白皮书

机器学习和大数据在量化投资中的应用

揭示金融和另类数据中的模式

MathWorks®

概要

机器学习使计算机或机器可直接从数据中学习，而无需通过编程来构建输入数据和输出的关系。从本质上讲，机器学习模型在挖掘数据背后的非线性关系上要强于传统模型。您需要使用现代工具协同处理大数据和机器学习，才能将藏匿于大型数据集中的有价值信息提取出来。

在投资中，您能借助机器学习和大数据来做些什么？

- 资产配置和优化
- 应用自然语言处理来进行投资情绪分析
- 异常值和欺诈检测
- 金融预测和价格预测

所有这些应用都采用新颖的方式，使用机器学习来应对投资中遇到的常见挑战。然而，成功应用机器学习技术需要数据科学技能，这是一项特殊技能，人才稀缺，但需求量极大。

就数据科学技能人才的缺口而言，2018 年 8 月的《领英劳动力报告》指出，美国数据科学家的短缺已达到 151,717 人，并且蔓延到了金融和科技以外的其它行业。[1]

同时这也带来一个机遇，可在传统 IT 系统或云端部署机器学习模型，通过实时分析来支持决策以及实现决策自动化。

本白皮书展示了如何应用机器学习和大数据技术来解决投资问题并提高投资绩效。

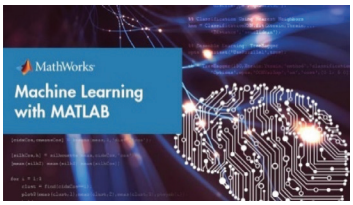
什么是机器学习？

*机器学习*技术使计算机或机器无需大量编程即可直接从数据中学习。在金融领域，机器学习提供了一套新的方法来开发预测模型。其与目前传统的模型相比，它能够更好地处理非线性关系的数据。

机器学习在业内流行着多种分类法。一般来说，根据我们试图解决的问题类型，对机器学习进行分类。例如：

- **监督式学习：从标记数据中学习。**在监督式机器学习中，结果表现为对数据标注正确（或所需）的响应。监督式学习使用分类和回归技术开发预测模型。
- **无监督学习：从未标记的数据中发现模式。**在无监督机器学习中，结果表现为不对数据标注响应。聚类是一种最常用的无监督学习技术。这种技术可通过探索性数据分析发现数据中隐藏的模式或分组。
- **强化学习：学习行为或动作。**强化学习的目的是建立一个模型，该模型能够执行一系列动作以最大化累积奖励。强化学习不是使用已知的输入和输出集，而是优化相对于奖励函数的动作。从根本上说，强化学习就像试错一样，智能体根据其动作从积极和消极的奖励中学习。

[1] *https://news.linkedin.com/2018/8/linkedin-workforce-report-august-2018*



了解更多

» ***使用 MATLAB 进行机器学习*** - 电子书

如何确定使用哪种算法?

常用的机器学习算法有几十种，而且每一种都用不同的学习方法。因此选择正确的算法看似一项艰巨的任务（图 1）。

根本就没有一种最佳方法或万全之策。试错是找到正确算法过程的一部分。— 即使是经验丰富的数据科学家，也很难说出某种算法是否无需试错即可使用。在其他因素中，算法的选择取决于您要处理的数据的大小和类型，以及您要从数据中了解的信息。

SUPERVISED LEARNING		UNSUPERVISED LEARNING	REINFORCEMENT LEARNING
CLASSIFICATION	REGRESSION		
Logistic Regression	Linear Regression, GLM	K-Means, K-Medoids Fuzzy C-Means	SARSA[2]
Random Forests	Gaussian Process Regression	Hierarchical	Policy Gradient
Discriminant Analysis	Boosted Decision Trees	Gaussian Mixture	DDPG[3]
Naive Bayes	Decision Trees	Hidden Markov Model	Actor Critic
Nearest Neighbors	Support Vector Machines	DBSCAN[1]	Q-Learning
Boosted Decision Trees	Neural Networks and Deep Learning	Spectral Clustering	Neural Networks and Deep Learning
Random Subspace		Neural Networks and Deep Learning	
Support Vector Machines			
Error-Correcting Output Code			
Neural Networks and Deep Learning			

1. DBSCAN = Density-Based Spatial Clustering of Applications with Noise
2. SARSA = State–action–reward–state–action
3. DDPG = Deep Deterministic Policy Gradient

