南 京 理 工 大 学

毕业设计(论文)开题报告

|  |  |  |  |
| --- | --- | --- | --- |
| **学生姓名:** | 王凡 | **学 号：** | 913106840624 |
| **专业(方向)：** | 软件工程 | | |
| **设计(论文)题目**： | 云计算环境下具有可变方差的虚拟机价格预测方法 | | |
| **指导教师:** | 蔡志成 | | |

2017 年 1 月 5 日

开题报告填写要求

1．开题报告（含文献综述）作为毕业设计（论文）答辩委员会对学生答辩资格审查的依据材料之一。此报告应在指导教师指导下，由学生在毕业设计（论文）工作前期内完成，经指导教师签署意见及所在专业审查通过后生效；

2．“文献综述”应按论文的格式成文，并直接填写在本开题报告第一栏目内，学生写文献综述的参考文献应不少于15篇（不包括辞典、手册）；

3．有关年月日等日期的填写，应当按照国标GB/T 7408—2005《数据元和交换格式、信息交换、日期和时间表示法》规定的要求，一律用阿拉伯数字书写。如“2014年3月15日”或“2014-03-15”。

**毕 业 设 计（论 文）开 题 报 告**

|  |
| --- |
| 1．结合毕业设计（论文）课题情况，根据所查阅的文献资料，每人撰写2000字左右的文献综述： |
| **文 献 综 述**  1.虚拟机实例价格预测的存在意义及研究现状  近年来，随着移动互联网的兴起，云计算资源的需求激增，越来越多的云计算资源服务提供商开始提供各种类型的云计算服务。国际上较知名的有微软的Azure、亚马逊的AWS等[4]。而在国内，阿里巴巴的ECS、腾讯公司的CVM以及新浪的SAE等虚拟机服务则占据了大片市场，成为许多个人开发者不可或缺的工具，为数以万计的人们提供云计算服务[3]。  作为一种特殊商品，虚拟机实例的价格并不是一尘不变的。随着时间变化，虚拟机实例的价格也在一定范围内产生波动。而消费者需要权衡价格、性能和可靠性之间的各种关系[4]。通过对实例历史价格的统计分析，我们期望得到价格的未来走势。随着越来越多的虚拟机商业化，对实例价格走势的分析显得愈加重要和有意义。通过对大量数据的建模分析，虚拟机实例提供商可以更加合理的调控价格，实现利益最大化。而消费者也迫切需要衡量价格、性能和可靠性之间的各种关系，通过统计分析，我们也希望能找到其平衡点[5]。  通过历史数据来预测未来走势的行为早已不是什么新鲜事，对于普通的统计分析，我们有一套非常成熟的建模方法。如AR模型、MA模型和ARMA模型等[8]。然而多数的分析方法对历史数据有较高的要求，它们普遍认为历史数据具有较强的规律性，通过寻找这些规律，利用相关数学模型得到预测数据。而我们知道，虚拟机实例价格受到非常多的因素的影响，并不能简单地认为它是有较强规律性的历史数据[5]。相反的，虚拟机实例价格存在一定的随机性，这就使得“找规律”式建模方式不再适用。为此，本课题希望通过利用非平稳时间序列的随机分析来对虚拟机实例价格进行预测。  2.虚拟机实例价格预测方法及数学模型  虚拟机实例的价格作为一种随时间变化的随机变量，在统计学科中我们将其称为时间序列。时间序列是统计学科的一个重要分支。其主要研究随着时间变化的事物发生、发展的过程，寻找事物发展变化的规律，并预测未来的走势[8]。  当我们得到一个虚拟机实例的历史价格即观察值序列之后，首先要检验它的平稳性和纯随机性，而这两个值的检验称为时间序列的预处理。根据检验结果，可以将序列分为不同的类型，对不同的类型我们采取不同的分析方法[8]。其中，平稳性是时间序列所具有的一种统计特征，一般情况下有两种定义方式。分别是“严平稳”和“宽平稳”[13]。所谓严平稳，就是一种条件较为苛刻的平稳性定义，其定义了只有当序列的所有统计性质都不会随着时间的推移而发生变化时，该序列才能被认为是平稳的。但是，因为在实践中获得随机序列的联合分布是非常困难的，而且即使知道随机序列的联合分布，计算和应用也相当的不便。所以严平稳时间序列通常只具有理论意义，在实践中用的更多的是条件较为宽松的宽平稳时间序列。