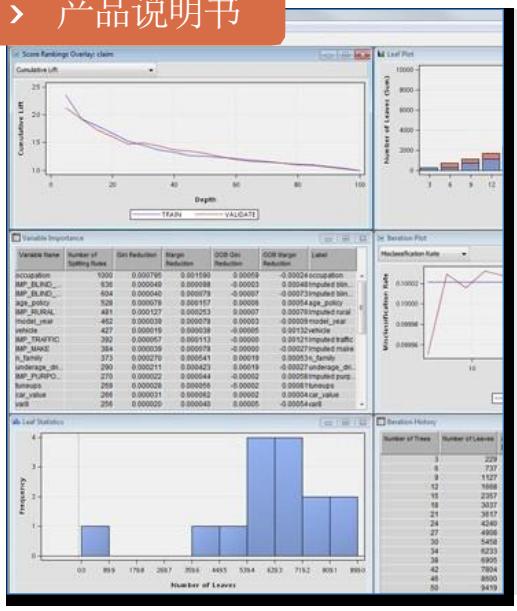
› 产品说明书



SAS®高性能分析可以做什么?

采用 SAS 高性能分析产品，可基于大量多样化数据开发和处理模型。这些产品–用于统计、数据挖掘、文本挖掘、计量经济学和优化–可配置在高度可扩展的分布式内存处理架构中。

SAS®高性能分析为什么重要?

通过分析大数据可以获得更加准确的洞察，及时制定业务决策。您能够解决各种业务难题、验证更多概念并评估复杂的场景，有助于企业把握新机遇，减少不确定因素的影响。

SAS®高性能分析为谁而设计?

这些产品专门为需要快速有效开发和处理模型的分析人员而设计 (如数据挖掘人员、统计师、数据科学家和业务分析师)。同时，为 IT 人员提供高度可扩展的可靠架构，便于管理和处理分析任务。

SAS®高性能分析产品

更快速地生成更准确的洞察，解决最复杂的业务问题

解决复杂的业务问题需要先进的高端分析技术，同时需要能够整合大数据源，包括庞大的文本数据。

SAS 提供五种高性能分析产品，基于分布式内存环境进行分析运算。您可以处理前所未有的大量数据，快速准备和探索数据、针对多种场景进行建模，以近实时方式 (一般几分钟，而非几小时) 快速生成准确的洞察结果。

如果能够将分析工作的处理时间从几天或几小时，缩短到几分钟甚至几秒钟，您可以提出更多假设。迅速调优模型，重新进行分析。

在分析中同时使用结构化和非结构化数据，利用更多变量，进行多次模型迭代–比以往更快–可以显著提高预测能力。

主要优点

- **更快更自信地把握新机遇，探测未知风险，做出正确选择。**SAS高性能分析产品可以利用所有可用计算资源，更快速地进行统计建模和模型筛选，无论单机环境，还是分布式计算环境。您可以获得更精细更准确的结果，为企业带来新的业务机会。
- **利用所有数据 (包括非结构化数据)，采用高级建模技术进行更多次模型迭代，帮助解答业务难题。**针对所有数据进行综合分析可以提高洞察精确性，从而制定更好的决策。采用最佳建模技术，增加模型迭代次数了。结合结构化数据与文本数据，揭示先前未发现的内在联系，提高模型预测能力。
- **以前所未有的速度获得洞察，做出高价值且时间紧迫的决策。**缩短分析模型处理时间，快速得出洞察结果，全面提高企业决策能力。SAS 高性能分析产品分析速度极快，可对快速多变的市场环境大量场景进行评估，及时提出合理建议。
- **利用高扩展、高可靠性分析基础架构，基于全部数据检测更多概念，测试多种场景。**分析人员可在不受基础架构限制的条件下，利用内存环境解决最复杂的业务问题。IT 可以有效满足当前及今后的需求，提供更强大的分析处理能力。

sas

THE POWER TO KNOW®

产品概览

SAS® 高性能分析产品可供企业分析大数据，几分钟即可生成更加准确的洞察结果。这些高性能分析产品包括：

- 统计分析
- 数据挖掘
- 文本挖掘
- 计量经济学
- 优化

除每种产品的特定功能之外，五种产品还具有核心通用程序，用来帮助准备和汇总数据。

单机或分布式模式

SAS 高性能分析产品既可以在单个服务器上运行，也可以在分布式计算机集群环境下运行。无论单机模式还是分布式计算环境，所有高性能程序都支持多线程处理，充分利用所有可用内核资源。