图 1.常用机器学习算法。

了解更多

» ***精通机器学习：MATLAB 分步实施指南*** - 电子书

什么是大数据？

现有的技术需要在内存中进行运算。大数据通常是指很难运用现有技术进行处理的大批量数据。金融机构使用或存储的大数据包括：

- 过去 10 年中交易的 1000 多支证券的历史记录数据
- 数十亿笔信用卡交易
- 与过去 10 年中所有证券相关的新闻数据
- 1000 名客服人员的通话记录

值得注意的是当数据大于内存容量时，分组计算平均值和标准差的算法，其复杂程度远超数据小于内存容量的情况。

必须利用现代工具来处理大数据以及应用机器学习在数据中查找模式，才能提取藏匿于大数据中的有价值信息。

数据越大，见解越多，效果越好

金融业中最基本的数据应用示例是技术分析，通过分析价格、交易量和时间之间的关系，作为预测未来价格变动的技术指标。

随着市场越来越高效，我们可以看到使用技术指标的优势在减弱，特别是对于收盘数据等常见的数据。投资者不会单独使用收盘数据，而是将短期数据（如盘中数据和每笔成交价格及时间数据）整合到投资分析中。此外，投资者利用许多不同类型的数据（如新闻、社交媒体帖子、卫星图像和销售点数据）提升投资业绩。这些数据的共同特征是数据集体量大、不便于及时处理。

大数据速览

数值型数据是金融业最常用的数据类型，其次是文本型数据。虽然在有些情况下也会使用成像数据（例如，卫星图像），但这类数据在金融业中的应用并不广泛。

大数据处理的首要挑战是确定如何访问大型数据集，因为数据集形式各异且存储在各类系统中：

- **文件：** 许多数据集由大量的中小型文件组成。文件数量会迅速增长，而且文件通常无法存入一台计算机的内存中。这些文件通常位于共享驱动器上的一个或多个目录中，并且可能包含分隔文本、电子表格、图像、视频和各种专有格式。
- **数据库：** 可以使用多种类型的数据库来存储和管理金融业中的大型数据集，包括关系数据库、图形数据库和文档数据库。
- **Hadoop：** Hadoop® 是一个基于分布式计算和存储原理来存储和处理大数据集的系统。它包含两个主要的子系统，而这两个子系统共存于一个计算机集群上（图 2）：
  - Hadoop 分布式文件系统 (HDFS)：一个大型的抗故障文件系统
  - YARN：应用程序调度框架，管理在 Hadoop 上运行的应用程序，包括批处理框架（如 MapReduce 和 Spark™）和 SQL 接口（如 Hive 和 Impala）

