??????

- 一、敏捷挖掘流程
  - 概述
  - 商业目标
  - 数据准备
  - 建模
  - 评估
  - 应用
- 二、应用案例
  - 概述
  - 分类预测
  - 关联分析
  - 聚类分析
  - 时间序列
  - 数据清洗
  - 指定时间区间
  - 文本挖掘
  - 社会网络分析

?????????

一、敏捷挖掘流程

??

AOSP-SM（即 “应用为导向的敏捷挖掘标准流程 ”的缩写）是结合专家团队长期的商业实践经验并基于 IBM 的跨行业数据挖掘过程标准（ CRISP-DM）和 SAS 的数据挖掘方法 (SEMMA) 总结而来的一种面向应用的用于指导大数据挖掘工作的方法。

作为一种方法，它包含项目中各个典型阶段的说明、每个阶段所包含的任务以及这些任务之间的关系的说明。
作为一种流程化建模方法论， AOSP -SM 概述了数据挖掘的整个过程。

整个流程由五个阶段组成，阶段之间并不一定要严格遵守顺序。实际上，大多数项目都会根据需要在这些阶段之间反复中优化。

AOSP-SM

模型具备灵活性，并不一定完全按照这几步完成。比如，如果一个产品总代理发现近年来二级代理的订单违约的情况越来越多，管理层想到这可能和销售员有关系，于是分析人员可以去分析每个销售员的订单违约情况是否有差异，进而验证订单违约与销售员是否存在某种关联。这种情况一般都不需要进行建模。

从各个阶段工作难度和重要性来看，商业目标和评估两个阶段最为重要，如下图：

从各个阶段工作量来看，数据准备和应用两个阶段最多，如下图所示：

一、敏捷挖掘流程

????

??????

这是我们开始数据挖掘的第一步。在我们确定是否要启动数据挖掘项目时，首先要对相应的业务进行梳理。通过对业务进行了解，可以明确业务中存在问题，以及是否可以通过数据挖掘的手段解决相应的问题。

确定业务目标通常有两种办法：一是业务调研，二是理解数据。通过业务调研对业务用户进行访谈以掌握其业务中的痛点，同时也可以为今后实施方案的设计提供思路。但并不能仅作业务调研，之后还必须对相应的数据进行探察，以了

解数据资源情况以及数据质量，为项目启动的可行性提供决策参考。通过详细说明业务问题、数据情况和业务资源，可以更快的明确数据挖掘的业务目标。

这部分工作往往是数据价值挖掘的基础能力，也是一个企业数据挖掘团队中往往容易缺乏的能力之一，因此这一部分工作最好由业务专家和项目经理直接参与。

举例，有一家做百货的公司，他们希望得到更好的促销方案，以提升公司效益，那他们的业务目标就是希望得到更好的促销方案提升促销收益，以及要对某个品牌促销的话，应该把信息推荐给哪些客户。

????

既然已经有了一个明确指定的目标，现在应该评估处于什么情况。这一步骤需要询问一些问题，例如：

- ○ ○ 什么类型的数据可供分析?
  - ○ 是否具有完成此项目所需的人力资源?
  - ○ 所涉及的最大风险因素是什么?
  - ○ 对于这些风险，是否具有相应的应急计划?

????????

既然已经明确了业务目标，现在应该将其转换为数据挖掘目标。例如，“促销分析”的业务目标可以转换为包含下列信息的数据挖掘目标：

客户都是什么类型的？客户的购买力如何?

客户的品牌忠诚度指数是多少？可以得到哪些产品的关联规则?

正如所见到的，业务和技术必须紧密配合才能获得有效的数据挖掘。

??????

□□□

一、敏捷挖掘流程

????

????