单机模式下，高性能程序可根据机器的 CPU (内核) 数目确定并发线程数量。简言之，单机模式意味着客户端采用多线程模式。

高性能程序在分布式模式下运行时，分布式计算环境中的多个节点参与运算。数据分布到集群中的每个机器上，集群的大规模计算能力可用于解决单个大型分析任务。分布式模式下，集群中的多台机器可同时执行分析任务，每台机器可以并发执行多个线程。

单机模式下，高性能建模过程利用单机所有内核实现扩展性。分布式计算环境中，这些过程能够并发访问数据，充分利用所有内核以及大量可用内存。

SAS® 高性能统计

利用 SAS 高性能统计，可比以往更加迅速地建立并运行模型。建模方法包括回归、逻辑回归、广义线性模型、线性混合模型、非线性模型和决策树。这些程序提供模型选择、降维、识别重要变量等常用分析功能。

SAS®高性能数据挖掘

SAS 高性能数据挖掘可采用拖放界面，强大的描述型、预测性机器学习方法，分析大量多样化数据。支持大量建模技术，包括贝叶斯网络、随机森林、支持向量机、神经网络、聚类等，同时提供数据准备、数据探索和模型评分功能。由于能够更加快速地建立并运行更多模型，因此可以提出并解决更多问题，将更多概念应用到数据挖掘过程中。(SAS 高性能数据挖掘包含 SAS 高性能数据统计)

SAS®高性能文本挖掘

利用 SAS 高性能文本挖掘，您可以迅速获得大量非结构化数据的洞察结果，包括大量文档、电子邮件、备注、报告文件、社交媒体等。支持的功能包括文本解析、实体抽取、自动抽取词干、同义词检测、主题发现和奇异值分解 (SVD) 等。文本挖掘结果可用作高性能数据挖掘的输入，提高模型的预测能力。

SAS®高性能计量经济

SAS 高性能计量经济建模方法包括线性回归、一元和二元 logit/probit 模型、随机前沿模型、删/截回归、样本选择模型、事件数模型和损失分布模型。其中的部分方法可用于面板数据。同时，还提供用来模拟分布的工具，包括多元 copula 以及复合分布模型。

SAS®高性能优化

高性能优化适用于解决确定线性、混合整数线性和非线性问题。提供的多起点 (非线性)、分解 (线性、混合整数线性)、调优选项 (混合整数线性) 和全局/局部搜索优化等重要优化选项可并行执行，大大缩短完成整体优化的时间。

主要特点

SAS®高性能分析产品核心功能

高性能数据概括

- 通过一系列并行化程序支持大规模数据探索和汇总。
- 以 SAS 输出数据集形式，极其快速地大规模生成描述统计量。
- 生成均值、最小值、最大值、极差、分散度和集中度等统计量，以及变量基数、汇总和水平等信息。

高性能 DS2

- 提供分布式内存计算环境下，通过 Base SAS 会话并行执行 DS2 代码的工具。
- 可以控制执行节点的并行化水平以及所用节点数量。

高性能数据挖掘数据库

- 创建重要输入数据源汇总统计结果，包括汇总、计数、最小值、最大值、标准差和不对称测量值。

高性能相关分析

- 对含有大量行列的大数据进行相关分析。

高性能抽样

- 执行高性能简单随机抽样或分层抽样。

高性能数据分箱

- 桶式 (等长) 分箱方法。
- Winsorized 分箱方法和 Winsorized 统计。
- 伪分位数分箱方法，类似于分位数分箱。
- 根据所选分箱方法给出映射表。
- 提供基础统计信息表，包括最小值、最大值、均值、伪中位数等。
- 柱状图表显示输出的映射统计结果。
- 伪分位数估计表。
- 根据分箱结果计算证据权重(WOE)和信息值(IV)。

