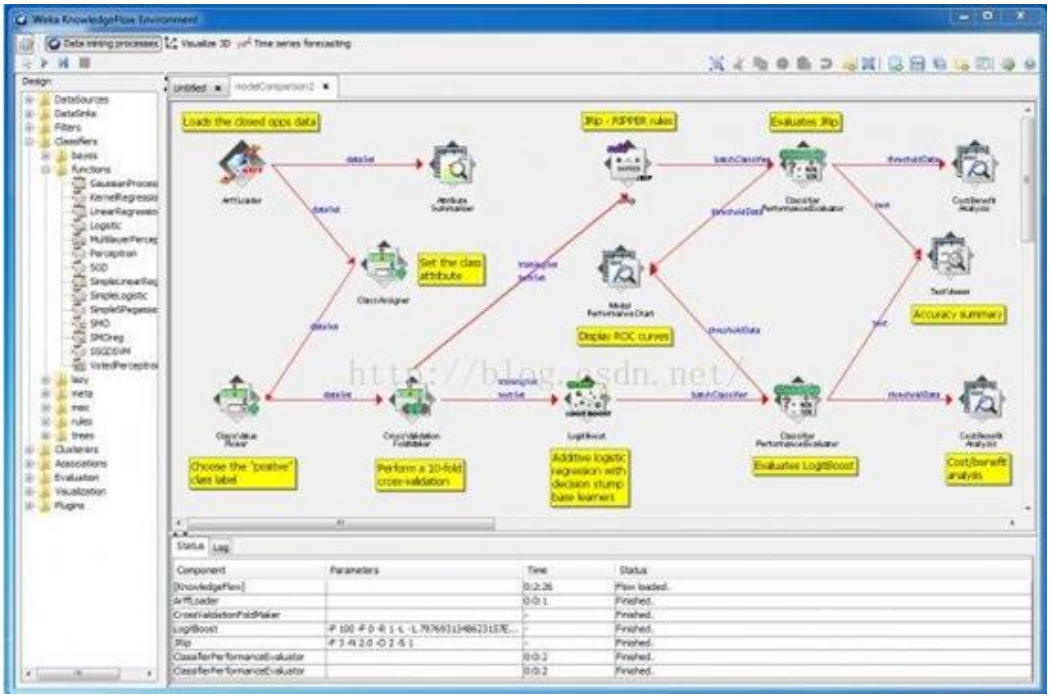
机器学习可视化工具调研分析

经典的机器学习可视化工具包括：RapidMiner、KNIME、DataInsight、东软 RealRec、明略以及 WEKA。

WEKA

WEKA 是一款免费开源的机器学习和数据挖掘可视化工具软件，其操作简便，运行速度快，尤其适合小规模的机器学习建模，是机器学习入门的不二选择。



特点

(1) 可移植性。WEKA 基于 Java 编程语言进行操作，从而几乎任何现代计算平台都可以运行。

(2) 支持标准的机器学习任务，包括数据预处理、聚类、分类、回归以及特征选择。各方面任务特点如下：

ü 数据预处理任务从数据库、CSV 文件等输入数据，并使用过滤算法对数据进行预处理。这些过滤器可用于转换数据（例如连续型数值属性变成离散型），从而根据具体的标准删除实例和属性。

ü 关联分析任务提供了不同的关联规则学习算法，譬如 Apriori 或 FP-growth 等，方便找出数据属性之间所有重要的关联关系。

ü 分类任务提供了贝叶斯、决策树、随机森林等不同的分类器进行数据分类学习，此外，使用户应用分类和回归算法的结果数据集，去评估预测模型产生结果的准确性，并可视化错误预测、ROC 曲线和模型本身(如果模型是适合可视化的，如贝叶斯)。

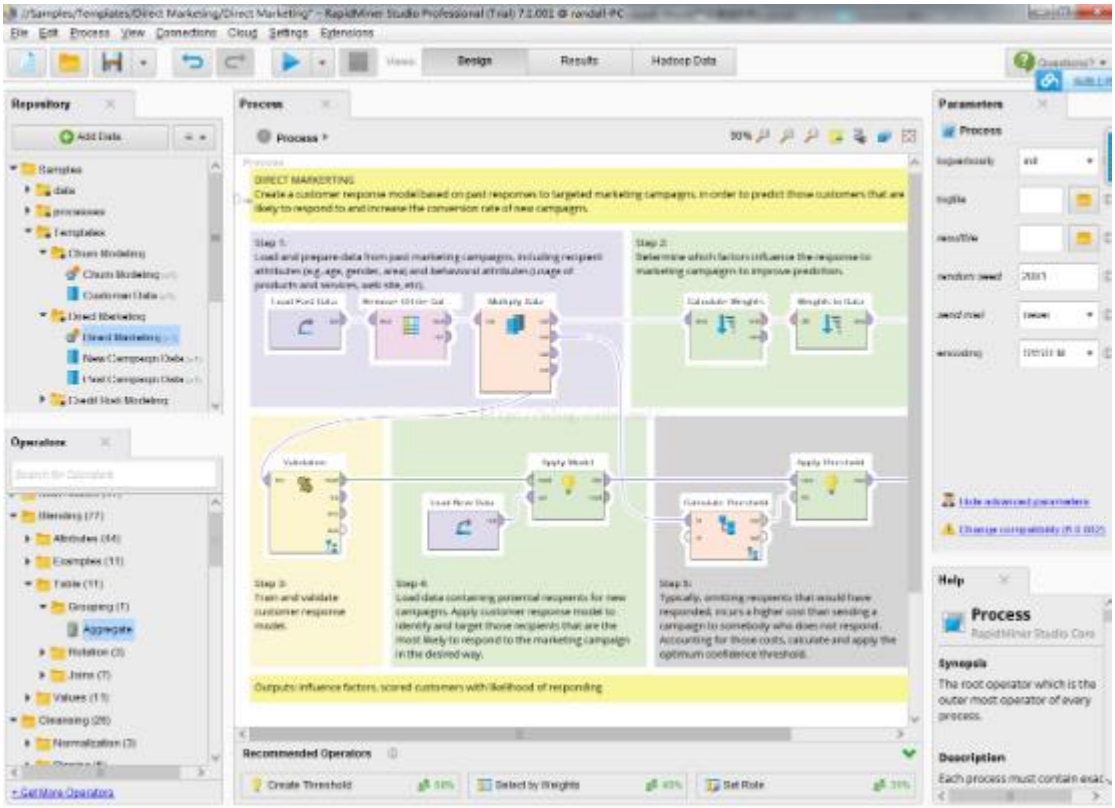
ü 聚类面板给 WEKA 提供了聚类技术，如简单的 K-Means 算法，也可以用期望最大化算法进行混合正态分布的学习。

