405.244)

[大数据时代数据挖掘在银行中的应用](http://kns.cnki.net/kns/detail/detail.aspx?QueryID=1&CurRec=11&recid=&FileName=TXSJ201904079&DbName=CJFDTEMP&DbCode=CJFQ&yx=&pr=&URLID=)

[基于数据挖掘分类模型对住房价值进行实证分析](http://kns.cnki.net/kns/detail/detail.aspx?QueryID=1&CurRec=12&recid=&FileName=TXSJ201904146&DbName=CJFDTEMP&DbCode=CJFQ&yx=&pr=&URLID=)

[基于大数据时代的数据挖掘预处理技术研究](http://kns.cnki.net/kns/detail/detail.aspx?QueryID=1&CurRec=28&recid=&FileName=NASH201911188&DbName=CJFDTEMP&DbCode=CJFQ&yx=&pr=CJFY2019;&URLID=)

[基于大数据挖掘的新能源汽车主要指标及现状研究](http://kns.cnki.net/kns/detail/detail.aspx?QueryID=1&CurRec=63&recid=&FileName=SZSZ201901005&DbName=CJFDLAST2019&DbCode=CJFQ&yx=&pr=CJFR2019;&URLID=)

[信用评分模型比较综述——基于传统方法与数据挖掘的对比](http://kns.cnki.net/kns/detail/detail.aspx?QueryID=1&CurRec=72&recid=&FileName=HNJG201902013&DbName=CJFDLAST2019&DbCode=CJFQ&yx=Y&pr=&URLID=41.1407.f.20190307.1451.024)

[基于改进聚类算法的Web异常数据挖掘软件设计](http://kns.cnki.net/kns/detail/detail.aspx?QueryID=1&CurRec=84&recid=&FileName=XDDJ201908017&DbName=CJFDTEMP&DbCode=CJFQ&yx=A&pr=&URLID=61.1224.TN.20190409.1334.017)

[基于大数据时代的数据挖掘预处理技术研究](http://kns.cnki.net/kns/detail/detail.aspx?QueryID=4&CurRec=16&recid=&FileName=NASH201911188&DbName=CJFDTEMP&DbCode=CJFQ&yx=&pr=CJFY2019;&URLID=)

[网络日志数据的预处理技术研究](http://kns.cnki.net/kns/detail/detail.aspx?QueryID=8&CurRec=7&recid=&FileName=YNSK201804006&DbName=CJFDLAST2018&DbCode=CJFQ&yx=&pr=&URLID=)

[基于聚类算法的遥感云用户行为分析研究](http://kns.cnki.net/kns/detail/detail.aspx?QueryID=8&CurRec=14&recid=&FileName=1018877609.nh&DbName=CMFDTEMP&DbCode=CMFD&yx=&pr=&URLID=)

[基于决策树模型实现预测用户的购买行为](http://kns.cnki.net/kns/detail/detail.aspx?QueryID=11&CurRec=16&recid=&FileName=TXSJ201904144&DbName=CJFDTEMP&DbCode=CJFQ&yx=&pr=&URLID=)

[基于不平衡分类的……..预测方法的研究](http://kns.cnki.net/kns/detail/detail.aspx?QueryID=11&CurRec=40&recid=&FileName=XJYY201904028&DbName=CJFDTEMP&DbCode=CJFQ&yx=&pr=&URLID=)

[离群点挖掘在网购中的应用研究](http://kns.cnki.net/kns/detail/detail.aspx?QueryID=13&CurRec=7&recid=&FileName=1018210954.nh&DbName=CMFD201802&DbCode=CMFD&yx=&pr=&URLID=)

[离群点检测算法研究](http://kns.cnki.net/kns/detail/detail.aspx?QueryID=13&CurRec=9&recid=&FileName=1018960380.nh&DbName=CMFD201802&DbCode=CMFD&yx=&pr=&URLID=)

[基于偏离特征的离群点挖掘方法研究](http://kns.cnki.net/kns/detail/detail.aspx?QueryID=13&CurRec=17&recid=&FileName=1018081892.nh&DbName=CMFD201801&DbCode=CMFD&yx=&pr=&URLID=)