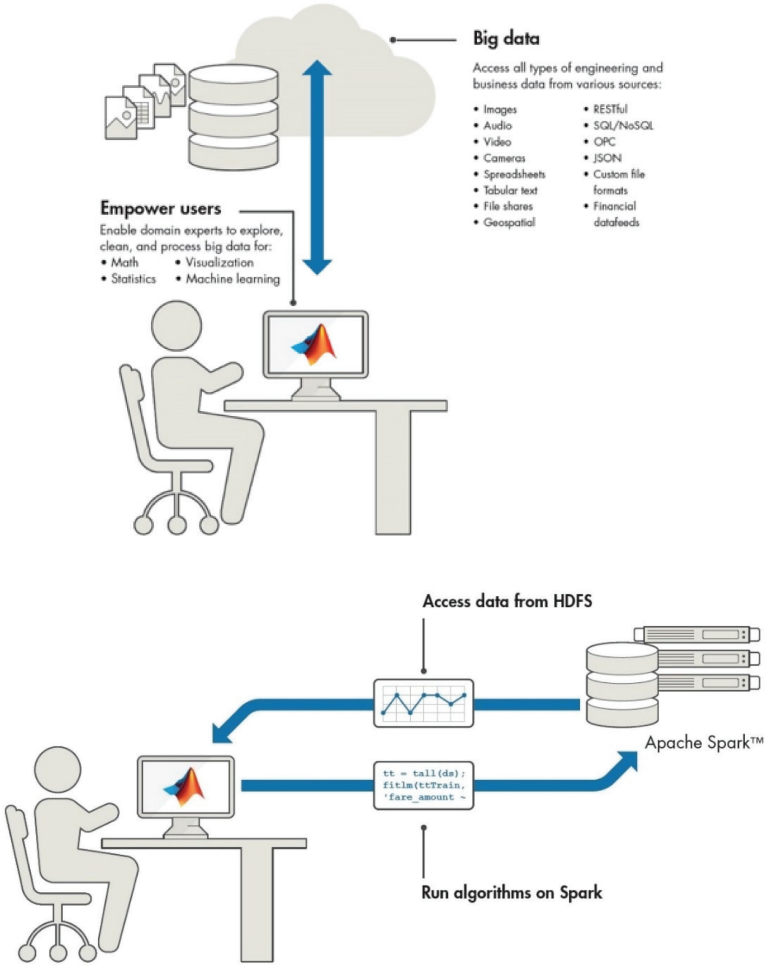


图 2.在 HDFS 和 Spark 中使用 MATLAB 处理数据。

在投资中，您能借助机器学习和大数据来做些什么？

遇到的复杂任务或问题中涉及大量数据和变量，但没有现成的处理公式或方程式，这时可以考虑使用机器学习和大数据技术。例如，以下情况适合采用机器学习和大数据技术：

- 数据的性质是非结构化的（例如，文本、图像、音频或视频的组合）。
- 需要快速响应大量数据或高速数据，比如交易执行过程中的数据。
- 专家知识、手写规则和方程式对于模型而言太过复杂，比如针对新闻的投资情绪分析。
- 数据本身在不断变化，程序也必须适应这种变化，比如资产配置（图 3）、自动交易、能量需求预测和价格趋势预测等。