ü 特征选择属性任务提供了数据集中大多数预测属性的识别算法。

(3) WEKA KnowledgeFlow 的所有技术是建立在数据可作为一个单一的平面文件或关系这个假设前提上的，其中每个数据点被设计成一个固定数量的属性（通常是数字或名义的属性，但一些其它的属性类型也被支持）。因此 WEKA Knowledge Flow 不能进行多位关系数据挖掘，但有独立的软件可以将连接的数据库表转换成一个单一的表，使其可以使用 WEKA 进行处理。

RapidMiner

RapidMiner 是集数据挖掘，机器学习，预测分析和商业智能为一体的可视化工具软件，包括文本挖掘、多媒体挖掘、功能设计、数据流挖掘、集成开发方法和分布式数据分析等。



特点

RapidMiner（以 RapidMiner V7 为例）具有以下特点：

(1) 包含完全集成的机器学习库 WEKA，提供关于数据集成、转换和建模方法的最全面的机器学习解决方案。

(2) 免费提供大量的数据抽取功能。包括 Oracle、IBM DB2、MS SQL Server、 MySQL、Ingres、Postgres、Teradata 等所有常见的数据库。支持 Excel、SPSS、CSV、Dbase、Arff、 DasyLab 等多种格式的数据源，以及 ASCII,、PDF、HTML 和 XML 格式的文本文档和和网页、时间序列数据等。

(3) 具有强大直观的图形用户界面设计分析过程。通过许多尖端的高维数据实现可视化建模和数据探索功能，形成在线的 1D、2D、3D 图，以及 Andrews、平行、偏差和 SOM 视图等。

(4) 具有模块化系统，使分析过程具有极大的灵活性和扩展性。RapidMiner 拥有超过 500 个数据集成和转化，分析和评估的模块工具。其中数据集成和分类工具包括支持向量机（SVM）、规则学习者、决策树、贝叶斯、高斯过程、神经网络、优化评估、boosting 算法、Apriority、FPGrowth 以及聚类等。评估工具包括交叉检验、leave-one-out、滑动时间窗、回溯测试、显着性检验以及 ROC 等。

(5) 模块化的运行概念奠定了机器学习模型创建的设计流程。元运算允许这些过程自动优化，因此使用者不需要手动去适应每个步骤和参数。优化运算包括自动参数优化、自动属性设置优化、循环、控制结构、宏、断点调试以及更多。

(6) 模型的快速成型和超越。从第一次探索分析到现成解决方案只需几步。快速成型允许机器学习模型过程中的关键决策可以尽可能早的实现。使用 RapidMiner 可以使你在短时间内设计好一个原型，从这些原型中，优化这些过程引导你得到解决方案。

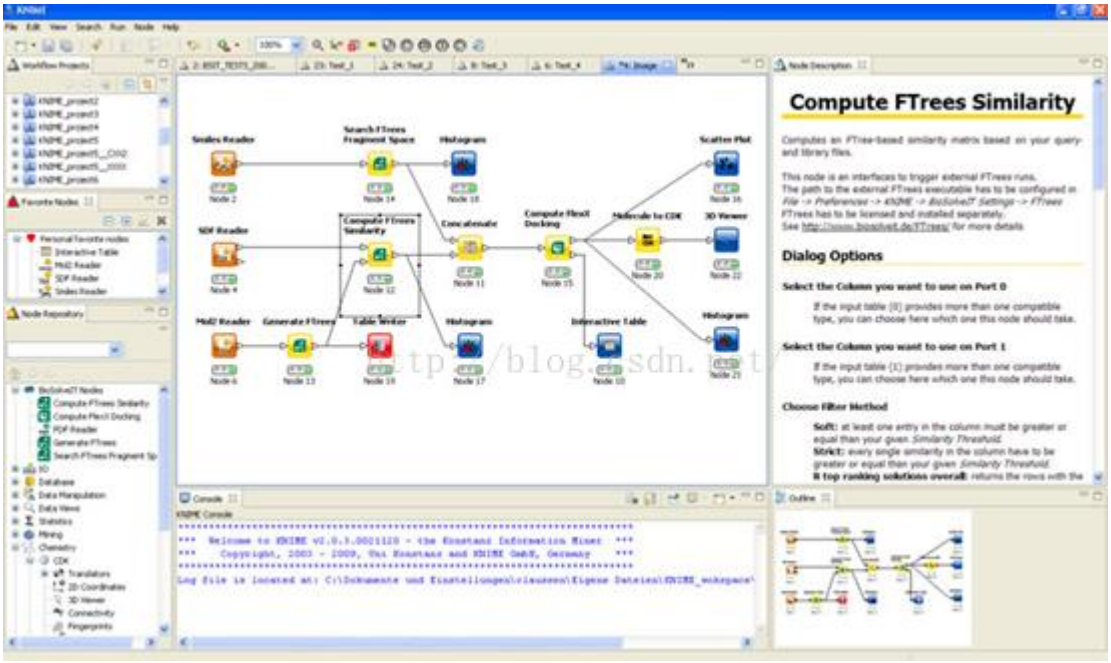
(7) 使用 Java 代码，可通过 GUI 模式或 Java API 进行操作，也可以用简单脚本语言自动进行大规模进程操作。并且内部 XML 保证了标准化的格式来表示建模过程。