高性能数据补缺

- 用给定值对数值型变量进行高性能补缺。
- 也可以用均值、伪中位数或非缺失值最小值与最大值之间的随机值代替数值型缺失值。

SAS®高性能统计

高性能逻辑回归和模型选择

- 预测二值型、二项和多项输出。
- 提供建模语法，包括 CLASS 和基于效应的 MODEL 语句。
- 提供多种链接函数，支持多项响应变量分类建模，包括有序分类或无序分类。
- 在输出数据集中给出预测值并生成评分代码。

高性能线性回归和模型选择

- 支持广义线性模型和标准参数化分类效应。
- 提供多种模型效应选择方法。
- 提供建模语法，包括 CLASS 和基于效应的 MODEL 语句。
- 支持数据分区，将数据分成训练集、验证集和测试集。
- 支持大量效应选择(成千上万)。

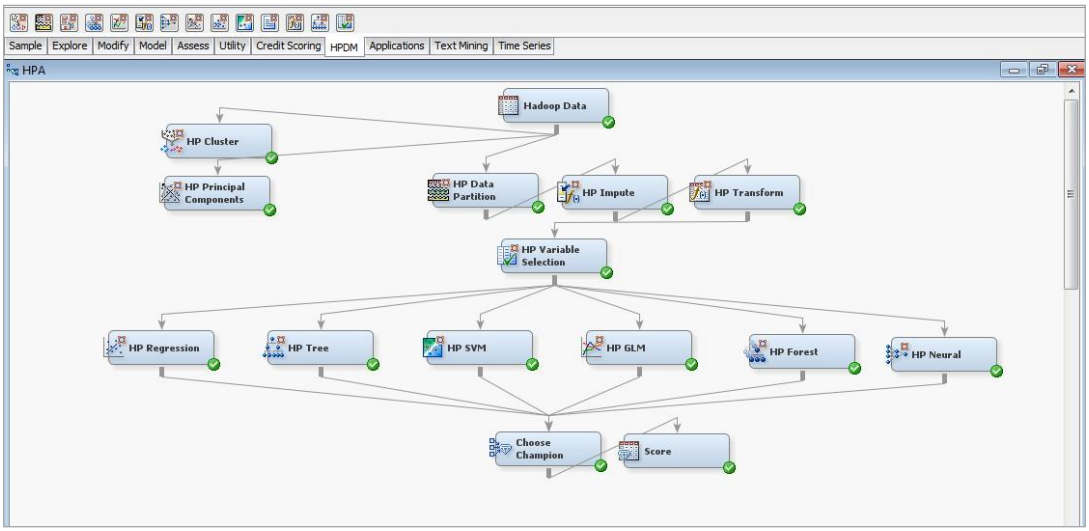


图 1: 采用高性能节点的 SAS 高性能数据挖掘流程图。

主要特点(续)

- 提供基于多种模型评估标准的终止规则。
- 支持基于外部验证和留一交叉验证的终止和选择规则。
- 在输出数据集中给出预测值并生成评分代码。

高性能非线性回归

- 采用最小二乘法和最大似然估计法进行参数估计。
- 提供多种优化技术进行参数估计。
- 基于用户自定义函数计算参数置信度。

高性能混合线性模型

- 多种协方差结构，包括方差分量、复合对称、非结构化、AR(1), Toeplitz 和因子分析。
- 利用各种优化技术进行 REML 和最大似然估计。
- 支持含有大量对象的数据。

高性能偏最小二乘

- 支持广义线性模型和标准参数化分类效应。
- 支持任意阶数交叉效应，包括分类变量和连续变量。
- 支持数据分区，将数据分为训练集和测试集。

高性能分位数回归分析

- 支持单级或多级分位数回归。
- 支持广义线性模型和标准参数化分类效应。
- 支持任意阶数(交叉效应)和嵌套效应。

高性能广义线性模型和模型选择

- 采用最大似然估计法估计估计广义线性模型参数。
- 分析支持模型训练、验证和测试。
- 提供建模语法，包括 CLASS 和基于效应的 MODEL 语句。
- 提供多种链接函数和分布，包括 Tweedie 系列分布。