**可以从以上方面的方法、算法及已有算法在具体领域中的应用进行选题，但不限于此；也可选某一数据挖掘竞赛题目的某一小题进行求解成文；或者根据自己感兴趣的挖掘方法，根据某一领域数据的特点，写一篇关于数据挖掘的论文，题目自定（可对书上算法进行改进或书上算法在具体领域的应用方法）。要求分析领域数据的特点，数据预处理方法和步骤，挖掘的原理和步骤（或原算法的思想与步骤，改进的思想），挖掘程序设计，结果分析与评价（对比）等。**

**需提交文件：**

1. **课程论文**

**打印版+电子版（命名：学号姓名.doc）**

**2、综合性实验（电子版，命名学号姓名.doc或rar，必要的程序或数据应包括在内）**

说明：

（1）从网上下载或相互抄袭的一律不及格

（2）提供查重（截图）

**下面有课程论文的模板，提交版从下页开始；**

**请参照模板写你的课程论**文

**《数据挖掘原理与实践》**

**课 程 论 文**

**基于小波去噪的灰色和支持向量机组合预测模型**

|  |  |
| --- | --- |
| **姓名：** |  |
| **学号：** |  |
| **学院：** | **金融数学与统计学院** |
| **专业** | **信息与计算科学专业** |
| **提交日期：** |  |

**查重结果粘贴处：**

**2020年5月19日**

基于小波去噪的灰色和支持向量机组合预测模型

**摘要**：金融时序数据具有非平稳、含噪声等特点，采用单一的预测模型时预测精度并不理想。据此，文章构建了基于小波去噪的灰色和支持向量机组合预测模型，并对预测难度较高的我国热钱流动规模进行了预测。实验结果表明：与经典GM(1,1)幂模型相比，组合预测模型的平均相对误差由46.14%降到2.95%，预测精度有显著性提高。

**关键词**：小波去噪；灰色系统；支持向量机；组合模型

1 引言

金融时间序列数据往往具有高流动性、高敏感性和高干扰性，对金融时间序列的准确预测一直是个热点问题。有学者研究表明，单一或者不加去噪的预测算法精度较差[1]，难以满足要求。因此，本文尝试首先在数据预处理过程中去噪，并采取组合预测模型，最后针对热钱流动规模这一复杂问题进行算法性能测试。小波变换发展于20世纪80年代，它克服了短时傅里叶变换的缺点，可以对信号进行多尺度的分析，特别适合分析像热钱流动这样的非平稳数据[2]。基于阈值处理的小波去噪算法[3]由于计算量小、去噪效果好而被广泛研究和应用，它是一种直观而有效的去噪算法。灰色系统理论是1982年我国学者邓聚龙教授创立的一种研究少数据、贫信息不确定性问题的新方法。灰色预测方法独特的数据生成方式可以弱化数据的随机性，增强数据的规律性，其中GM(1,1)模型为灰色预测模型中的典型代表。GM(1,1)幂模型[4-5]是近年来才被广泛关注的一种非线性灰色模型，它可以通过寻找与实际数据最匹配的幂指数，从而使得模型能够较好地反映数据的非线性特征。它比传统的GM(1,1)模型和灰色Verhulst模型具有更广的适用范围和更高的模拟精度，但直接应用于非平稳序列在预测精度上并不理想。支持向量机是机器学习领域研究的最新成果，是上世纪90年代由Vapnik等在统计学习理论的基础上提出的一种新的机器学习方法[6]，它建立在统计学的VC维理论和结构风险最小原理基础上，克服了神经网络过拟合等缺陷，具有较强的自学习能力，在解决非线性问题上有着诸多优势，已成功应用于分类、函数逼近和时间序列预测等方面[7]。

……………………………………………………………………………………………..

2 小波阈值去噪原理

……………………………………………………………………………………………………..

硬阈值函数表达式为：

 （1）

软阈值函数表达式为：

 （2）

式（1）、（2）中，*λ*为阈值，*ωj,k*为小波系数，为处理后的小波系数。

3）小波重构：利用处理后的小波系数进行小波重构，则可以得到去噪后的信号，从而达到去噪的目的。

小波阈值去噪原理流程图如图1所示：



图1 小波阈值去噪原理流程图

3 GM(1,1)幂模型

**定义**1[14] 设，，如上所述，则称

， （3）

为GM(1,1)幂模型。将一阶微分方程

， （4）

称为GM(1,1)幂模型的白化方程。

通过求解GM(1,1)幂模型的白化方程，可得

 （5）

其中，幂指数、和为未知参数。现首先根据文献[4]提供的方法来估计幂指数，可得





， （6）

由得到幂指数的最优值。

……………………………………………………………………

4 支持向量机回归

对给定的训练样本集，其中是维输入变量，是对应的输出值，是训练样本的个数。构造回归函数如下：

 （10）

其中映射函数满足Mercer条件，映射的完成由构成的核函数实现。，分别是权向量和偏差量。根据结构风险最小化（SRM）原理，和可通过求解如下优化模型得到：



*s.t.*  （11）

.