(8) RapidMiner 与市面上主流的大数据分析平台支持非常友好，譬如不同版本的 CDH 集成，包括同一平台的不同版本的支持，不需要应对复杂的配置文件，因此能够方便应对大规模分布式机器学习应用场景。

(9) RapidMiner 有成熟的社区和扩展交易市场（marketplace），用户可以根据自己的需要，快速获取很多定制化的功能。

KNIME

KNIME 是一个集数据集成、处理、分析和开发于一体的开源数据分析平台，对常用的机器学习算法也提供了友好的支持，并在严格的软件工程实践中进行开发和利用。本文主要对 KNIME 3.1 进行比较研究。



特点

(1) 可视化的工作平台集成了数据访问、数据转换、数据探索和预测分析等常用的机器学习功能。

(2) 集成了数百个处理结点来进行数据输入与输出、预处理和清洗、建模、分析、数据挖掘以及制作各种互动的视图（如散点图、平行坐标和其他视图）。

(3) 可集成所有的分析模版到众所周知的 WEKA 数据挖掘环境中，并有额外的插件模块允许 R-脚本运行，还提供了广大统计例程库接口。

(4) 基于 Eclipse 平台开发，并且通过其模块化的 API 可轻松进行扩展。因为 KNIME 在后台可进行智能自动的数据缓存，同时最大限度地提高吞吐量性能，所以这种模块化和可扩展性允许 KNIME 在商业的生产环境以及教学和研究原型设置工作中得到应用。

(5) 提供超过 1000 个数据分析例程，无论是在本地或通过 R 和 WEKA 都可以进行，如单元和多元统计、数据挖掘、时间系列、图像处理、Web 分析、文本挖掘以及社会化媒体分析等。

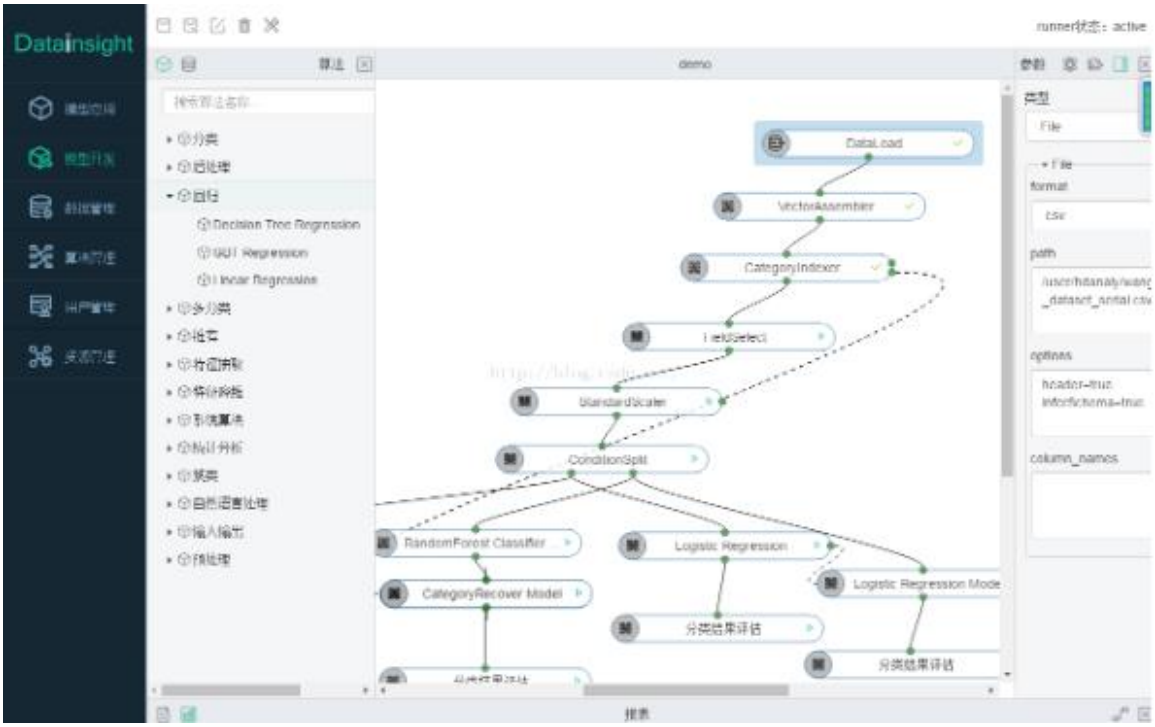
(6) 机器学习建模工作流程不仅可以通过交互式用户界面运行而且执行批处理模式，使数据分析过程可以很容易地定期集成到本地工作运行的管理中去。

(7) KNIME 提供了大量的行业应用模板和定制化化的算子，便于特定应用行业的数据分析，譬如生物医药行业。

(8)具有 HiLite 功能，允许用户在节点结果中标记感兴趣的记录，并进一步展开后续探索。

明略

明略可视化机器学习平台 DataInsight 本质是一种 MLAAS 平台，用户无须在客户端安装平台工具，通过浏览器即可进行拖拽，交互式数据探索，完成机器学习模型的训练、部署和应用。DataInsight 不仅集成了 Spark MLlib 分布式机器学习能力，还定制了高效的分布式机器学习算法。



特点

明略 DataInsight 平台基于 BS 架构，DataInsight 通过提供一体化、并行化的高效模型应用平台，能帮助企业有效降低机器学习的应用曲线和落地成本。具有如下特点：

