北京航空航天大学计算机学院

学位论文文献综述

论文题目：数据驱动的期货投资人风险评估模型研究

专　　业：计算机软件与理论

研究方向：数据挖掘

研 究 生：邓雅琳

学　　号：ZY1606101

指导教师：葛声

北京航空航天大学计算机学院

2017 年 12 月 20 日

目　　录

数据驱动的期货投资人风险评估模型研究

1 摘要

机器学习及深度学习技术近年来发展迅猛，计算机技术的发展为金融交易带来了巨大的变革，如近年来区块链技术、股票与期货等金融衍生品的高频交易平台、程序化交易技术等。从业的交易员和分析员长期接触实际交易，对于各种现象和问题有深入的了解，但是大量的金融交易和金融数据仅依靠少数的交易员和分析员，无法满足对海量数据和实时更新的分析需求。因此，应用计算机领域数据挖掘相关技术，对金融交易数据进行分析是十分必要的。

国内外研究使用支持向量机、神经网络等多种机器学习技术，在价格指数、波动率等预测、交易策略实践等方面取得了众多研究成果；在风险分析方面经济学领域使用了一些实证方法，机器学习技术在如金融欺诈、自然灾害等风险预测中也有突出的表现。然而，使用交易数据及市场价格数据对期货公司投资人风险进行分析评估的相关问题还有待研究，期货公司需要使用先进的机器学习算法，以数据为驱动建立智能化的风险预测模型，以此作为风险防范的有效手段。

本文将基于机器学习相关技术对期货公司投资人特征进行分析，对其杠杆利用因素以及账户风险进行研究。本文首先介绍对研究问题和使用的数据，然后介绍了课题研究背景和意义，下个部分是研究中使用的相关算法介绍，最后是总结和展望以及参考文献。

关键词：机器学习，风险分析，期货，杠杆

Abstract:

In recent years, machine learning and deep learning has developed rapidly in recent years. The development of computer technology has brought tremendous changes in financial transactions. The traders and analysts have long-term exposure to actual transactions and they have in-depth knowledge of various phenomena and

issues. However, only relying on them is unable to meet the analytical needs for massive amounts of data and the requirements of real-time updating. Therefore, it's necessary to apply the machine learning technology to financial transaction data.

Domestic and foreign research in forecasting price index and volatility have made a lot of results, using machine learning technology such as Support Vector Machine and neural network. Some empirical approaches are used in economics research for risk analysis. Machine learning also has outstanding performance in risk prediction, such as financial fraud and natural disasters. However, the issue of using transaction data and market price to analyze the risk of futures companies is still unsolved. The futures companies need to use machine learning algorithms and build intelligent risk prediction models which are driven by data. The models are effective tools of risk prevention.

This study will use machine learning technologies and the futures company's transaction data, for analyzing the leverage factors and account risk. In this paper, the data and research questions are briefly introduced. Then, the research background and significance will be introduced briefly. Then introduce the algorithms used in the study. Final part concludes the summarize, prospect and references.

关键词：Machine learning, risk analysis, futures, leverage

2 论文选题的背景与意义

本文对期货投资者风险和杠杆影响因素进行分析。使用的数据包括期货公司投资人的交易数据和期货交易市场每日价格数据。交易数据包含账户信息、每日资金结算数据、每日持仓数据、每日平仓数据、每日成交数据、每日委托数据等；期货市场价格数据主要来自大连、上海和郑州交易所。

计算机技术的发展为金融交易带来了巨大的变革，如近年来区块链技术、股票与期货等金融衍生品的高频交易平台、程序化交易技术等。从业的交易员和分析员长期接触实际交易，对于各种现象和问题有深入的了解，但是大量的金融交易和金融数据仅依靠少数的交易员和分析员，无法满足对海量数据和实时更新的

分析需求。因此，应用计算机领域数据挖掘相关技术，对金融交易数据进行分析是十分必要的。数据挖掘一般指从大量的、存在噪声或缺失的、随机的、模糊的数据中，发现人们未知的、隐含的、规律性的有价值信息和知识，涉及的技术包括统计学、情报检索、机器学习、专家系统和模式识别等方法。本文拟结合统计学、机器学习和专家系统中的方法，解决期货投资人风险分析的问题。

期货是包含金融工具或未来交割实物商品销售的金融工具，按照现货标的物种类，可以分为商品期货和金融期货两大类。期货交易市场实行保证金制度，仅需要少量资金就可以做大宗买卖，具有极强的“杠杆效应”，并且存在做空机制，因此是一个大风险的市场。经过 30 年左右的探索和实践，我国期货市场规模快速增长，运行质量不断提升，市场功能逐步发挥，产品体系不断完善，期货市场服务实体经济的能力逐步得到认可。然而，与欧美发达的期货市场相比，我国期货市场在发展环境、运行机制、市场规模、品种体系等方面都还存在差距。与我国实体经济整体发展相比，我国期货市场的发展仍相对滞后，形成的市场价格体系层次不够丰富，现有的风险管理工具种类不够多样，不能很好的满足实体经济风险管理的需要。