高性能决策树

- 建立决策树模型。
- 支持区间型和名义型输入变量和目标变量。
- 决策树生长方法可选择熵、Gini、FastCHAID、CHAID、信息增益比率和卡方(名义型目标变量)。
- 回归树生长方法可选择方差、CHAID 和 F 检验(区间型目标变量)。
- 支持决策树的生长和修剪。
- 支持 C4.5 修剪。
- 生成英文规则描述决策树模型结果。

高性能有限混合模型

- 单变量有限混合模型支持最大似然估计。
- 部分模型支持马尔科夫链蒙特卡洛模拟。
- 提供多种内置链接和分布函数。
- 使用混合概率进行分类和回归建模。

高性能主成分分析

- 多元分析技术，用于检视定量变量之间的关系。
- 计算特征值、特征向量和主成分分数。

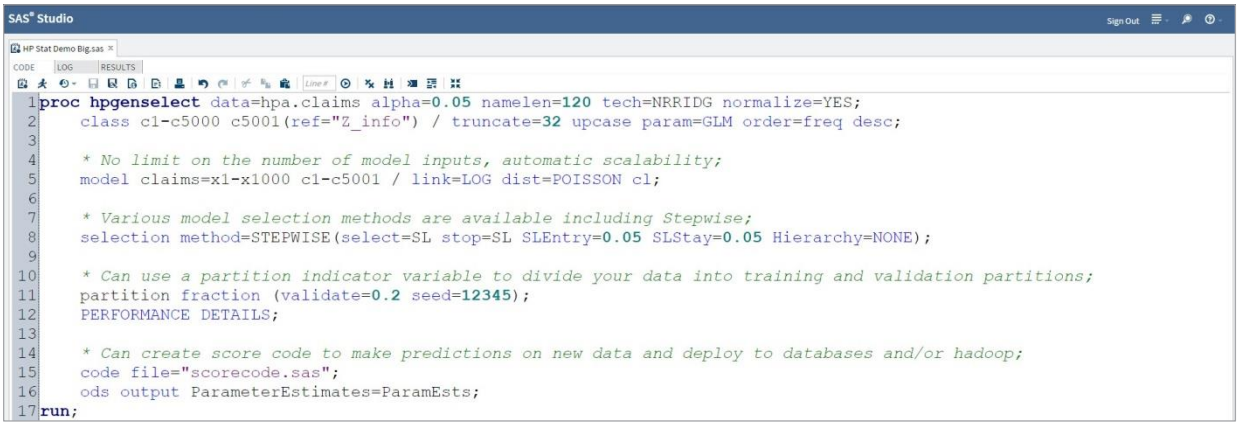


图 2: 采用 SAS 高性能统计，可在成千上万输入变量中进行选择，同时利用内存计算技术建立广义线性模型。

主要特点(续)

高性能典型 差别分析

- 支持降维。
- 计算类均值之间的马氏距离。
- 生成典型系数和典型变量行分。

SAS®高性能数据挖掘

高性能变量归约

- 结构化输入变量降维，选择原始变子量集。
- 通过指定一组变量共同解释最大量数据方差(协方差分析)，实现无监督变量筛选。
- 支持分布式计算，以及 CORR、COV 或 SSCP 矩阵输出。
- 利用 CLASS 语句支持类别型输入变量。
- 输出统计量和矩阵信息，用于后续统计过程。

高性能时序降维

- 通过相似度、聚类等方法降维。
- 支持三种时序输入数据格式：事务、转置和列式。
- 降维后的时序可以三种格式输出：事务、转置和列式。
- 事务格式输入数据可以处理多个时序变量。