.

…

5 应用研究

为了检验本文模型的有效性，我们收集了1983年­至2012年中国热钱流动数据，并参照了最新的热钱流动规模修正公式[15]进行了修正，得到最终的数据如表1所示。

表1 1983年­至2012年中国热钱数据

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 年份 | 热钱 | 年份 | 热钱 | 年份 | 热钱 |
| 1983 | 2.52 | 1993 | -152.48 | 2003 | 694.74 |
| 1984 | -6.96 | 1994 | -151.08 | 2004 | 1127.12 |
| 1985 | 66.10 | 1995 | -362.11 | 2005 | 1545.93 |
| 1986 | 86.93 | 1996 | -292.70 | 2006 | 651.80 |
| 1987 | 19.86 | 1997 | -591.35 | 2007 | 4666.92 |
| 1988 | 43.86 | 1998 | -874.85 | 2008 | 2624.79 |
| 1989 | 42.72 | 1999 | -668.59 | 2009 | 2279.68 |
| 1990 | -88.71 | 2000 | -201.85 | 2010 | 1811.39 |
| 1991 | -46.73 | 2001 | -368.56 | 2011 | 1307.78 |
| 1992 | -200.21 | 2002 | -181.53 | 2012 | 738.56 |

数据来源于国家统计局商务部、国家外汇管理局

……………………………………………………………………………….

最后，将GM(1,1)幂模型得到的预测值和残差预测值结合，可计算得2005-2012年热钱的预测值及相对误差。同时为了便于比较本文方法的有效性，还利用单一的GM(1,1)幂模型对原始数据进行建模预测，得到的结果见表3。

表3 2003-2012年热钱预测值及相对误差

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 年份 | | 原始值 | GM（1,1）幂模型 | | 组合模型 | |
| 预测值 | 相对误差（%） | 预测值 | 相对误差（%） |
| 2003 | | 694.74 | 694.74 |  |  |  |
| 2004 | | 1127.12 | 1066.41 | 5.39 |  |  |
| 2005 | | 1545.93 | 1862.40 | 20.47 | 1852.74 | 19.85 |
| 2006 | | 651.80 | 2559.47 | 292.68 | 653.90 | 0.32 |
| 2007 | | 4666.92 | 2861.14 | 38.69 | 4665.26 | 0.04 |
| 2008 | | 2624.79 | 2711.59 | 3.31 | 2623.28 | 0.06 |
| 2009 | | 2279.68 | 2270.11 | 0.42 | 2280.96 | 0.06 |
| 2010 | | 1811.39 | 1739.12 | 3.99 | 1805.72 | 0.31 |
| 2011 | | 1307.78 | 1253.15 | 4.18 | 1307.48 | 0.02 |
| 平均相对误差 | | |  | 46.14 |  | 2.95 |
| 2012 | 738.56 | | 866.50 | 17.32 | 706.25 | 4.37 |

从表3可以看出，组合模型预测精度显著高于单一的GM（1,1）幂模型。

7 结论

在偶然因素的影响下，金融时间序列具有较大的随机性，并非理想光滑的曲线，对其准确预测具有重要的意义。本文结合小波分析、GM(1,1)幂模型和支持向量机三种方法的优势，提出了一种基于小波去噪的灰色和支持向量机组合模型，并应用于金融时间序列中我国热钱时序数据的预测分析。通过与单一灰色GM(1,1)幂模型的对比，结果表明组合模型显著提高了预测精度，从而证明了其在金融时序预测中的有效性和实用性。

参考文献

[1]兰秋军，马超群，文凤华．金融时间序列去噪的小波变换方法．科技管理研究[J]，2004（06）：117-120．

[2]王瑞，张友纯．新阈值函数下的小波阈值去噪．计算机工程与应用[J]，2013,49（15）：215-218．

…………………………………………………