宽平稳是使用序列的特征统计量来定义的一种平稳性。它认为序列的统计性质主要由它的低阶矩决定，所以只要保证序列低阶矩（二阶）平稳，就能保证序列的主要性质近似稳定[13]。  综上所述，当我们得到虚拟机实例的价格观察值序列之后，首先判断其平稳性。通过平稳性检验，可以把价格观察值序列分为平稳序列和非平稳序列两大类。  对于平稳序列，在统计上，我们通常建立一个线性模型来拟合该时间序列的发展情况，借此来提取其中有用的信息[2]。目前最常用的平稳序列拟合模型是ARMA模型。ARMA模型的全称是自回归移动平均模型，它又可以细分为AR模型、MA模型和ARMA模型三大类[13]。  对于非平稳序列，又可以分为确定性时序分析和随机时序分析两大类[13]。确定性时序分析虽然是非平稳的，但是这些时间序列具有非常显著的趋势，我们可以分解并找到这种趋势，并利用这种趋势对序列的未来发展做出合理的预测[6]。常用的时间序列分解方法有Wold分解定理和Cramer分解定理，而趋势分析方法则有趋势拟合法和平滑法等[13]。但是，就虚拟机实例价格分析这一课题而言，确定性因素分解方法虽然原理简单、操作简便且易于解释，但它只能提取强劲的确定性信息，对随机性信息则浪费严重。而我们知道，虚拟机实例价格的时间序列多是由随机性信息组成的，因而并不适用。此外，确定性因素分解方法把所有序列的变化都归结为四因素的综合影响，却始终无法提供明确的方法判断各大因素之间确切的作用关系[6]。这些问题显示出确定性因素分解方法在分析虚拟机实例价格方面的不足。而对于非平稳序列的随机分析就是为了弥补上述不足，提供更为精确的时序分析。  3.拟采用的研究路线  虚拟机实例价格在统计学科上是一种非平稳序列的随机分析时间序列。对于非平稳序列的随机分析，通常我们要通过有效手段提取随机序列中所蕴含的确定性信息。通过差分运算，我们可以校验非平稳序列随机分析中蕴含的平稳性[13]。如果非平稳序列在差分后显现出平稳序列的性质，我们可以将其称之为差分平稳序列[13]。这为我们预测分析非平稳序列提供了理论基础。  对于差分平稳序列可以采用求和自回归移动平均模型进行拟合（简称ARIMA模型）。ARIMA模型于1970年由Box和Jenkins提出，它已成为最经典的时间序列拟合模型之一[13]。但和传统的确定性因素分解方法比较,ARIMA模型仍有不足。它使用差分方法提取确定性信息，差分方法的优点是对确定性信息的提取比较充分，缺点是很难对模型进行直观解释[8]。而我们知道，由于节假日促销等因素的影响，虚拟机实例价格极有可能呈现出显著的季节效应和确定性趋势[7]。这时我们期望使用确定性因素分解方法拟合，却又极易造成残差信息的浪费[1]。为了解决这个问题，我们引入了残差自回归模型。1982年Engle在分析赢过通货膨胀序列时，发现经典的ARIMA模型始终无法取得理想的拟合效果。经过对残差序列的仔细分析，他发现残差序列具有异方差性[9]。即随机误差序列的方差不是常数，而是随着时间的变化而变化，这种情况叫做异方差。  综上所述，本课题拟采用条件异方差模型对虚拟机实例价格序列进行拟合。条件异方差模型分为ARCH模型（1982年由Engle提出）和GARCH模型。ARCH模型的实质是使用残差平方序列的q阶移动平均拟合当期异方差函数值。由于移动平均模型具有自相关系数q阶截尾性，所以ARCH模型实际上只适用于异方差函数短期自相关过程[13]。考虑到实际中虚拟机的价格序列的异方差函数可能具有的长期自相关性，这时如果使用ARCH模型拟合异方差函数，将会产生很高的移动平均阶数，增加参数估计的难度并最终影响拟合精度[13]。  为此，本课题最终决定采取GARCH模型对异方差函数进行拟合。GARCH模型全称广义自回归条件异方差模型，由Bollerslov于1985年提出。其在ARCH模型的基础上增加考虑了异方差函数的p阶自相关性，因而可以有效拟合具有长期记忆性的异方差函数[13]。  4. 