» ***资产配置 - 分层风险平价（代码示例）***

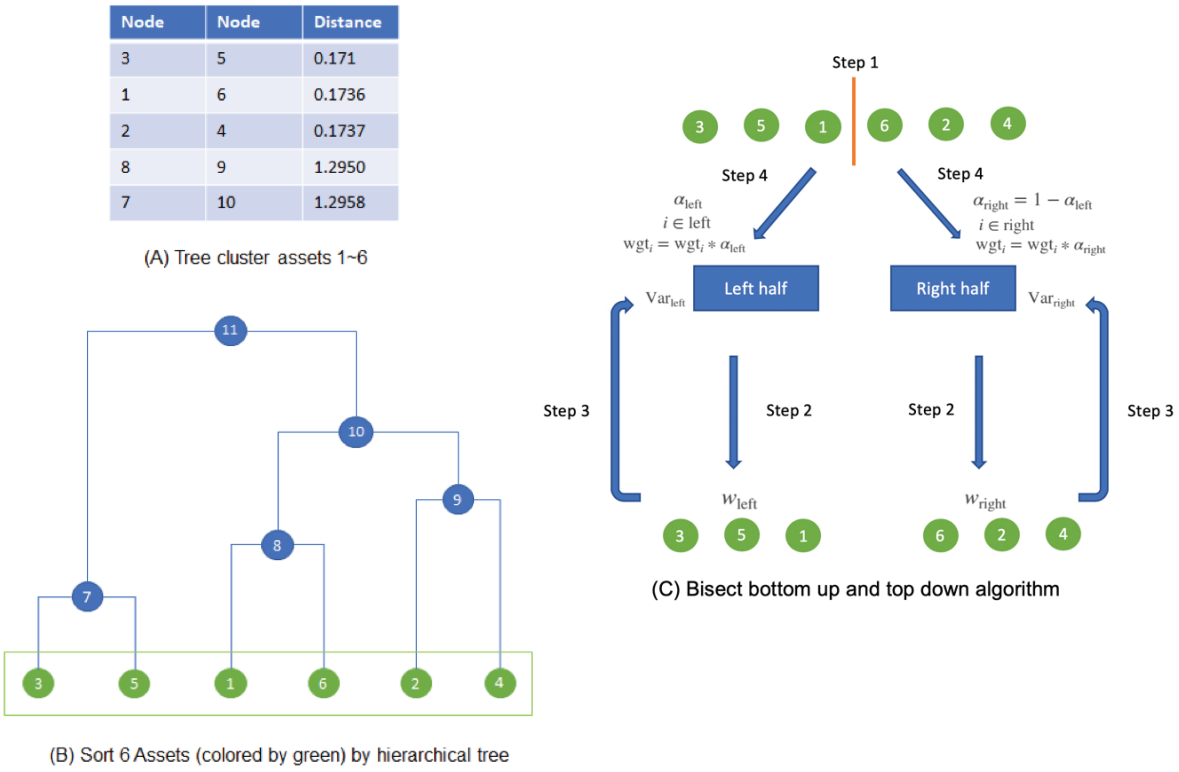


图 3.使用 MATLAB 分层风险平价分析 (HRP) 的资产配置工作流程

在投资方面，机器学习比人类更高效吗？

尽管投资团队越来越多地采用机器学习来进行投资和大数据处理，但鉴于基于大型数据集进行机器学习的计算特性，其主要用途通常是量化投资分组。在投资组合管理中，机器学习主要应用于识别交易信号，或利用大量数据来针对价格变动创建交易指标。实际上，对于新闻趋势、社交媒体数据或卫星图像，这类数据量不断增加，而所需的反应时间在一毫秒乃至一微秒内。因此此类大数据的产生，迅速推动着投资团队对机器学习的需求，进而对数据进行统计推断，并创建预测模型。

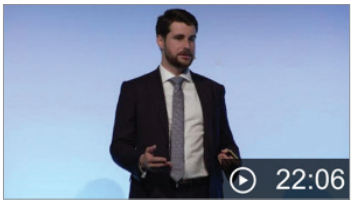
然而在某些领域如预测长期趋势或政权更迭，人类的判断分析依然强于机器预测。人类分析师将继续发挥关键作用。

哪些技术最有效呢？

表 1 对比基本面分析和量化方法（两者都由分析师进行）与机器学习和大数据方法的有效性。

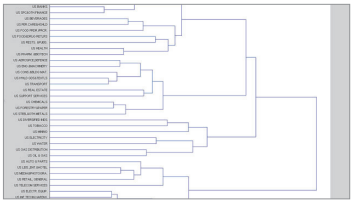
技术是否有效：	基本面分析	传统量化方法	机器学习和 大数据技术
一定时间段内的预测			
长期趋势	是	否	否
短期趋势	否	是	是
当日趋势（高频数据）	否	可能	是
不同数据类型的预测			
结构化数据	是	是	是
非结构化数据	否	否	是
少量数据	是	是	否
大量数据	否	否	是
模型解释难易程度	容易	适中	困难

机器学习示例



Aberdeen Standard ***的资产配置、机器学习和高性能计算*** - 视频

了解 Aberdeen 如何使用机器学习来分析金融市场趋势并推动创新的资产配置策略。



Banque Cantonale Vaudoise ***提高金融分析工作速度*** - 用户案例

了解 BCV Asset Management 如何使用聚类模型对工业指数进行建模。

从量化分析师转型为数据科学家

谁将进行分析并构建方法论? 要成功创建机器学习模型，您需要具备以下三类知识的人才：

- **机器学习:** 了解如何应用统计分析和机器学习技术来解决问题。
- **计算:** 掌握关于编程、数据管理和计算基础设施的基本知识。
- **专业领域知识:** 需要不断发展进步以了解项目背后的金融建模。

数据科学家通常具有机器学习和计算方面的综合知识，但可能不具备专业领域知识。具备这三项技能的人才凤毛麟角，而对具备这些技能的人才争夺异常激烈，导致招募此类人员难度颇大。几乎每周都有文章报道数据科学家短缺，而这已经是一个公认的全球性问题。

在金融服务行业，量化分析师在金融建模中发挥着不可或缺的作用。与数据科学家所需的知识相比，量化分析师已经具备了金融和计算技能领域的专业知识。他们需要补充的只是机器学习知识。

许多金融机构并未竞相聘请可能不具备相应领域专业知识的数据科学家，而是招募和培训了量化分析师来从事数据科学工作。他们可以快速了解机器学习的工作原理，并应用这些技术来解决金融问题。

通过 MATLAB 进行机器学习和处理大数据

使用 MATLAB®，团队可以尝试更多的想法，与其他替代方案（例如，Python® 和 R）相比，可以在更短的时间内获得更有效的成果。

为了最大限度地提高效率，请将团队的工作重点放在以下四个步骤上：

1. 访问和探查大数据

MATLAB 旨在处理大量数据和各种数据类型，包括数字、文本、新闻和社交媒体数据。

“我们之前的系统非常繁琐，而且我们的数据集非常庞大，如果不借助 MATLAB 及其处理大数据的能力以及直接与 Bloomberg 和我们的数据库进行交互的能力，我认为几乎不可能完成我们的工作。”

— Ananthi Jegan, Olam CFSG

2. 预处理大数据

MATLAB 提供的工具可以轻松地对大型数据集进行快速处理。此外，其数值运算可直接扩展到集群和云端进行并行处理。

“我具有金融领域的专业知识，而不是编程。如果要对大量数据进行复杂分析，我需要一个易于使用且包括许多所需功能的软件。有了 MATLAB 后，我能在一个环境中完成所有工作，这确实非常有用。”