高性能神经网络

- 支持输入变量和目标变量自动标准化。
- 智能设定大部分神经网络参数(如激活函数和误差函数)。
- 自动选择和使用校验数据子集。
- 当验证误差不再提高时，自动停止模型训练。
- 支持每项观测单独加权。
- 支持以非结构化文本信息作为输入，提高预测能力。
- 可使用任意数量隐藏层，支持深度学习。
- 可指定 Poisson 和 gamma 误差函数以及指数输出层激活函数，支持计数数据建模。
- 可指定隐藏层和输出层激活函数(identity、tanh 或 sin)。

高性能随机森林

- 建立数百棵决策树封装模型，预测单个目标变量。
- 并行训练分别在不同网格节点上运行的数百棵决策树。
- 在所有可用输入变量中，随机选择用于切分节点的输入变量。
- 仅考虑与目标变量最相关的单个变量进行切分。
- 支持以非结构化文本信息作为输入，提高预测能力。

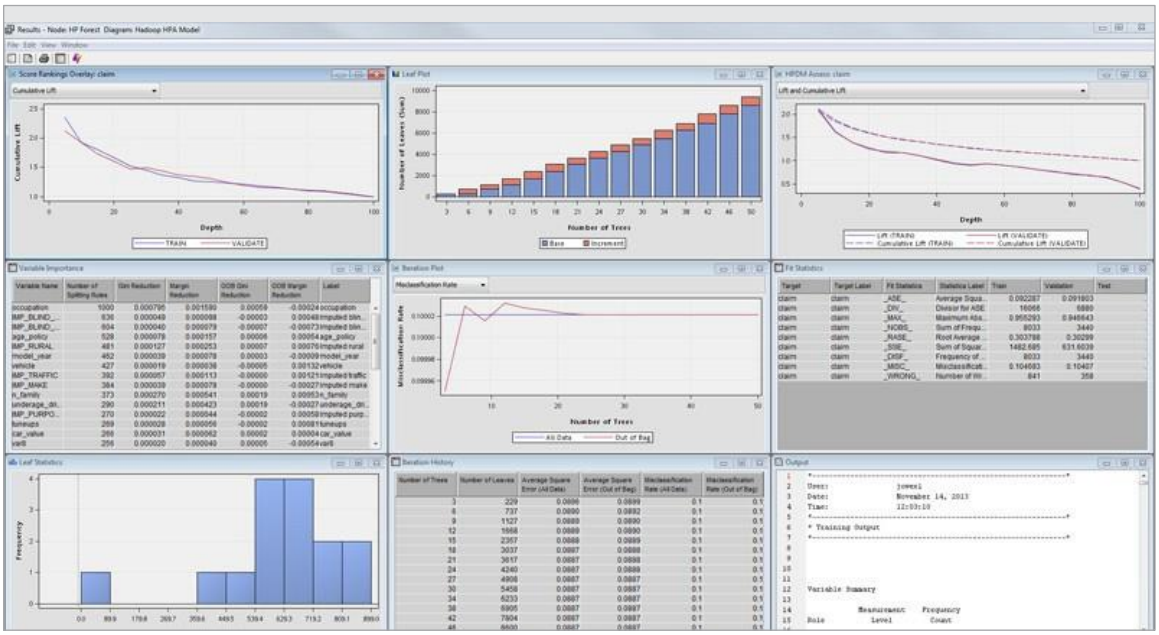


图 3: 利用 SAS 高性能数据挖掘，可采用随机森林等先进建模技术快速获得解决复杂问题的答案。

主要特点(续)

高性能随机森林评分

- 对通过 HPFOREST 程序训练生成的随机森林进行评分。

高性能决策

- 基于用户指定的决策矩阵、先验概率、建模程序输出(可以是区间型目标变量的预测值)生成最优决策。
- 决策矩阵中的列(决策变量)对应每个决策，行(观测) 对应目标值。决策变量的值表示特定目标，可以是利润、损失或营收。

高性能贝叶斯网络

- 构建贝叶斯网络模型。
- 建立不同结构的贝叶斯网络，包括朴素贝叶斯、树增强型朴素贝叶斯(TAN)、贝叶斯网络增强型朴素贝叶斯(BAN)、父子贝叶斯网络和马尔科夫覆盖。
- 通过独立测试有效选择变量，基于验证数据集根据指定参数自动选择最佳模型。
- 生成 SAS 数据步代码对数据集进行评分。