虽然我国在期货市场推出阶段进行了理论和技术层面的研究和准备，但由于发展历程较短、市场监管和交易机制不完善，中介机构风控措施单一，投资者甚至期货从业人员的专业化程度仍参差不齐，存在浓厚的投机氛围。

对于期货公司而言，如果很多客户同时出现极端亏损，可能导致大量强行平仓问题，不能充分满足投资人的需求，另一方面也对公司资金的稳定性造成影响，因此监控注册会员的账户风险十分重要。期货公司一般在市场保证金比例的基础上设置更高的公司保证金比例，以保证客户账户金额可能不足时存在缓冲区，但对于不同的客户公司没有设置相同的比例，因此对于投资人保证金杠杆的利用研究也值得关注。

3 国内外研究现状及发展动态

国内外研究现状及发展情况从国内外金融领域研究现状、国内外风险分析研究现状两个部分进行介绍。

3.1 国内外金融领域研究现状

股市预测一直是金融领域研究的热点，最广泛应用的模型之一是线性平稳时间序列自回归模型（AR），例如[1]应用 AR 模型分析中国股指动态收益，[2]将分解股票收益并利用 AR 模型进行预测。但股票价格高度非线性和非平稳限制了 AR 模型的适用性，因此[3]对印度基准指数使用了基于 SVM-KNN 的股票市场趋势变化分析方法，[4]使用隐马尔科夫模型（HMM）进行股票走势的非线性预测。随着神经网络逐步发展，在股市预测领域也出现了一些应用，比如[5]对比分析 ARIMA 模型和神经网络在股票价格预测走势，贝叶斯正则化的人工神经网络[6]也应用于股市预测，[7]结合 SVM、人工神经网络和随机森林提出了一个用于股价预测的模型。

LSTM[8]能够解决时间序列长短时间依赖的问题，所以在其被提出后得到广泛应用。RNN 和 LSTM 也在金融预测方面获得了关注，比如[9]结合 RNN 与 AR 预测股市收益，[10]LSTM 通过处理数字和文本数据预测股票价格，[11]LSTM 模拟交易策略。考虑到股市中存在长期和短期交易者，[12]受到离散傅里叶变换的启发，提出多频率交易模式 SFM 神经元结构，预测股票长期和短期价格。除此之外，[13]提取新闻词向量并使用 CNN 对 S&P 500 指数及个股价格进行预测，另外，他们还使用 NLP 技术处理事件文本信息和知识图提供的实体、属性等信息预测股市波动率。[15]使用混合神经网络以及遗传算法等组成的启发式算法预测股票价格，[16]使用回归树、自组织映射（SOM）及聚类方法，处理股票收盘价格，利用推荐系统为用户买卖操作提出建议。

在金融领域，[17]介绍和总结了大宗商品期货的研究发展，[9]使用 Logistic 回归对不同机构类型或个人买卖决策因素进行分析，也有许多其他关于投资者行为的研究[19,20,21]。

3.2 国内外风险分析研究现状

风险分析是一个跨学科的研究内容，从经济领域中的财务风险、上市风险、信用风险等到计算机领域的信息安全、工程系统风险，再到自然资源管理等方面，

风险无处不在，风险分析包括风险控制、风险预测、风险评估、风险识别等，不同问题的分析方法也存在差异。如[22]采用 Zigbee 通信收集环境温度、湿度等数据，使用 Apriori 算法和 Pearson 相关系数分析指数间的关系，对健康风险、行业风险等进行了预测。[23]设计了一个混合专家系统用于预测森林火灾的规模。[24]使用贝叶斯网络作为软件项目开发中需求工程风险评估的自动化工具。

在金融风险分析方面也有多种研究方法。Bellanger 和 Serge（1933）认为按照风险来源可以讲股指期货风险分为市场风险和非市场风险，他们认为市场风险即价格风险，包括流动性风险、信用风险、操作风险等，是股指期货市场的主要风险，非市场风险指法律和政策风险。随着统计学广泛应用，T.Linsmeier 和 N.Pearson（2001）研究和归纳了 Value at risk 方法的定义和计算方法，计算方法包括历史模拟法、蒙特卡洛模拟法和方差-协方差法[25]。为了准确的刻画价格的波动性，Engle（1982）提出来自回归条件异方差（ARCH）模型[26]；Bollerslev 于 1986 年在此基础上提出广义自回归条件异方差（GARCH）模型[27]；1991 年，Nelson 提出了指数 GARCH 模型（EGARCH），假设条件方差是非线性、非对称的，并且不要求参数非负[28]。将神经网络应用于金融风险分析的相关研究，包括使用神经网络预测银行倒闭风险[29,30]、分析公司破产风险[31,32]，[33]使用 NN 对商业银行信用风险进行判别，[34]使用贝叶斯神经网络对股指期货风险进行预警。