企业数据来自各种不同的数据源，例如：

**企业内部数据。 **这包括大量不同的数据，例如用户信息、交易数据、调查数据、 Web 日志、网页内容等。要考虑现有数据是否足以满足需要。

**外部行业数据。 **如人口统计数据、 GDP 数据、天气数据、合作单位数据?

**互联网数据。 **为了更好地提升分析效果，还可以考虑通过网络爬虫等技术采集互联网数据。

基于已经梳理的数据资源，我们为了提高数据的效率，降低分析大数量的难度，我们首先应该想尽办法降低数据量，尽快排除与数据挖掘目标不相关的数据。通常，分以下几步选择数据：

第一步，按表进行筛选。首先从企业内部数据中选择与分析主题相关的表，然后考虑是否可以从外部行业数据中补充数据，最后再考虑从互联网中采集一些数据进行完善。不过要注意，互联网的数据质量比较低，一定要谨慎使用。

第二步，再对选择的每一张表，进行行和列的筛选。

选择属性或特征（列）：如交易频次、最近一次交易时间、交易金额等。选择合适分析对象（行）：如要包含哪些卡号、产品类型或日期时间。

????

经过上一步的处理，我们的数据量明显小了很多，为接下来的分析提高了不少效率。数据审核的工作是非常必要的，因为数据从来都不是完美的，只有保证输入的质量，才可以保证模型输出成果的可靠。在这一部分我们不仅要对有质量问题的数据进行处理，同时还要记录数据的基本特征。

需要查找以下这些问题：

- ○ ○ 数据是否缺失（例如 $null$、?）？如果有，是填充缺失值还是删除缺失值；
  - ○ 数据度量类型是否正确？比如日期格式是否正常，是否把销售额等数值字段存储为了字符型；
  - ○ 编码是否一致？例如同时使用 M 和 male 表示性别男；
  - ○ 字段取值是否异常？比如，购买金额是否为负，退货次数是否过多；
  - ○ 数据量多大？记下每张表的大小；
  - ○ 各变量分布特征如何？比如名义值个数，最大值最小值；

    - □□□□

????

□□□

一、敏捷挖掘流程

??

????

数据准备主要是面向模型而准备，当您制定了项目实施方案时，基本已确定分析思路和计划使用的算法。那么完成数据准备后，水到渠成的开始建模。

建模时通常会执行多次迭代。通常，数据挖掘人员会使用默认参数运行多个模型，然后再对这些参数进行微调或回到数据准备阶段以便执行所选模型所需的操作。仅使用一个模型且仅执行一次就能圆满地解答组织的数据挖掘问题，这样的情况几乎不存在。这就是数据挖掘如此有趣的原因，可以使用多种方法来考虑某个已知的问题。

??????

尽管可能已经知道哪种类型的建模方式可以尝试，但现在还无法确定使用哪个模型是最好的。通常，将会基于下列因素确定最适用的模型：

**可用于挖掘的数据类型。 **例如，感兴趣的字段是字符型还是数值型?

**数据挖掘目标。 **
是否能挖掘出令用户感兴趣的购买模式？或者是否需要生成一个得分，例如用于表明会员再次购买的的倾向?

**应用要求。 **模型是否准确并且稳定?

??????

一个好的模型不仅要准确更要稳定。为了最终测试模型的优劣，需要提前设计一个测试标准。主要包含两个部分：

- ○ ○ 描述模型的 “优异性 ”
  标准：模型的优异性可以通过多种方法度量。对于监督式模型，例如决策树模型和逻辑回归模型，优异性的度量方法通常是估计特定模型的错误率。对于非监督式模型，例如聚类模型，度量方法可以包括易于解释、部署或所需处理时间等标准。
  - ○ 定义将要对其测试这些标准的数据。通常使用分区节点将数据分成训练集和测试集，训练集用于训练模型，测试集用于测试模型。另外，采用交叉验证方法测试模型的稳定性。

记住，模型构建操作是一个迭代的过程。这意味着通常需要测试多个模型的结果才能决定使用和部署哪些模型。

????

□□□

一、敏捷挖掘流程

??