高性能聚类

- 基于一个或多个定量变量计算的距离进行聚类分析。将观测分入不同的聚类中，每个观测仅属于一个聚类。
- 进行 k-均值聚类，以区间变量为输入。
- 提供称为 aligned box criterion (ABC)的新技术估计数据集的聚类数目。

高性能支持向量机

- 使用线性和非线性核函数训练模型。
- 支持两种优化技术：内点法和有效集法。
- 支持以连续变量和类别变量作为输入，训练二值型目标变量模型。
- 内点法可用于单机模式或分布模式，有效集法仅用于单机模式。

SAS® Enterprise Miner™ 支持高性能节点

- HP 数据分区
- HP 数据探索
- HP 数据变换
- HP 变量选择
- HP 回归
- HP 数据贝叶斯网络
- 神经网络
- HP 随机森林
- HP 插补
- HP 决策树

SAS®高性能文本挖掘

自然语言处理 (NLP)

- 自动标注词性(系统中定义了 15 种以上不同词性)。
- 支持抽取标准实体，如位置、时间、日期、地址等，共 17 种预定义选项。
- 检测名词词组和多词词组，在机器学习过程中视为单个词条来处理。
- 检测不同词干，无需人工干预。
- 利用同义词检测，自动发现词条的多种变形。

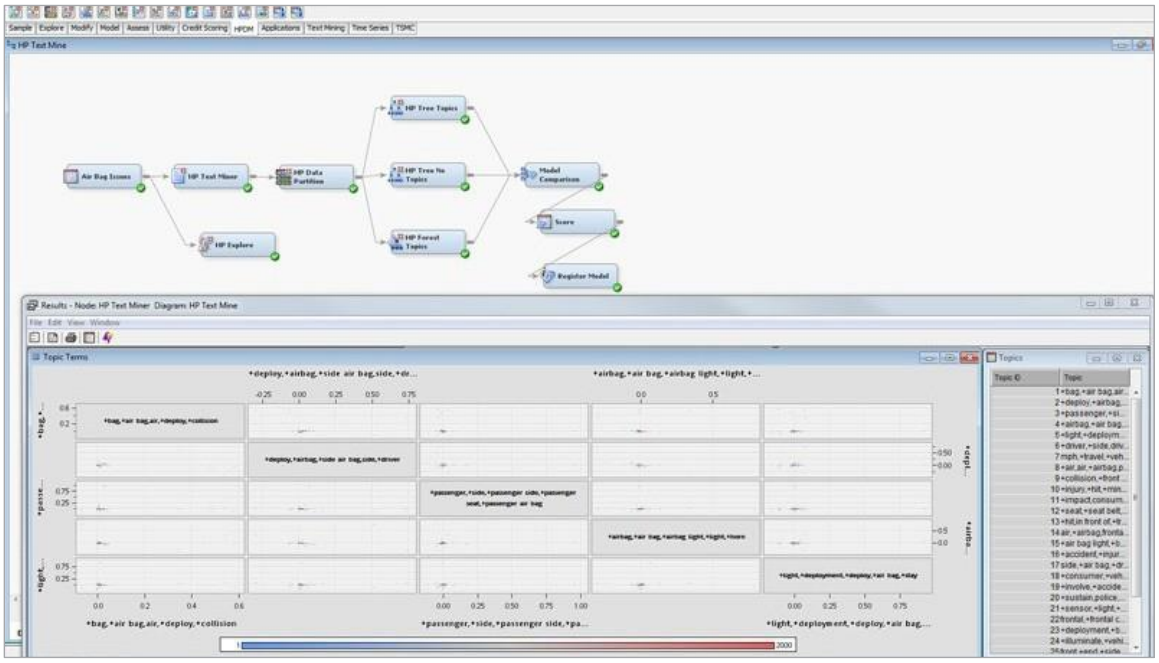


图 4: 采用内存计算技术，轻松创建含有文本挖掘主题的预测模型。

主要特点(续)