机器学习是一种重要的金融科技创新手段，近年来在国内外金融机构和金融科技企业中被尝试应用于风险防范、反欺诈等领域。例如花期银行、汇丰银行等广泛应用逻辑回归、神经网络等技术以提升欺诈识别能力，在大规模数据上进行全方位的综合考虑，挖掘深层次业务场景特征而建立监督、无监督等类型的学习模型。

4 相关算法

首先，期货投资人账户资金分为两个部分，一部分为公司要求的持仓保证金、另一部分为账户余额，当账户资金不能达到持仓保证金要求时，公司会通知客户平仓或追加资金。因此，分析投资人账户余额全部亏损的概率能够为此情况提供

预警信息，对不同风险程度的投资人采取不同的措施。

第二部分，期货交易的基本机制是杠杆，对投资人的杠杆影响因素进行分析，如操作品种数量、做多做空频次、年龄、自然人还是法人等因素，能够获得影响因素的重要程度排序，并且杠杆越高的人风险程度越高。

最后，除了以上用到的显著因素外，结合用户交易习惯，如在保证金不足时投资人倾向于平仓还是注资等多种因素，预测投资人是否会达到极端亏损，分析用户的风险程度。

本文基于期货公司投资人的交易数据进行研究，主要研究目标分为以下几个部分：

（1）分析投资人账户余额全部亏损的概率，评价其风险；

（2）分析影响杠杆利用程度的影响因素，高杠杆投资人风险较高；

（3）使用以上相关特征及用户习惯特征，预测投资人是否会达到极端亏损。

交易数据需要经过预处理和提取才能作为特征使用，比如用户年龄、地区需要从身份证中截取并计算获得，持仓时间、交易频次、交易量大小等也需要划定时间长度经过计算获得。其中存在一些总交易次数过少的数据需要过滤去除，对于用户没有交易的时间段相关数据也需要经过过滤。另外，投资人杠杆等于当日持仓总价值除以当日保证金数量；投资人风险程度等于当日公司保证金数量除以当日权益，该值大于 1 时风险水平为 1，该值大于 1.2 时风险水平为 2，这两种情况均定义为极端亏损。

4.1 Var 模型、Garch 模型、聚类算法

在对账户余额全部亏损风险概率进行分析时，拟使用 Var 模型、Garch 模型、EWMA 模型和聚类方法。采用 Var 模型和 GARCH 模型分析风险，并对风险结果及资金量等重要特征使用 K-means 聚类，将不同风险的投资人分开。

（1）Value at Risk 模型估计亏损风险

Var（Value at risk）即风险价值，假设市场价格波动率的分布符合正态分布。Var 定义为在一定的置信水平下，某资产组合在未来特定一段时间内，发生不超过某一范围亏损的最大值，如图所示。用公式表示为：