(1) 扩展性强。明略 DataInsight 平台基于 Hadoop 和 Spark 的并行化平台，计算能力随着大数据平台计算能力的扩展而扩展。其提供了多种数据预处理的并行化算法，以及大量并行运行于 Spark 之上的数据挖掘和机器学习算法。

(2) 模型工作流。使用工作流的概念表示整个建模过程，每个建模步骤看作一个算子，使得整个建模过程形成一幅有向无环图，建模过程将原始的输入通过一系列算子组合得到最终的业务结果。

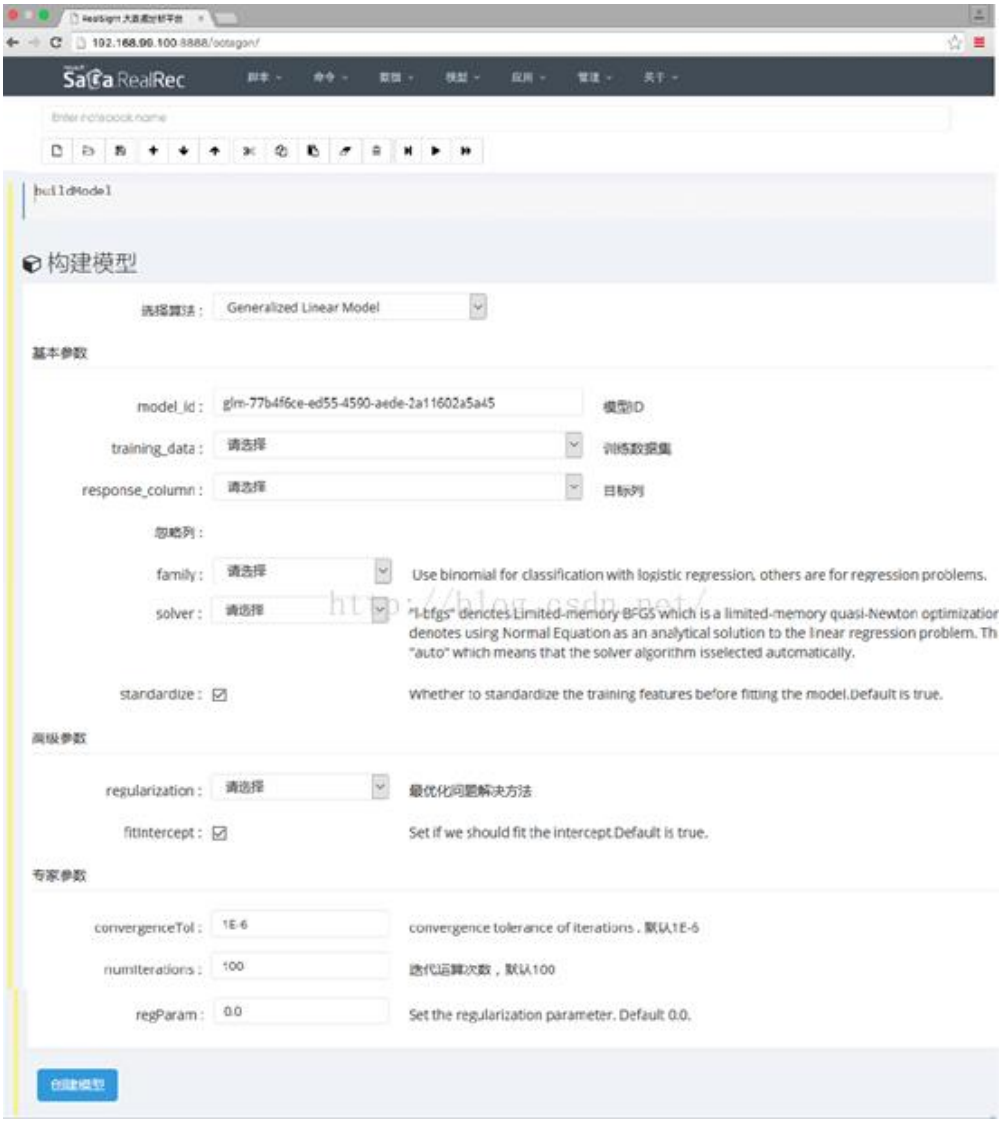
(3) 交互式探索。明略 DataInsight 提供了交互式数据探索工具供用户对数据进行实验性的探索工作，帮助用户实时的对数据进行探索和实验。同时，明略 DataInsight 通过可视化的方法，提供了常用的数据统计和分析的图表，供用户能够直观的从图形中发掘数据背后的意义。

(4) 模型应用管理。提供模型应用的版本管理，能够方便的进行模型的维护和更新，提升工作效率。并且对模型的应用管理提供了用户和角色的支持，方便权限控制。

(5) 模型即服务。DataInsight 平台通过 Restful API 向企业其他生产系统提供服务，外部系统可以通过 Restful API 实现模型的运行和更新等操作。

东软 RealRec

东软数据科学平台定位于企业级数据科学平台，通过简化复杂机器学习算法的使用成本，提高企业构建智能应用的能力和效率，帮助企业实现数据驱动的商业模式。产品组成如下：



特点

东软 RealRec 主要通过在机器学习算法和模型方面的能力建设，使企业能够快速构建智能应用，开展创新数据服务和业务。

主要具备如下特点：

(1) 特征分析实现了多维分析和分布统计，通过对数据进行过滤、去重、拆分、合并，实现数据可视化展现，并实现了数据的特征抽取、管理和复用；

(2) 模型训练实现了自动化的模型选择，模型的交叉验证和可视化展现，并提供全流程的引导，训练数据可以是实时数据、批量数据或文件；

(3) 在模型最终部署应用时提供跨平台 REST 接口、POJO 导出/UDF 导出，并实现了模型管理复用和任务调度管理；

(4) 机器学习通过提取原始数据的特征，选择合适的算法，最终实现模型的自动化选择和导出。

（一）机器学习算法的比较

上述可视化机器学习工具对常用的机器学习算法都提供了支持，但具体对某一算法的支持程度则有所不同，譬如以支持向量机算法为例：KNIME仅支持LibSVM，SparkLinear SVM，Rapid Miner可以支持LibSVM，Linear SVM，Evolutionary SVM以及PSO SVM，以及Spark SVM，明略支持Spark SVM和LibSVM，而WEKA仅支持LibSVM。