参考文献：  [1] Galante G, Bona L C E D, Mury A R, et al. An Analysis of Public Clouds Elasticity in the Execution of Scientific Applications: a Survey[J]. Journal of Grid Computing, 2016, 14(2):193-216.  [2] Juve G, Chervenak A, Deelman E, et al. Characterizing and profiling scientific workflows[J]. Future Generation Computer Systems, 2013, 29(3):682-692.  [3] Ben-Yehuda O A, Ben-Yehuda M, Schuster A, et al. Deconstructing Amazon EC2 Spot Instance Pricing[C]// IEEE International Conference on Cloud Computing Technology and Science Proceedings. IEEE, 2011:304-311.  [4] Singh V K, Dutta K. Dynamic Price Prediction for Amazon Spot Instances[J]. 2015:1513-1520.  [5] Zhang Q, Zhu Q, Boutaba R. Dynamic Resource Allocation for Spot Markets in Cloud Computing Environments[C]. IEEE Computer Society, 2011:178-185.  [6] Dadashov E, Cetintemel U, Kraska T. Putting Analytics on the Spot: Or How to Lower the Cost for Analytics[J]. Internet Computing IEEE, 2014, 18(5):70-73.  [7] Lorido-Botran T, Miguel-Alonso J, Lozano J A. A Review of Auto-scaling Techniques for Elastic Applications in Cloud Environments[J]. Journal of Grid Computing, 2014, 12(4):559-592.  [8] Javadi B, Thulasiram R K, Buyya R. Statistical Modeling of Spot Instance Prices in Public Cloud Environments[C]// 2013 IEEE/ACM 6th International Conference on Utility and Cloud Computing. IEEE, 2011:219-228.  [9] Li X, Cai Z. Elastic Resource Provisioning for Cloud Workflow Applications[J]. IEEE Transactions on Automation Science & Engineering, 2015:1-16.  [10] 沈虹, 李小平. 带准备时间和截止期约束的云服务工作流调度算法[J]. 通信学报, 2015, 36(6):183-192.  [11] 刘少伟, 孔令梅, 任开军,等. 云环境下优化科学工作流执行性能的两阶段数据放置与任务调度策略[J]. 计算机学报, 2011, 34(11):2121-2130.  [12] 陈冬林, 陈玲, 马明明,等. 云计算IaaS现货实例定价方法研究[J]. 计算机工程与设计, 2013, 34(10):3366-3370.  [13] 王燕. 时间序列分析——基于R[M]. 基于R应用的统计学丛书,2015,3:44-173 |