$$P(\Delta P \Delta t \le Var) = a$$

其中 P 表示资产价值损失小于可能损失上限的概率；ΔP 表示资产在一定持有期 Δt 的价值损失额；Var 为可能的损失上限；a 为给定的置信水平。

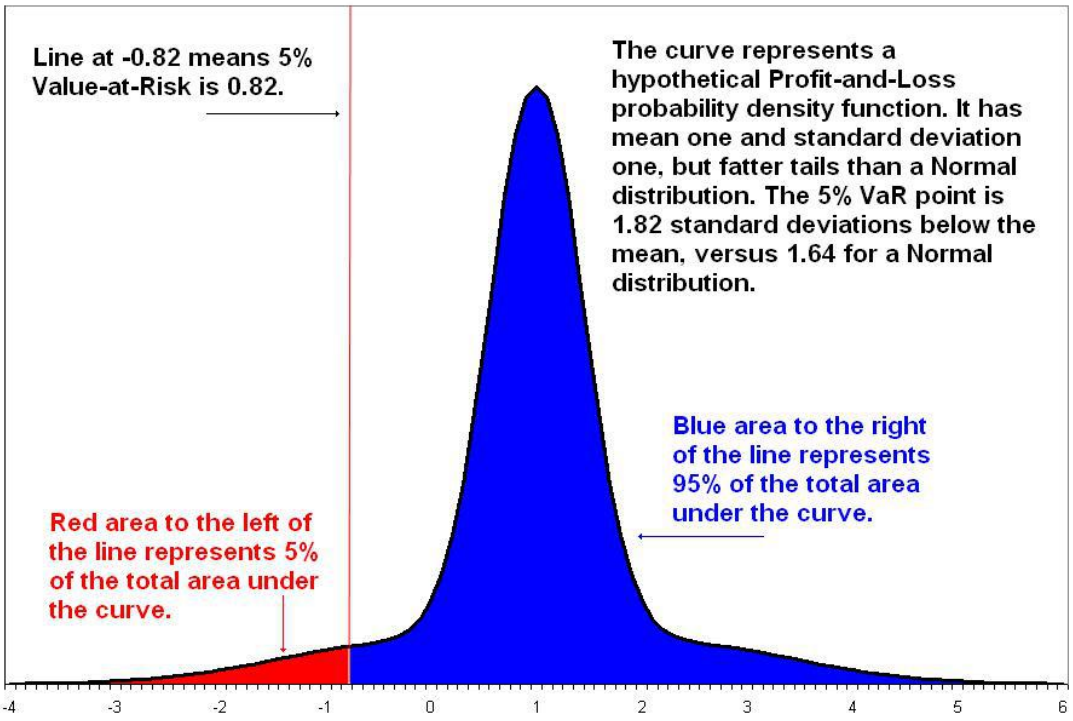

图 1 Var 模型

（2）GARCH 模型估计合约波动率及相关系数

使用 GARCH 模型最大似然估计每日每个合约的波动率，并计算两两合约的相关系数，然后对每个投资人当日持仓组合计算组合的波动率和均值；将波动率和均值带入 Var 模型，根据账户余额分析全部亏损的概率。

GARCH 用于预测波动率，输入为合约每日结算价，估计值为波动率：

$$\sigma_n^2 = \omega + \alpha u_{n-1}^2 + \beta \sigma_{n-1}^2$$

其中 u 表示结算价，σ 表示波动率，ω、α、β 为估计的参数，一般需要满足 $\omega > 0$ 且 $\alpha + \beta < 1$。GARCH 模型的参数使用最大似然法估计获得。另外，需要根据两两合约间的相关系数和投资组合占比计算整体投资持仓的波动率。

在期货投资人余额全部亏损分析中，GARCH 模型存在约束条件 $\omega > 0, \alpha + \beta < 1$，在不满足两个条件时不能使用该模型，此时使用指数加权移动

平均模型（Exponentially Weighted Moving Average, EWMA）进行波动率估计 $\sigma_i^2 = \lambda u_i + (1-\lambda)\sigma_{i-1}^2$。在使用 Var 模型时存在假设——波动率是符合正态分布的，但大量研究发现波动率的分布相对于正态分布存在“尖峰厚尾”的特点，因此对于波动率的分布可以考虑采用神经网络拟合。另外，投资组合的波动率均值采取持仓价值加权平均波动率，并且由于买卖开仓同时涨价带来的波动是相互抵消的，因此对卖仓设置-1 的系数。

在账户余额风险分析中，由于波动率存在随交割期临近而增大的规律，因此用于计算波动率的结算价选取同一品种同期最近三个月的数据，保证波动率的稳定性。其他特征如成交频次、交易种数等，包括用户行为习惯构造特征均需要通过定义时间窗并经过计算获得。

（3）聚类算法

在分析了投资人资金全部亏损的概率风险后，需要一定的表现形式将投资人风险展示出来，仅基于风险值对用户进行划分并不合理，比如 A 投资人和 B 投资人风险均为 70%，但 A 投资人账户余额为 100 万，B 投资人账户余额仅为 10 万。因此，需要结合如资金量、持仓价值、某些客户属性等和风险对投资人进行划分，也就需要使用聚类算法。

聚类算法主要包括：

——划分法（partitioning methods），给定一个有 N 个元组或者纪录的数据集，分裂法将构造 K 个分组，每一个分组就代表一个聚类，K<N。而且这 K 个分组满足每一个分组至少包含一个数据纪录、每一个数据纪录属于且仅属于一个分组。大部分划分方法是基于距离的。给定要构建的分区数 k，划分方法首先创建一个初始化划分；然后，它采用一种迭代的重定位技术，通过把对象从一个组移动到另一个组来进行划分，一个好的划分的一般准备是：同一个簇中的对象尽可能相互接近或相关，而不同的簇中的对象尽可能远离或不同。除距离外，还有许多评判划分质量的其他准则。传统的划分方法可以扩展到子空间聚类，而不是搜索整个数据空间，当存在很多属性并且数据稀疏时是有效的。为了达到全局最优，基于划分的聚类可能需要穷举所有可能的划分，计算量极大。实际上，大多