各算法支持功能强弱可参考下表：

机器学习算法	WEKA	RapidMiner	KNIME	明略	东软RealRec
支持向量机	中等	强	较强	中等	下一版本支持
决策树	中等	强	强	中等	强
贝叶斯	中等	强	较强	中等	强
回归	中等	强	较强	中等	强
神经网络	弱	较强	较强	弱	强
深度学习	不支持	很弱	很弱	不支持	强
K-Means	中等	较强	中等	弱	强

说明：

ü 目前RAPID MINER正在集成JAVA深度学习开源项目DeepLearning4J，可以支持深度学习建模，但是目前尚不成熟。

ü 目前WEKA不支持分布式机器学习算法，其他工具均支持分布式的机器学习算法。其中RAPIDMINER和KNIME均通过软件扩展集成Spark，通过Spark MLLIB的算法支持分布式机器学习算法。明略DataInsight目标即是面向大数据的分析和建模。

（二）可视化功能的比较

软件产品的使用容易程度至关重要，一款好的可视化工具应能够实现训练数据的可视化探索、模型的可视化、模型训练的可视化、模型验证及应用的可视化，能够自带多行业的模板和样例，便于数据分析人员的快速上手，提升模型建立和训练的效率。具体表现在数据的导入/处理、模型的构建、模型的易理解性等方面。

比较项目	WEKA	RapidMiner	KNIME	明略	东软RealRec
数据抽取	只能通过数据文件、URL地址以及数据库抽取收取，不支持其他数据源数据抽取，功能单一。	支持各种格式文件、数据源的抽取，图形化的抽取算子，并进行数据的交互式探索	也具备常用文件格式和数据源的抽取，但数据的交互式探索功能不方便	能够支持多个文件格式、数据库抽取，对大数据平台数据抽取能力强	支持文件上传、HBase数据、HDFS数据、关系型数据库数据，可对接大数据平台
数据转换	提供常用的数据过滤、归一化等功能	数据的归一化、降维和格式化等	各种数据格式化功能，数据融合、数据过滤等	提供常用的数据过滤、归一化等功能	提供常用的数据过滤、归一化、切分等功能
数据建模	提供简洁的拖拽等可视化建模能力。	类似工作流的可视化建模，支持嵌套。	也提供工作流的可视化建模，但是设置选型多，比较复杂	也提供工作流的可视化建模，但功能还不完善，正在改善。	提供notebook(记事本)式的数据建模，以命令行的方式运行模型
模型验证及评测结果展示	支持常用的图表展示，但是图表美观性差	各种图形的可视化展示，切换容易	各种常见图形的展示	支持的图表有限，目前正在改善。	支持常用的图表展示，但是图表美观性一般

（三）不同用户群学习及使用难易度的比较

不同机器学习工具具有不同功能特点，有的灵活性强，有的功能简洁，这些都会影响不同用户群的学习曲线以及工具使用的难易程度。我们根据不用用户群知识结构特点以及上述工具产品的功能特点综合分析比较，不同用户群掌握工具难易程度如下：

比较项目	WEKA	RapidMiner	KNIME	明略	东软RealRec
开发人员	操作界面比较简洁和直观；工具是轻量级的，对系统资源消耗少，运行快捷；对于开发人员入门简单，使用便捷	对运行的硬件资源要求比较高，运行速度不是很快；界面设计比较清晰直观；对于开发人员入门也非常容易	工具操作界面基于eclipse开发，对于熟悉eclipse开发工具的研发人员比较简单；由于选项多、灵活性大，掌握难度大	完全基于WEB的操作界面，界面风格简洁直观，对于开发人员入门简单；由于WEB版，运行于浏览器，对硬件资源需求小，操作便利	完全基于WEB的操作界面，对开发人员入门简单
专业数据分析人员	对于熟悉数据分析专业知识人员，工具使用简单，操作便利	由于界面比较直观，对于专业数据分析人员比较简单，门槛低	由于界面灵活，配置也相对复杂，使用相对复杂，需要摸索	WEB操作界面比较简单直观，功能比较单一，用户入门门槛不高	WEB操作界面比较简单直观，功能比较单一，用户入门门槛不高
一般业务分析人员	界面描述均是专业术语，一般分析人员无法轻松使用，学习周期长	界面虽然比较直观，但还具有一定的专业性，需要专业指导	由于功能提供了较多专业的配置项，用户需要花费时间学习配置，学习周期比较长	界面虽然比较直观，但还具有一定的专业性，需要专业指导	界面虽然比较直观，但平台定位于计算引擎，一般业务分析人员需要专业指导

（四）主流大数据平台支持能力的比较

为了提升机器学习的效率以及数据分析的规模，当前主流的机器学习工具均提供了与大数据分析平台的集成，但在兼容性方面存在比较大的差异，具体比较如下：

比较项目	WEKA	RapidMiner	KNIME	明略	东软RealRec
大数据平台的兼容性	目前不提供与大数据平台的集成	支持主流的大数据平台Clouddera和Hortonworks以及MapR， CDH 4.x和5.x， HDP 1.x和2.x， MapR 4.1， Hadoop 1.x和2.x等	支持主流的大数据平台Clouddera和Hortonworks以及MapR，支持 CDH 5.x， HDP 2.1 和 2.2， MapR 4.1， Hadoop 2.4.0等	兼容性较差，主推自己的大数据平台，目前也支持其他的大数据平台，但需要验证。	支持大数据平台Clouddera，支持Spark 1.6，自行定制的Tachyon 0.8
与大数据平台集成难易度	目前不提供与大数据平台的集成	集成快捷，提供配置模板，对已有大数据环境无侵入	配置相对复杂，对已有大数据环境无侵入	需要在已有大数据环境安装相应软件套件，有侵入性	配置相对复杂，对已有大数据环境无侵入