????

现在，数据挖掘项目已经完成了一大半。而且，根据之前定义的数据挖掘成功标准，还确定在建模阶段构建的模型从技术上说是正确而且有效的。但是，前面的模型还不能直接运用到商业中去。在建模阶段的模型评估主要基于技术角度，

一个有着很强拟合能力的模型不一定有着同样好的预测能力。众所周知，数据分析中常常会遇到“过拟合”的问题，我们做数据挖掘的目的不仅是在总结过去的特征或者规律，更主要是为了用于预测未来，那么我们通过挖掘技术从历史信息中挖掘出的规律应该是既适用于过去，又适用于现在，还同样适用于未来。如果我们对历史信息“过度总结”，那很可能会造成我们挖掘出了一些规律只代表过去的特征，这就会导致用这样的规律去拟合历史的数据会很准确，却使得预测未来的效果很差，这就是“过拟合”现象。另外，我们在实际应用中，还可能会遇到模型“不稳定”，也就是有时模型预测准确，有时差距很大，这种“不稳定”导致我们无法信任模型。

因此，在继续之前，还应该使用在项目开始时设立的业务成功标准评估模型。这是确保实现业务目标的关键环节。

????

在这个阶段，将对项目结果是否达到业务成功标准的评估进行规范。此步骤要求对声明的业务目标有清晰地了解，因此请确保在项目评估时包含关键决策制定者。在这里主要回答以下几个问题：

三个模型中哪个模型最优？

选择使用单一模型部署还是使用混合模型部署？

单一模型和混合模型各自的优势和不足分别是什么？模型未来的风险有多大，对业务影响有多大？

应对措施是什么？

将对业务有多大改善？

今后如何改善模型？

??????

至今为止，已经生成了结果并且评估了数据挖掘过程，此时应该考虑，接下来该做些什么？这一阶段将根据数据挖掘业务目标回答上述问题。实际上，此时具有两个选择：

一是，继续进入部署阶段，将模型结果并入业务过程中，从而生成最终的报告。

二是，返回到前面的步骤改进或替换模型。如果发现结果几乎可以算是（但并非）最佳结果，可以考虑另一轮建模。可将在此阶段中了解到的信息用于改进模型并生成更好的结果。

此时的决定关系到建模结果的准确度和相关性。如果结果实现了数据挖掘目标和业务目标，则已经可以进入部署阶段。不论作出什么决定，都请确保记录下完整的评估过程。

一、敏捷挖掘流程

??

??

数据挖掘可以生成两种类型的成果：在前期阶段中选定的最终模型；从模型本身以及数据挖掘过程中得出的任意结论或推论，包括派生的一些重要的 KPI指标、数据质量或业务流程优化方案、未来规划等。

成果部署的方式有两种：一是撰写研究报告；二是开发 WEB 应用系统，通过 PC、IPAD 或移动端进行分享。第一种较为简单，一般是针对一个具体的阶段性问题进行研究，最终形成分析报告就行了。第二种稍微复杂，但是一般重要的项目最终都会开发成智能的系统，这样的系统有诸多好处：

一是，方便用户浏览数据挖掘成果，还可以开发计划任务，自动化完成成果应用。比如，我们可以为百货商场建设一个自动营销系统，只要用户指定促销的产品，系统可以自动评估并输出精准营销名单，并按照模版通过短信通知会员；二是，可以结合人工智能的思想，设计系统的自学习能力，不断优化模型；

三是，方便用户对数据挖掘成果进行可视化评估，降低模型应用风险；

四是，系统可以建立高扩展机制，不断扩展系统的能力，轻易集成今后研究发现的更多成果，增加系统功能或者改善模型；

五是，有利于企业对积累的数据挖掘成果进行规范化管理，并追踪挖掘成果的效益。

??????