数应用都采用了流行的启发式方法，如 k-均值和 k-中心算法，渐近的提高聚类质量，逼近局部最优解。这些启发式聚类方法很适合发现中小规模的数据库中小规模的数据库中的球状簇。使用这个基本思想的算法有：K-MEANS 算法、K-MEDOIDS 算法、CLARANS 算法。

——层次法（hierarchical methods），这种方法对给定的数据集进行层次似的分解，直到某种条件满足为止。具体又可分为“自底向上”和“自顶向下”两种方案。例如，在“自底向上”方案中，初始时每一个数据纪录都组成一个单独的组，在接下来的迭代中，它把那些相互邻近的组合并成一个组，直到所有的记录组成一个分组或者某个条件满足为止。层次聚类方法可以是基于距离的或基于密度或连通性的。层次聚类方法的一些扩展也考虑了子空间聚类。层次方法的缺陷在于，一旦一个步骤（合并或分裂）完成，它就不能被撤销。这个严格规定是有用的，因为不用担心不同选择的组合数目，它将产生较小的计算开销，然而这种方法不能更正错误的决定。代表算法有 BIRCH 算法、CURE 算法、CHAMELEON 算法等。

——基于密度的方法（density-based methods），基于密度的方法与其它方法的一个根本区别在于它并非基于各种各样的距离，而是基于密度，这样就克服了基于距离的算法只能发现“类圆形”的聚类的缺点。这个方法的指导思想是，只要一个区域中的点的密度大过某个阈值，就把它加到与之相近的聚类中去。代表算法有 DBSCAN 算法、OPTICS 算法、DENCLUE 算法等。

——基于网格的方法（grid-based methods），首先将数据空间划分成为有限个单元（cell）的网格结构，所有的处理都是以单个的单元为对象的。这种处理方法的一个突出的优点就是处理速度很快，通常与目标数据库中记录的个数无关，它只与把数据空间分为多少个单元有关。代表算法有 STING 算法、CLIQUE 算法、WAVE-CLUSTER 算法。

4.2 决策树和随机森林

回归投资人杠杆影响因素和对投资人极端亏损进行预测时需要分别使用回归模型和分类模型。回归分析中影响因素包括个人信息类（如自然人、法人、年

龄、地区等）、个人资金类特征（如可用资金量大小等）、操作统计类特征（如持仓时间、做多做空频次、交易种数等）以及操作行为类特征（如动量交易类型和逆向交易类型、余额不足时习惯于注资还是平仓等）。

在分析杠杆影响因素和投资人风险中，如果需要了解想要控制投资人杠杆最重要的影响因素，因果分析就显得十分重要。[42]提到能够预测因变量的特征可能有很大的显著性，但显著性只能表明关联关系而不能代表因果关系。因此除使用线性回归、决策树、随机森林等可解释的模型外，还需要尝试能够体现因果关系的其他方法。

（1）决策树

决策树是一种基本的分类和回归方法，分类决策树如 ID3、C4.5 算法[35,36]等，回归决策树如 CART 算法[37]等。在回归分析杠杆影响因素中，需要使用回归决策树。CART 本身即可用于分类也可用于回归，它假设决策树是二叉树，对回归树用平方误差最小化准则，对分类树用基尼指数（Gini index）最小化准则，由生成和剪枝两步组成。在回归树生成过程中，采用启发式的方法选择变量和它的取值，作为切分变量和切分点，并寻找最优的切分变量和切分点递归地生成最小二乘树。

决策树是随机森林的基础，用于分类的决策树中，ID3 算法核心是在决策树各个结点上使用信息增益选择特征，C4.5 算法对 ID3 算法进行了改进，使用信息增益比来选择特征。对于每个随机变量 X，定义熵为：

$$H(X) = -\sum_{i=1}^{n} p_i \log p_i$$

熵越大表示随机变量的不确定性越大；定义随机变量 X 给定条件下随机变量 Y 的不确定性由可以条件熵表示：

$$H(Y \mid X) = \sum_{i=1}^{n} p_i H(Y \mid X = x_i)$$

信息增益的定义为：特征 A 对训练数据集 D 的信息增益 $g(D, A)$，定义为集合 D 的经验熵 $H(D)$ 与特征 A 给定条件下 D 的经验条件熵 $H(D \mid A)$ 之差，即：

$$g(D,A)=H(D)-H(D\mid A)$$

信息增益比定义为：特征 A 对训练数据集 D 的信息增益比 $g_R(D,A)$ 定义为其信息增益 $g(D,A)$ 与训练数据集 D 的经验熵 $H(D)$ 之比：