- 原生支持英文或德文，保留作者原义。

文本处理选项

- 选择所需文本解析选项，定制机器学习文本模型。
- 多种词条重要性加权方式，可基于词条在单个文档中出现的频率，以及在整个文档集合中出现的频率。
- 选择词频加权抑制词条出现过多产生的影响。
- 通过词条加权区分出更重要的词条。
- 重复文档识别和关键词控制处理可用于文本预处理和评分。

文本过滤

- 指定启用词表对应的 SAS 数据集，在文本解析和后续处理中仅处理指定的词条。
- 指定停用词表对应的 SAS 数据集，在文本解析和后续分析中排除指定的词条。
- 通过添加、删除和编辑词条调整启用词表和停用词表，包括加入多词短语。
- 指定包含词条的最小文档数，在后续分析中限制选定的词条。

主题生成

- 机器学习主题由词条-文档矩阵生成，可作为文档集合的结构化数字表达方式。
- 语义分析生成的主题可用作高性能结构化数据挖掘节点或过程的输入，展开进一步数据挖掘分析。
- 可以在其他 SAS 应用中使用标记的机器生成主题。

图表输出

- 通过查看词条特征明细表检查文本挖掘模型中的词条。
- 查看图表中显示的词条在文档集合中出现的频率、词条权重、包含词条的文档数，有助于评估文本挖掘模型或调优。
- 根据角色或属性表示的词频图形评估文档集合。
- 利用图形向导可以自行创建结果图表，或修改现有结果图表。

SAS®高性能计量经济

高性能事件数回归

- 拟合回归模型，其中的因变量为事件数量。
- 支持零堆积泊松回归和负二项回归模型，零堆积分布可以拟和各自的回归模型。

高性能严重性模型

- 拟和先验利润损失随机事件严重性(量级)的参数概率分布。

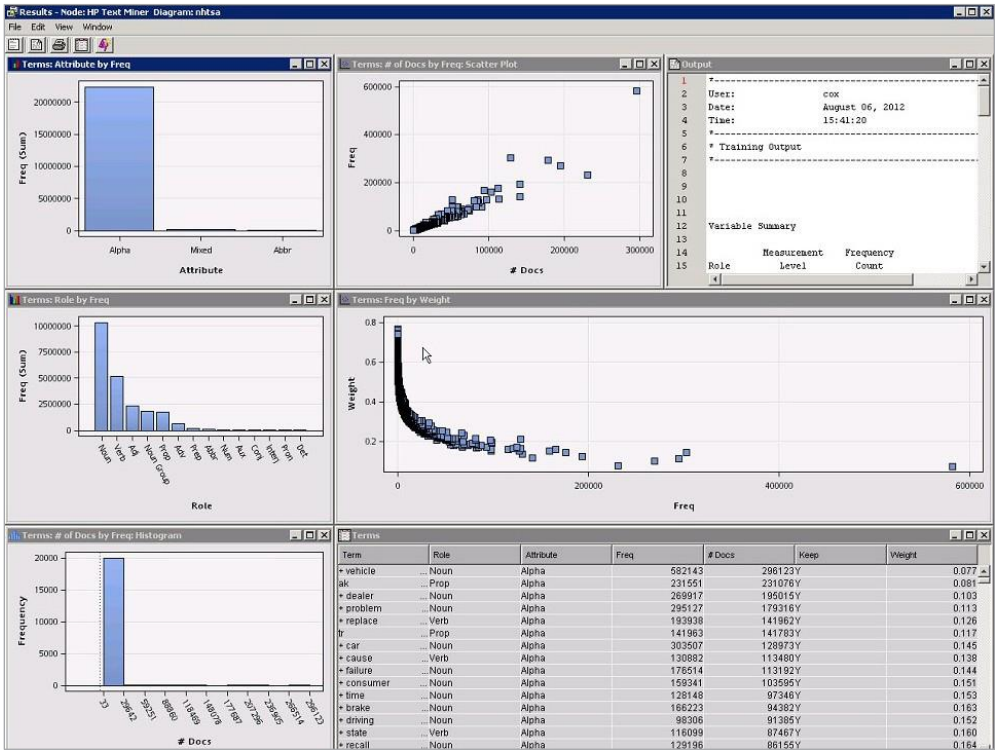


图 5: 在 SAS 高性能文本挖掘中查看词条属性分析的图形结果。

主要特点(续)