0

（五）软件工具自身成熟度比较

本节主要从软件自身的稳定性、文档的完整性、用户群、社区生态系统以及开源协议支持等五个方面对WEKA、RapidMiner、KNIME以及明略DataInsight进行对比，综合评定各软件总体成熟度。

比较项目	WEKA	RapidMiner	KNIME	明略	东软RealRec
软件稳定性	将近二十年的发展和优化历史；功能稳定、扩展性强但	超过十五年的发展历史，功能稳定并且全面、扩展性强。目前正在研发纯WEB版的分析平台	超过十年发展历史，功能稳定全面、扩展性强，缺少WEB版的分析平台	发展历史比较短，功能不够全面，产品还处于发展中，不够稳定	发展历史比较短，定位于底层支撑平台，产品还处于发展中
文档完备性	是缺乏规范的文档列表	文档规范并且齐全	文档规范并且齐全	文档不够规范，用户应用不方便	文档规范，但不太齐全
用户群及应用领域分布	在高校和研究机构，有着广泛的用户群体，用户基数大。	产品有比较广泛的用户群体，成功应用于汽车、金融、保险等领域	国内用户群比较窄，用户群主要分布在生命科学、政府、金融等领域	目前用户群体基数小，用户主要分布在电商和金融行业	目前用户群体基数小，用户主要分布在客服和金融行业
社区生态系统	由于用户群基数大，社区发展比较成熟，文档和手册指南比较成熟，容易获取社区帮助	社区发展比较成熟，扩展多，可以获取很多免费的视频等学习资料，存在专业的论坛提供帮助和交流	社区规模比较丰富和规范，扩展多可以获取很多免费的视频等学习资料，也有专门的论坛提供交流和指导	由于是纯商业软件，目前没有形成成熟的社区，难以获取社区帮助，需要依赖厂商	由于是纯商业软件，需要依赖厂商
开源协议	GPL，协议灵活，代码完全开放	单机版支持协议AGPL-3.0，协议灵活，代码完全开放	单机版支持协议GPL-V3，协议灵活，代码完全开放	商业软件	商业软件

总上，RapidMiner发展历史久，功能稳定完备，用户群分布广、社区生态成熟，因此产品成熟度也最高。

工具使用成本比较：

WEKA 是完全免费的开源软件，无须支付任何软件费用，使用成本低；RapidMiner 和 KNIME 单机版完全开源，无需支付费用，但是分布式模型训练和处理组件以及面向特定应用领域的扩展组件均是商业版本，需要收费，license 主要按照用户数和使用期限收费，费用不菲；明略 DataInsight 和东软 RealRec 是纯商业软件,成本比较昂贵。

基于上述六个层面的比较分析，我们可以看出 RapidMiner 在机器学习领域应是一个全面综合的软件工具，在算法和可视化效果方面都很突出，特别适用于不同数据的多方面分析和研究，适用于数据和业务分析人员。WEKA 适合小规模的机器学习能力，譬如科研探索和机器学习入门人员等。KNIME 比较接近 RapidMiner，但由于界面比较复杂，因此比较适合开发人员，尤其是具备 Eclipse 开发经验的人员。明略 DataInsight 具有较强的任务管理和模型管理能力，并提供角色管理，适用于工程领域的研发人员。东软 RealRec 算法非常全面，适合复杂场景的分析和训练，但是可视化建模能力弱，使用复杂，不建议一般业务人员使用。

一款优秀的可视化机器学习工具应具备如下功能：

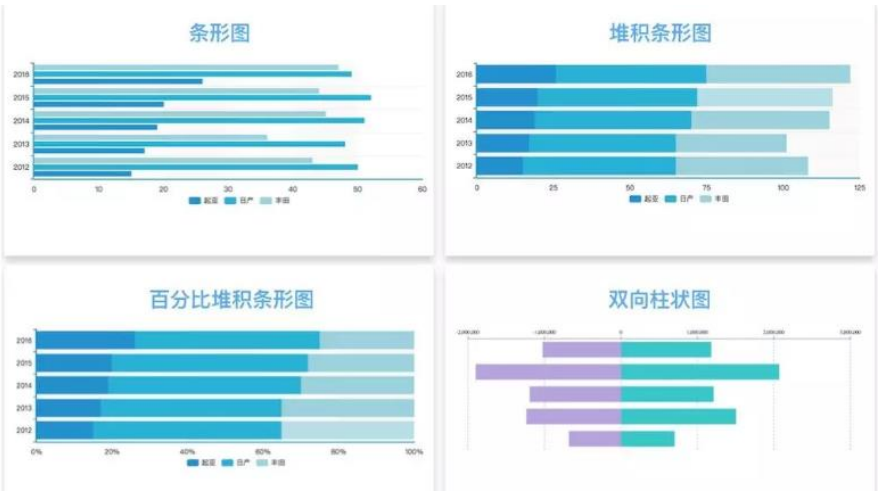
1. 首先，具备良好的可视化能力，包括可视化数据探索、可视化数据预处理、可视化建模、可视化验证、可视化评估、可视化结果展示等。
2. 操作界面简洁直观，对分析组件进行模块化，有效降低组件之间的数据格式的兼容性，让数据分析变得敏捷化。
3. 对常用的机器学习算法都提供了友好支持，不仅具备算法支持的深度，还可以扩展支持同一种算法的不同实现，以应对不同分析领域。
4. 尽可能面向更广的用户群，譬如能够覆盖一般业务分析人员、专业的数据分析人员或者技术开发人员，用户学习成本均比较低或仅需简单培训。
5. 具备完善的社区生态系统，提供软件工具使用过程中所需帮助信息，此外，社区还能够提供特定领域分析的扩展或插件。
6. 能够和企业自身大数据分析环境无缝连接，对不同大数据系统及版本均提供了友好的支持，易于部署和管理。
7. 工具训练的模型能够轻松的向生产环境迁移，且便于模型的运营和升级维护。
8. 最后，一款好的可视化机器学习工具还需在主流行业中得到相对广泛的应用，具有较大的用户群，经历过实践检验。