**毕 业 设 计（论 文）开 题 报 告**

|  |
| --- |
| 1. 本课题要研究或解决的问题和拟采用的研究手段（途径）： |
| 1. 问题描述    1. 竞价实例及时间序列的定义   竞价型实例是指可以根据自身需要启动任意数量的虚拟服务器、配置安全和网络以及管理存储。竞价型实例指定用户愿意为运行特定实例类型支付的最高小时价格。现货价格随供需浮动，但用户支付的价格不会超过您所指定的最高价。  按照时间顺序把随机事件变化发展的过程记录下来就构成了时间序列。在统计研究中，常用按时间排序的一局随机变量来表示一个随机事件的时间序列，简记为或。用表示该随机序列的n个有序观察值，称为序列长度为的观察值序列。   * 1. 本课题拟解决的问题   本次毕业设计的目的主要是学习时间序列的数学模型推导思想和建立过程，利用广义自回归条件异方差模型（GARCH）对虚拟机实例的价格序列进行建模，设计基于合适的的竞价型实例价格预测算法，开发软件界面，进行大规模数据验证。   1. 拟采用的技术路线    1. 总体技术路线   设计基于GARCH的虚拟机价格预测算法，利用R语言进行统计分析，开发可视化的Java程序界面，进行数据的加载和预测结果的展示。  设计数据分析算法  学习GARCH模型  Java开发可视化界面  R语言统计分析  数据结果展示  调用  数据文件输出   * 1. 具体实现方法      1. 对不同的时间序列进行分类，针对不同类型的时间序列采用不同的分析方法。虚拟机实例的价格作为一种随时间变化的随机变量，在统计学科中我们将其称为时间序列。时间序列是统计学科的一个重要分支。   对于异方差序列，有两种处理思路：假如已知异方差函数的具体形式，则进行方差齐性变换；假如不知异方差函数的具体形式，则拟合条件异方差模型。  在分析虚拟机实例价格序列时，我们只能根据残差图及残差平方图所显示出来的特点，使用一些常用的函数形式估计异方差函数。然而根据宏观经济领域对金融序列的研究分析，证明这种假定过于单一化，价格序列的异方差特征很多，我们并不能通过对数变换将其转换成方差齐性序列。  因此本课题采用条件异方差模型对虚拟机实例价格序列进行拟合。   * + 1. 如果方差齐性假定不成立，即随机误差序列的方差不再是常数，而是随着时间的变化而变化，可以表示为关于时间的函数：   这种情况称之为异方差。  事实上，方差齐性假定在价格分析中并不总是被满足的，异方差普遍存在。为了提高我们的模型拟合精度，需要对残差序列进行方差齐性检验，并对异方差序列进行深入分析。  由于残差序列的方差实际上就是其平方的期望，即  所以残差序列是否具备方差齐性，主要考察的性质。如果方差齐性满足，则，这意味着应该在某个常数值附近波动，它不应该具有任何明显的趋势，否则就呈现出异方差性。   * + 1. 利用GARCH模型对虚拟机实例的价格序列进行建模，设计基于GARCH的虚拟机价格预测方法。GARCH模型可以修正ARCH模型在异方差函数具有长期自相关性下产生的拟合误差。它的结构如下：   式中,为的确定性信息拟合模型，。这个模型简记为GARCH(p,q)。   * + 1. R语言内置多种统计学及数字分析功能，本课题拟开发可视化的Java程序界面，通过调用R语言，满足数据的加载和预测结果的展示。   R的功能也可以通过安装包（Packages，用户撰写的功能）增强。因为S的血缘，R比其他统计学或数学专用的编程语言有更强的面向对象（面向对象程序设计）功能。  对于数据的加载和预测结果的展示采用Java进行开发。显然R是统计计算的强大工具，而Java是做应用系统的主流语言，两者天然具有整合的需要。关于整合，一方面，R中可以创建Java对象调用Java方法，另一方面，Java中可以转换R的数据类型调用R的函数，互相取长补短。Java调用R函数常见的有两种方式：JRI 和 Rserve。本课题初步拟采用 Rserve作为数据同步工具。这是一个基于TCP/IP的服务器，通过二进制协议传输数据，可以提供远程连接，使得客户端语言能够调用R。目前Rserve作为一个package发布在CRAN上，可以直接使用install.packages(“Rserve”)进行安装。需要使用时在R控制台下加载该包，然后输入命令Rserve()，开启服务器，就可以供客户端调用。   * 1. 自回归条件异方差模型   文献综述已经提到，1982年Engle提出了自回归条件异方差模型（ARCH模型）。假设在历史数据已知的情况下，零均值、纯随机残差序列具有异方差性  在正态分布假定下，有  异方差等价于残差平方的均值  使用残差平方序列的自相关系数，可以考察异方差函数的自相关性。  考察的结果有如下两种：  a).自相关系数恒为零，即  这说明异方差函数是纯随机的。此时历史数据对未来异方差的估计一点作用都没有，至今也没有有效方法提取其中的异方差信息，这种情况本课题不做讨论。  b).存在某个自相关系数不为零，即  这说明在残差平方序列中蕴藏着某种相关信息，可以通过构造适当模型提取这些相关信息，以获得序列异方差波动特征。  具有结构的模型称为q阶自回归条件异方差模型，即ARCH(q)。  