$$g_R(D,A)=\frac{g(D,A)}{H(D)}$$

CART 分类树采用基尼指数选择最优特征，同时决定该特征的最优二值切分点。假设有 K 个分类，样本点属于第 k 类的概率为 p_k，则概率分布的基尼指数定义为：

$$Gini(p)=\sum_{k=1}^{K}p_k(1-p_k)=1-\sum_{k=1}^{K}p_k^2$$

在特征 A 条件下，集合 D 的基尼指数定义如下，表示集合 D 经 $A=a$ 分割后集合 D 的不确定性。

$$Gini(D,A)=\frac{|D_1|}{|D|}Gini(D_1)+\frac{|D_2|}{D}Gini(D_2)$$

（2）随机森林（Random Forest, RF）

鉴于决策树容易产生过拟合，随机森林采用多个决策树投票机制进行改善[39]。RF 生成首先从样本集中采用 Bootstrap 方式产生 N 个样本，然后选取 A 个特征中的 k 个构建决策树，重复 M 次产生 M 棵决策树后，采用多数投票机制进行预测。

随机森林的主要思想是每一棵决策树是精通于某一个窄领域的专家（因为从 A 个特征中选择 k 让每一棵决策树进行学习），这样在随机森林中就有了很多个精通不同领域的专家，对一个新的问题（新的输入数据），可以用不同的角度去看待它，最终由各个专家，投票得到结果。

随机森林的优点在于能够处理很高维度的数据并且不用做特征选择，在训练完成后它也能给出不同特征的重要性，并且容易理解，具有很好的抗噪声能力，可以进行并行处理。

4.3 逻辑回归和深度学习

对于投资人极端亏损的预测分析需要使用分类算法，如分类决策树、随机森林、逻辑回归、神经网络等模型。逻辑回归是银行风控领域的经典算法，随机森林是一种集成学习算法，深度学习（NN）模型是包含多隐层的多层感知器系统，通过应用综合复杂结构和多重非线性变换构成的多个处理层及对数据进行高层抽象的一系列算法，建立具有数个隐层的多层感知网络并实现各种模式的识别和认知。不同资金量的投资人风险水平可能相同，但产生亏损的代价不一定等价，因此考虑采用将代价引入损失函数等方式优化模型，使分类模型能够对公司更关心的投资人预测准确度相对高。

逻辑回归（Logistic Regression）是一种分类方法，主要用于二分类问题，使用 Sigmoid 函数，其形式为：

$$g(z)=\frac{1}{1+e^{-z}}$$

2006 年，Hinton 提出了深层网络训练中梯度消失问题的解决方案：无监督与训练对权值进行初始化+有监督训练微调。其主要思想是先通过自学习的方法学习到训练数据的结构（自动编码器），然后在该结构上进行有监督训练微调。但是由于没有特别有效的实验验证，该论文并没有引起重视。2011 年，ReLU 激活函数被提出，该激活函数能够有效的抑制梯度消失问题。同年，微软首次将 DL 应用在语音识别上，取得重大突破。神经网络在分类问题如图像、NLP 领域取得了越来越多的成果，其激活函数通常采用 Sigmoid 函数、tanh 函数、ReLU 函数等，输出层常用 Sigmoid 或 Softmax 层，训练算法主要采用梯度下降法如 BP 算法[40]。常见的神经网络模型有深度置信网络（Deep Belief Network, DBN）、深度玻尔兹曼机（Deep Boltzmann Machine, DBM）、含有卷积层和池化层的卷积神经网络（Convolutional Neural Network, CNN）、循环神经网络（Recurrent Neural Networks, RNN）、长短期记忆网络（Long Short-Term Memory, LSTM）等。深度学习应用领域跨度较大，如图像识别主流算法 CNN，语音识别和机器翻译，文本模型、时序数据处理以及推荐系统中都有广泛的应用。对极端亏损情况的分类尝试使用合适的网络结构进行分析预测。

4.4 样本不均衡处理

在整体极端亏损的样本中，有意义的整体样本数量为 1731748 条，极端亏损分为两个级别，较低级别样本数量为 103057 条（5.95%），较高级别样本数量为 45709 条（2.64%），极端亏损总样本量占 8.59%，即正负样本比约为 91：9，样本数量不平衡，因此需要一定的不均衡样本处理方法。

解决方法主要为两个方面，一是从数据的角度出发，主要采用抽样的策略使数据相对均衡；二是从算法的角度出发，考虑不同误分类情况代价的差异对算法进行优化[41]。采样方法分为随机采样、过采样和欠采样，SMOTE（Synthetic Minority Oversampling Technique）是用合成少数类的过采样技术，根据少数类样本人工合成新样本添加到数据集中；欠采样算法主要有 EasyEnsemble 算法和 BalanceCascade 算法，其中 EasyEnsemble 算法类似于随机森林的 Bagging 方法，N 次有放回的抽样生成 N 份子集，少数类样本分别与其合并生成 N 份样本，得到 N 个模型的平均值作为预测结果。代价敏感学习算法（Cost-sensitive Learning）主要从算法层面上解决不平衡数据学习，其核心要素是代价矩阵，实际应用中不同类型的误分类情况导致的代价是不同的。实现从模型的角度有代价敏感版本的决策树、SVM、神经网络等；从贝叶斯理论出发可以把代价敏感学习看作分类结果的后处理，调整损失函数为 $H(x)=\arg\min_{i}(\sum P(j\mid x)c(i,j))$，优点在于不依赖具体的分类器，但缺点在于它要求分类器的输出值为概率。