— Omid Rezania, CalPERS

3. 运用机器学习算法分析大数据

积累的大数据包含复杂的信息。需要一个准确的预测模型来处理变量间的非线性关系，才能从中提取出有效见解。您可以快速选择和识别模型的正确特征，然后迭代其他模型以确定最佳预测算法。

MATLAB 可以通过专业的工具箱、预置函数以及一整套的数据统计和机器学习功能来快速实现这一过程。它还提供了如非线性优化、系统识别以及数千种预置算法等一系列先进的方法，用于金融建模、投资组合优化和风险管理（图 4）。

“在我们使用 MATLAB 之前，我们无法在合理的时间内生成聚类模型。我们根本不会这样做。MATLAB 为我们开辟了新的视野。”

— Pierre-Yves Boillat, Banque Cantonale Vaudoise

4. 部署机器学习模型

MATLAB 是一个集成系统，可让您在生产系统中部署机器学习模型并与数据提供商和现有 IT 基础设施进行集成。此外，MATLAB 还提供集成到企业系统、集群和云端的在线和实时部署。

“借助 MATLAB、MATLAB Production Server 和 MathWorks 培训服务，我们风险团队中拥有 C++ 或 Java® 编程经验的人员能够高效地开发用于金融分析的核心库，然后将其部署为 Web 应用程序，用在企业环境下的生产系统中。”

— Marcus Veltum, Helaba Invest

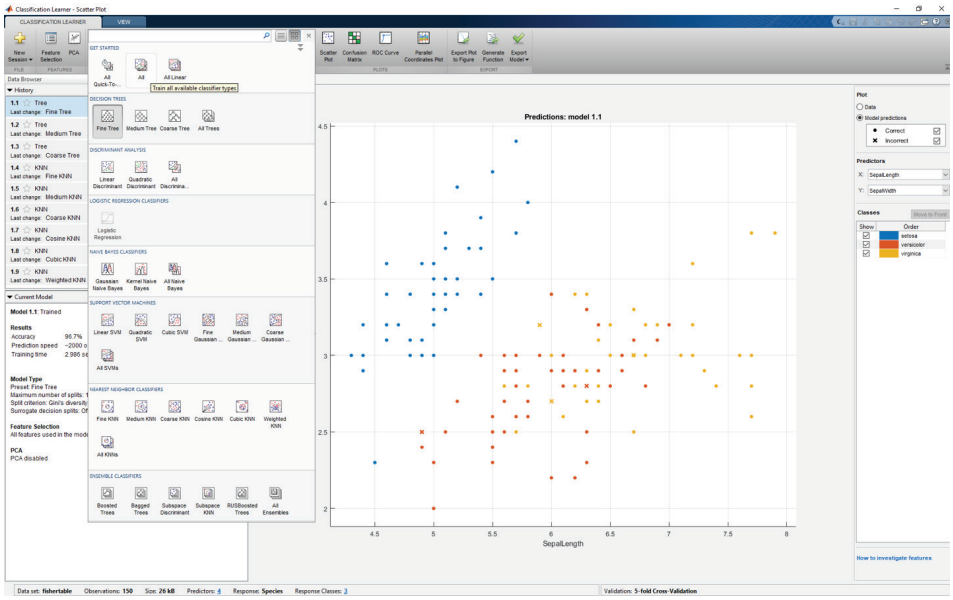


图 4.Classification Learner 应用程序，能够以交互方式训练、验证和调整分类模型。

结束语

借助机器学习和大数据，投资经理能够根据大数据提炼出信息，做出以往传统模型无法实现的预测，进而做出明智的决策。新的应用程序和投资见解交付给最终客户的时间也大大缩短。

MATLAB 提供了一个交互式环境以及预置函数和库，使量化分析师能够成为数据科学家并开发定制的机器学习模型。MATLAB有各种灵活的部署方案可供选择，您可以将用于生产的模型快速集成到现有 IT 基础设施中，从而节省时间并避免在向不同的编程环境转换时发生错误。

了解更多

机器学习在量化交易中的应用 (32:55) - 视频

机器学习一点通 (34:34) - 视频

使用 Regression Learner 应用程序预测比特币波动率 (5:56) - 视频

MATLAB 助力量化金融和风险管理 - 免费产品试用