ARCH模型旨在利用历史信息，得到条件方差信息  得到一个观察值序列之后，通常首先提取序列的水平相关信息，然后分析残差序列中所蕴含的波动相关信息。将这二者综合起来才能得到较为完整和精确的分析结果。ARCH模型提取异方差中所蕴含的相关信息完整结构为  式中，为的确定性信息拟合模型，。  信息提取后需要对序列进行拟合，而要拟合ARCH模型，需要进行ARCH检验。ARCH检验是一种特殊的异方差检验，它不仅要求序列具有异方差性，而且要求这种异方差性是由于某种自相关关系造成的，这种自相关关系可以用残差序列的自回归模型进行拟合。常用的两种ARCH检验统计方法是Portmanteau Q检验和LM检验。  a).Portmanteau Q检验  Portmanteau Q检验的构造思想是：如果残差序列方差非齐且具有集群效应，那么残差平方序列通常具有自相关性。所以方差非齐检验可以转化成残差平方序列的自相关检验。  Portmanteau Q检验的假设条件为  用表示残差平方序列的延迟k阶自相关系数，则该假设条件可等价表达为  Portmanteau Q检验统计量其实就是的LB统计量  式中，n为观察序列长度，为残差序列延迟i阶自相关系数，有  原假设成立时，Portmanteau Q统计量近似服从自由度为q-1的分布  当检验统计量的P值小于显著水平α时，拒绝原假设，认为该序列方差非齐且具有自相关关系。  b).LM检验  LM检验全称为拉格朗日乘子检验。其构造思想为：如果残差序列方差非齐且具有集群效应，那么残差平方序列通常具有自相关性。就可以尝试通过自回归模型拟合残差平方序列  于是方差齐性检验就可以转化为这个方程是否显著成立的检验。  拉格朗日乘子检验的假设条件为  对残差平方序列构造q阶自回归方程，假设条件等价为  记总误差平方和为，自由度为。回归平方和为，自由度为q。其中，SSE为回归方程残差平方和，，自由度为。则LM检验统计量为  原假设成立时，LM(q)近似服从自由度为的分布  当LM(q)检验统计量的P值小于显著水平α时，拒绝原假设，认为该序列方差非齐，并且可以用q阶自回归模型拟合残差平方序列中的自相关关系。  通过LM检验和Q检验，我们可以分析虚拟机实例价格序列方差是否非齐。如果检验显示出序列方差非齐且具有显著自相关关系，则可以利用ARCH模型提取残差平方序列中所蕴含的相关信息。如果检验显示出高阶模型显著成立，即ARCH(q)模型的阶数q较大，则说明残差平方序列具有长期相关性，可采用GARCH(广义自回归条件异方差)模型进行拟合。  GARCH模型可以修正ARCH模型在异方差函数具有长期自相关性下产生的拟合误差。它的结构如下：  式中,为的确定性信息拟合模型，。这个模型简记为GARCH(p,q)。   * 1. 实验手段   本课题采用R语言作为统计分析的工具。在R语言中，通过调用TSA包中的garch函数，可以进行ARCH模型和GARCH模型拟合。garch函数的命令格式为：  式中，-x为序列名称，-order为拟合模型阶数。拟合ARCH(q)模型时order=c(0,q)，拟合GARCH(p,q)模型时order=c(p,q)。从这个细节也可以看出GARCH模型是ARCH模型的更普遍模式。  GARCH模型其实就是在ARCH模型的基础上，增加考虑了异方差函数的p阶自相关性而形成的。显然ARCH(q)模型就是p=0的GARCH(p,q)模型，可以看作是GARCH模型的一个特例。GARCH模型可以有效拟合具有长期记忆性的异方差函数，这也是本课题采用GARCH模型的重要原因。   * 1. 实验进度   2016.11.01-2016.12.30 翻译文献资料及撰写开题报告  2017.01.01-2017.02.30 利用基础知识，设计预测算法  2017.03.01-2017.03.30 利用R语言，进行大规模算法验证  2017.04.01-2017.05.15 毕业设计论文写作   1. 课题整体（或软硬件系统）解决方案及其研究思路    1. 利用时间序列统计分析的数学方法，基于现有的科学研究手段，对竞价型虚拟机实例的价格序列进行合理的分析。理解并掌握时间序列拟合的各类方法，并在此基础上设计系统性的预测算法，对竞价型虚拟机实例价格序列的残差平方分布进行合理的预测。    2. 与同小组的同学进行讨论，交流和分析各自算法的优劣。定期与指导老师和小组成员沟通，时刻注意课题的进度，不断完善数学模型并优化算法，做到取长补短。学习使用常用的开发工具，共享学习的经验，彼此之间互相帮助。    3. 使用现代化的编程语言编写可视化软件，满足数据的加载和批量处理，并进行大规模的算法验证。开发过程中利用第三方代码仓库进行管理，熟练掌握代码版本的控制和项目的管理。    4. 评价竞价型虚拟机实例的价格波动对市场经济的作用以及对互联网的影响，正确认识本课题所开发的软件系统对社会可持续发展的积极作用，以及该系统在互联网大环境下所承担的责任。    5. 对于课题所涉及的数学模型，要从实用性、准确度等多个角度进行评价，明确时间序列分析预测的随机性。预测算法的预期结果不可能和实际结果完全一致，应当不断致力于提高其预测的精度。如今神经网络算法起步不久，此课题只是冰山一角，此类研究还有很大的学习空间。 |