在模型评价方面，对于不平衡样本采用准确率不能达到预期效果，需要结合 Precision、recall 等指标对预测效果进行评价，如 F1-Score、AUC 等。

5 总结与展望

随着机器学习、深度学习技术的飞速发展，金融服务产业模式不断革新，传统的经济学实证方法依赖于人工操作软件计算，不能同时关注大量数据和及时更新。金融公司需要使用先进的机器学习算法，以数据价值为驱动建立智能化的风险预测模型，以此作为风险防范的有效手段。

针对期货公司投资人运用杠杆和市场波动带来的风险问题，应用机器学习进行风险分析及预警系统。系统包括基础数据库、风险分析模型和预警模块。其中基础数据库基于期货公司交易数据以及爬虫抓取的期货交易所各合约结算价格

建立，为建模过程提供持续输入，并且存储模型的中间输出结果；风险分析模块包括账户资金余额全部亏损概率分析、投资人杠杆影响因素分析和投资人极端亏损预测；预警模块展示分析得到的投资人风险聚类结果，为公司决策提供支持。除采用决策树、随机森林、神经网络等机器学习模型外，风险分析模块也采用了计量经济学模型如 Var、GARCH 等，满足模型经济学理论坚实的同时提高性能。国内外研究使用 SVM、逻辑回归、LSTM、NLP 等多种机器学习技术，在价格指数、波动率等方面预测取得了众多研究成果，在交易策略方面有基于 ARIMA 等模型的相关实践，在风险分析方面经济学领域使用了一些实证方法，计算机领域在如金融欺诈、自然灾害等风险预测中也有突出的表现，但针对期货公司投资人风险进行分析评估的相关问题还有待研究。本文将基于机器学习相关技术对期货公司投资人特征进行分析，对其杠杆利用因素以及账户风险进行研究。

6 主要参考文献

[1] Li L, Leng S, Yang J, et al. Stock Market Autoregressive Dynamics: A Multinational Comparative Study with Quantile Regression[J]. Mathematical Problems in Engineering, 2016: 1-15.

[2] Xie H, Bian J, Wang M, et al. IS TECHNICAL ANALYSIS INFORMATIVE IN UK STOCK MARKET? EVIDENCE FROM DECOMPOSITION-BASED VECTOR AUTOREGRESSIVE (DVAR) MODEL ∗[J]. Journal of Systems Science & Complexity, 2014, 27(1): 144-156.

[3] Nayak R K, Mishra D, Rath A K, et al. A Naïve SVM-KNN based stock market trend reversal analysis for Indian benchmark indices[J]. soft computing, 2015: 670-680.

[4] Kavitha G, Udhayakumar A, Nagarajan D, et al. Stock Market Trend Analysis Using Hidden Markov Models[J]. arXiv: Statistical Finance, 2013.

[5] Adebiyi A A, Adewumi A O, Ayo C K, et al. Comparison of ARIMA and Artificial Neural Networks Models for Stock Price Prediction[J]. Journal of Applied Mathematics, 2014: 1-7.

[6] Ticknor J L. A Bayesian regularized artificial neural network for stock market forecasting[J]. Expert Systems With Applications, 2013, 40(14): 5501-5506.

[7] Patel J, Shah S, Thakkar P, et al. Predicting stock market index using fusion of machine learning techniques[J]. Expert Systems With Applications, 2015, 42(4): 2162-2172.

[8] Hochreiter S, Schmidhuber J. Long Short-Term Memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[9] Rather A M, Agarwal A, Sastry V N, et al. Recurrent neural network and a hybrid model for prediction of stock returns[J]. Expert Systems With Applications, 2015, 42(6): 3234-3241.

[10] Akita R, Yoshihara A, Matsubara T, et al. Deep learning for stock prediction using numerical and textual information[C]. international conference on information systems, 2016: 1-6.

[11] Qiyuan Gao. 2016. Stock market forecasting using recurrent neural network. Ph.D. Dissertation. University of Missouri-Columbia.

[12] Zhang L, Aggarwal C, Qi G J. Stock Price Prediction via Discovering Multi-Frequency Trading Patterns[C]. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2017:2141-2149.

[13] Ding X, Zhang Y, Liu T, et al. Deep learning for event-driven stock prediction[C]. international joint conference on artificial intelligence, 2015: 2327-2333.