常见各图表适用场景

1、柱状图

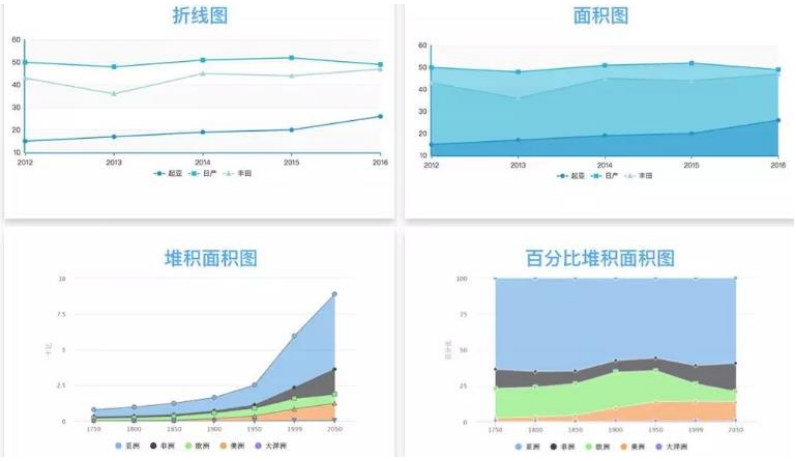
展示多个分类的数据变化和同类别各变量之间的比较情况。
适用：对比分类数据。局限：分类过多则无法展示数据特点。相似图表：
1）堆积柱状图。比较同类别各变量和不同类别变量总和差异。
2）百分比堆积柱状图。适合展示同类别的每个变量的比例。

2、条形图



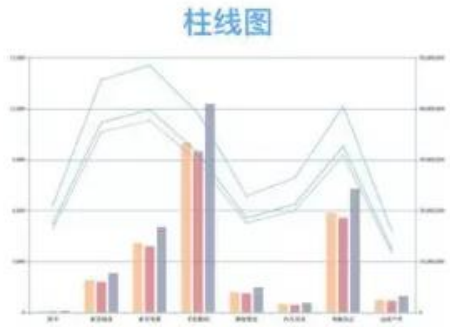
类似柱状图，只不过两根轴对调了一下。
适用：类别名称过长，将有大量空白位置标示每个类别的名称。局限：分类过多则无法展示数据特点 。
相似图表：
1）堆积条形图。比较同类别各变量和不同类别变量总和差异。
2）百分比堆积条形图。适合展示同类别的每个变量的比例。
3）双向柱状图。比较同类别的正反向数值差异。

3、折线图



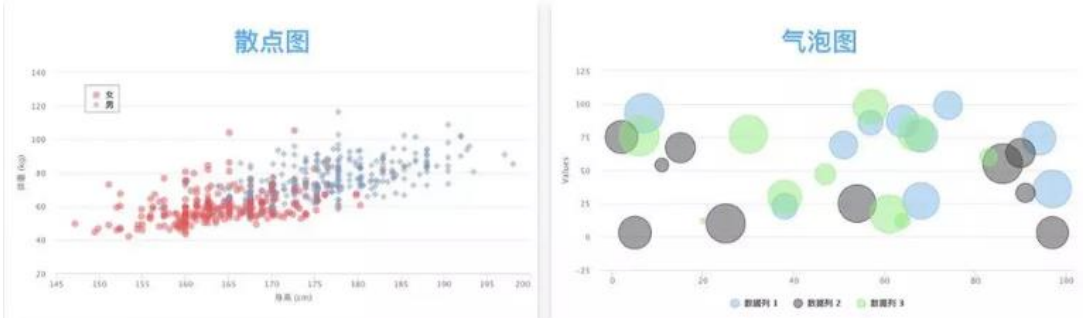
展示数据随时间或有序类别的波动情况的趋势变化。
适用：有序的类别，比如时间。局限：无序的类别无法展示数据特点。相似图表：
1）面积图。用面积展示数值大小。展示数量随时间变化的趋势。
2）堆积面积图。同类别各变量和不同类别变量总和差异。
3）百分比堆积面积图。比较同类别的各个变量的比例差异。

4、柱线图



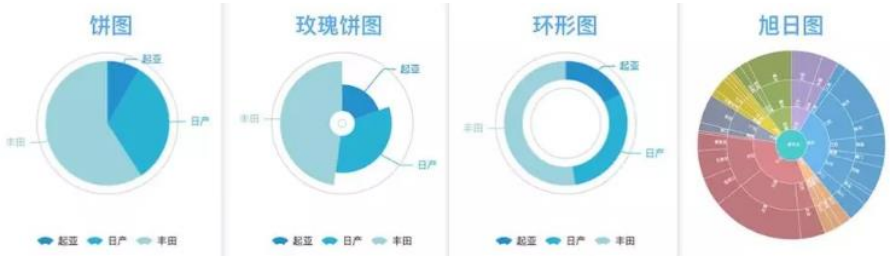
结合柱状图和折线图在同一个图表展现数据。
适用：要同时展现两个项目数据的特点。局限：有柱状图和折线图两者的缺陷。

5、散点图



用于发现各变量之间的关系。
适用：存在大量数据点，结果更精准，比如回归分析。局限：数据量小的时候会比较混乱。相似图表：
气泡图。用气泡代替散点图的数值点，面积大小代表数值大小。