- 根据严重性分布规模拟合回归模型。
- 可在九种概率分布中自动选择最适合的分布，同时支持用户自行选择适合的统计。
- 允许用户增加新的概率分布。
- 用户可以对数据截断(免赔额)或数据删失(保单限额) 进行缺失值建模

高性能定性模型与有限因变量模型

- 拟合线性模型、logit/probit 模型、删失或截断回归异方差模型、随机前沿生产和成本模型。
- 支持估计一元和多元响应模型。
- 贝叶斯工具可供用户查找参数的后验概率分布。

高性能面板参数模型

- 估计线性面板模型，使用单向或双向固定或随机效应。

高性能 copula 模拟

- 使用给定的相关结构等信息，基于给定的多元 copula 模拟数据。

高性能复合分布模型模拟

- 使用计数数据模型和严重性模型建立用于保险业和银行业的聚合损失分布模型。
- 可用于执行假设及其他场景分析，支持不同的前提假设。也可以用于银行上报某些损失类型的 VaR (风险值)。
- 全面灵活的语法可模拟特定业务规则，大量不同的保险业模式(免赔额、保单限额等)可以进行分层。

SAS® 高性能优化

- 全局/局部搜索优化用户自定义函数(非线性、不可微分等)，支持连续型和整数决策变量，以及线性和非线性约束。
- 采用更有效的分解算法解决块角结构线性和混合整数线性优化问题。支持单目标和多目标优化。
- 非线性优化问题存在多个局部最优解时，可采用多起点方法提高获得全局最优解的可能性。
- 特定混合整数线性优化问题(或一组问题)，优调选项便于确定最有效的优化求解选项设定。

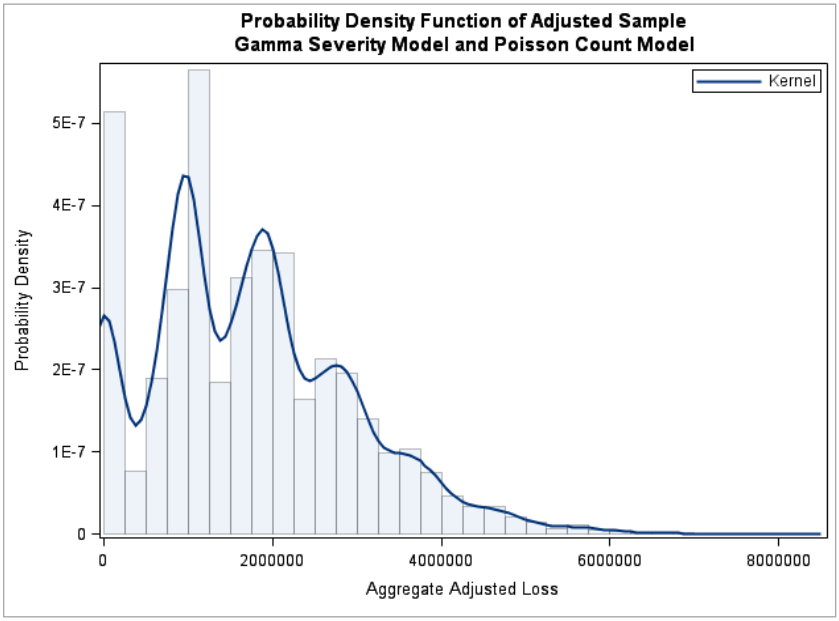


图 6: SAS 高性能计量经济可应用多种保险分层模式模拟聚合损失。

如需进一步了解SAS® 高性能分析产品的特性和技术要求，请访问 sas.com/hpanalytics。

105543_S138939.0715

THE POWER TO KNOW®