[15] Ghasemiyeh R, Sana S S, Sana S S. A Hybrid Artificial Neural Network with Metaheuristic Algorithms for Predicting Stock Price[M]. Taylor & Francis, Inc. 2017.

[16] Nair B B, Kumar P K S, Sakthivel N R, et al. Clustering stock price time series data to generate stock trading recommendations: An empirical study[J]. Expert Systems with Applications, 2017, 70:20-36.

[17] Rouwenhorst K G, Tang K. Commodity Investing[J]. Social Science Electronic Publishing, 2012, 4(1):447-467.

[18] Grinblatt M, Keloharju M. What Makes Investors Trade[J]. Journal of Finance, 2001, 56(2): 589-616.

[19] Bailey W, Kumar A, Ng D T, et al. Behavioral Biases of Mutual Fund Investors[J]. Journal of Financial Economics, 2011, 102(1): 1-27.

[20] Barber B M, Odean T, Zhu N. Systematic noise[J]. Journal of Financial Markets, 2009, 12(4):547-569.

[21] Nicolosi G, Peng L, Zhu N, et al. Do Individual Investors Learn from Their Trading Experience[J]. Ssrn Electronic Journal, 2004.

[22] Kim J, Chung K. Emerging risk forecast system using associative index mining analysis[J]. Cluster Computing, 2017, 20(1): 547-558.

[23] Neshat M, Tabatabi M, Zahmati E, et al. A hybrid fuzzy knowledge-based system for forest fire risk forecasting[J]. International Journal of Reasoning-based Intelligent Systems, 2016, 8(3/4):132.

[24] ISABEL MARÍA DEL ÁGUILA, JOSÉ DEL SAGRADO. REQUIREMENT RISK LEVEL FORECAST USING BAYESIAN NETWORKS CLASSIFIERS[J]. International Journal of Software Engineering & Knowledge Engineering, 2011, 21(02):167-190.

[25] Linsmeier T J, Pearson N D. Value at Risk[J]. Financial Analysts Journal, 2000, 56(2):47-67.

[26] Engle R F. Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation[J]. Econometrica, 1982, 50(4): 987-1007.

[27] Bollerslev T. Generalized autoregressive conditional heteroskedasticity[J]. Journal of Econometrics, 1986, 31(3): 307-327.

[28] Nelson D B. CONDITIONAL HETEROSKEDASTICITY IN ASSET RETURNS: A NEW APPROACH[J]. Econometrica, 1991, 59(2): 347-370.

[29] Tam K Y, Kiang M Y. Predicting bank failures: A neural network approach[J]. Applied Artificial Intelligence, 1990, 4(4): 265-282.

[30] Markham I S, Ragsdale C T. Combining Neural Networks and Statistical Predictions to Solve the Classification Problem in Discriminant Analysis[J]. Decision Sciences, 1995, 26(2): 229-242.

[31] Fletcher D, Goss E. Forecasting with neural networks : An application using bankruptcy data[J]. Information & Management, 1993, 24(3):159-167.

[32] Wilson R L, Sharda R. Bankruptcy prediction using neural networks[J]. decision support systems, 1994, 11(5): 545-557.

[33] 王春峰, 万海晖, 张维. 基于神经网络技术的商业银行信用风险评估[J]. 系统工程理论与实践, 1999, 19(9):24-33.

[34] 臧玉卫, 王萍, 吴育华. 贝叶斯网络在股指期货风险预警中的应用[J]. 科学学与科学技术管理, 2003, 24(10):122-125.

[35]Quinlan J R. Introduction of decision trees[J]. 1986(1):81-106.

[36]Quinlan J R. C4.5: programs for machine learning[J]. 1993, 1.

[37]Breiman L, Friedman J H, Olshen R, et al. Classification and Regression Trees[J]. Encyclopedia of Ecology, 2008, 40(3):582-588.

[38]Criminisi A, Shotton J, Konukoglu E. Decision Forests: A Unified Framework for Classification, Regression, Density Estimation, Manifold Learning and Semi-Supervised Learning[M]. Now Publishers Inc. 2012.

[39]Breiman L. Random Forests[J]. Machine Learning, 2001, 45(1):5-32.

[40]Lecun Y, Bengio Y, Hinton G. Deep learning[J]. Nature, 2015, 521(7553):436.

[41]He H, Garcia E A. Learning from Imbalanced Data[J]. IEEE Transactions on Knowledge & Data Engineering, 2009, 21(9):1263-1284.

[42]Kuang K, Cui P, Li B, et al. Estimating Treatment Effect in the Wild via Differentiated Confounder Balancing[C]. The, ACM SIGKDD International

Conference. ACM, 2017:265-274.