6、饼图

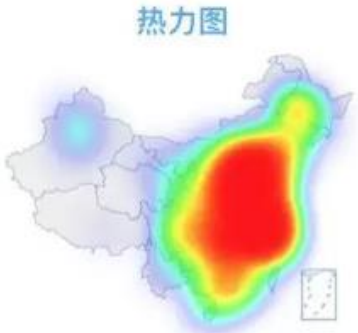


用来展示各类别占比，比如男女比例。
适用：了解数据的分布情况。缺陷：分类过多，则扇形越小，无法展现图表。相似图表：
1）环形图。挖空的饼图，中间区域可以展现数据或者文本信息。
2）玫瑰饼图。对比不同类别的数值大小。
3）旭日图。展示父子层级的不同类别数据的占比。

7、地图

用颜色的深浅来展示区域范围的数值大小。
适合：展现呈面状但属分散分布的数据，比如人口密度等。局限：数据分布和地理区域大小的不对称。通常大量数据会集中在地理区域范围小的人口密集区，容易造成用户对数据的误解。相似图表：
1）气泡地图。用气泡大小展现数据量大小。
2）点状地图。用描点展现数据在区域的分布情况。
3）轨迹地图。展现运动轨迹。

8、热力图



以特殊高亮的形式显示访客热衷的页面区域和访客所在的地理区域的图示。
适合：可以直观清楚地看到页面上每一个区域的访客兴趣焦点。局限：不适用于数值字段是汇总值，需要连续数值数据分布。

9、矩形树图

展现同一层级的不同分类的占比情况，还可以同一个分类下子级的占比情况，比如商品品类等。
适用：展示父子层级占比的树形数据。缺陷：不适合展现不同层级的数据，比如组织架构图，每个分类不适合放在一起看占比情况。

10、指标卡

突出显示一两个关键的数据结果，比如同比环比。
适合：展示最终结果和关键数据。缺陷：没有分类对比，只展示单一数据。

11、词云

展现文本信息，对出现频率较高的“关键词”予以视觉上的突出,比如用户画像的标签。
适合：在大量文本中提取关键词。局限：不适用于数据太少或数据区分度不大的文本。

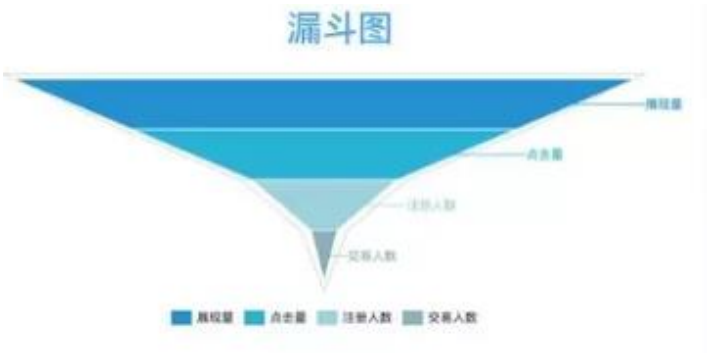
12、仪表盘

展现某个指标的完成情况。
适合：展示项目进度。局限：只适合展现数据的累计情况，不适用于数据的分布特征等。

13、雷达图

将多个分类的数据量映射到坐标轴上，对比某项目不同属性的特点。
适用：了解同类别的不同属性的综合情况，以及比较不同类别的相同属性差异。局限：分类过多或变量过多，会比较混乱。

14、漏斗图

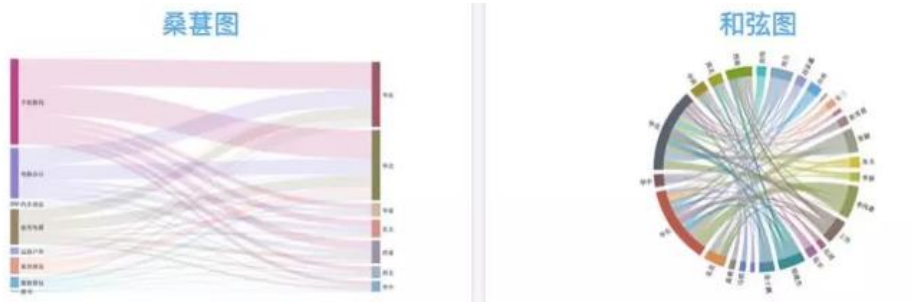


用梯形面积表示某个环节业务量与上一个环节之间的差异。
适用：有固定流程并且环节较多的分析，可以直观地显示转化率和流失率。局限：无序的类别或者没有流程关系的变量。

15、瀑布图

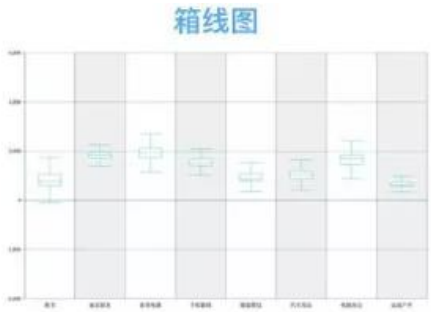
采用绝对值与相对值结合的方式，展示各成分分布构成情况，比如各项生活开支的占比情况。
适合：展示数据的累计变化过程。局限：各类别数据差别太大则难以比较。

16、桑葚图



一种特定类型的流程图，图中延伸的分支的宽度对应数据流量的大小，起始流量总和始终与结束流量总和保持平衡。比如能量流动等。
适合：用来表示数据的流向。局限：不适用于边的起始流量和结束流量不同的场景。比如使用手机的品牌变化。相似图表：1）和弦图。展现矩阵中数据间相互关系和流量变化。数据节点如果过多则不适用。

17、箱线图



是利用数据中的五个统计量：最小值、第一四分位数、中位数、第三四分位数与最大值来描述数据的一种方法。
适用：用来展示一组数据分散情况，特别用于对几个样本的比较。局限：对于大数据量，反应的形状信息更加模糊。