在系统开发完成后，一般需要一个月的测试，和 3 个月到半年的试运行（根据项目的大小不同，这个周期也会也不一样，并不都需要这么长的周期）。在这期间要做好监控和维护。因此，要根据监控的结果编写监控报告和评估报告。另外，还要对项目的整个开发过程的资料进行整理备案，以便今后研究、系统维护和经验分享。

同时需要系统建模一个验证子系统，不断监控模型状态并追踪模型应用效果，有利于及时发现模型出现的新问题以及模型优化的新思路。

????

在系统试运行结束后，可以正式将挖掘成果运用到商业环境中，为企业创造价值。

??????

二、应用案例

??

本模块案例数据都是事先安装在云版服务其中的数据，通过数据库节点导入工作流，配置如下：

二、应用案例

????

??????

专家系统是典型的大数据和人工智能结合的计算机程序系统，其内部含有大量的某个领域专家水平的知识与经验，能够利用人类专家的知识和解决问题的方法来处理该领域问题。也就是说，专家系统是一个具有大量的专门知识与经验的程序系统，它应用人工智能技术和计算机技术，根据某领域一个或多个专家提供的知识和经验，进行推理和判断，模拟人类专家的决策过程，以便解决那些需要人类专家处理的复杂问题，简而言之，专家系统是一种模拟人类专家解决领域问题的计算机程序系统。

专家系统的发展已经历了 3 个阶段，正向第四代过渡和发展。第一代专家系统（ dendral、macsyma 等）以高度专业化、求解专门问题的能力强为特点。但在体系结构的完整性、可移植性、系统的透明性和灵活性等方面存在缺陷，求解问题的能力弱。第二代专家系统（ mycin、casnet、prospector、hearsay 等）属单学科专业型、应用型系统，其体系结构较完整，移植性方面也有所改善，而且在系统的人机接口、解释机制、知识获取技术、不确定推理技术、增强专家系统的知识表示和推理方法的启发性、通用性等方面都有所改进。第三代专家系统属多学科综合型系统，采用多种人工智能语言，综合采用各种知识表示方法和多种推理机制及控制策略，并开始运用各种知识工程语言、骨架系统及专家系统开发工具和环境来研制大型综合专家系统。在总结前三代专家系统的设计方法和实现技术的基础上，已开始采用大型多专家协作系统、多种知识表示、综合知识库、自组织解题机制、多学科协同解题与并行推理、专家系统工具与环境、人工神经网络知识获取及学习机制等最新人工智能技术来实现具有多知识库、多主体的第四代专家系统。

本案例使用数据云敏捷挖掘工具，以决策树算法为背景，通过生动有趣的过程讲解，帮助读者了解依托大数据如何构建专家诊病模型，以及如何通过可视化探索数据，实现决策树同样的计算结果!

该案例的工作流如下：

????

业务理解：该案例所用的数据是一份医生诊病的数据，如下：

表 **1**: 数据视图

其中，年龄、性别、血压、胆固醇、钠、钾是病人的指标，而药物是医生针对病人的情况开出的药物。

□□□

□□□

用的话，多次运行可以测试数据分布规律和模型稳定性。

再次，使用类型节点指定目标变量的角色，将药物的角色设为目标。

还有一点要注意 ，这里分区即把数据集随机分成两份，通常训练集要比测试集大，通常分为 5：5，6：4，7：3，8：2，9：1 。数据集越小，训练集应该分的更多，原因是要保证模型的稳定性，参与训练的数据要足够多。实际中，各种比例都会尝试，目的有二：一是选择最佳的比例，二是测试模型的稳定性。

然后，进行分类预测建模，因为目标变量（药物）为分类型。此处我们选择 C4.5 决策树算法。节点连接如上图 3 所示，在决策树（训练）节点配置中选择目标变量药物。其中，决策树（训练）节点连接分区节点。决策树训练节点可以采用默认配置，无需修改配置，如下：

最后，使用分类评估节点评估模型的准确性。配置如下图所示：

评估结果：

表 **3**：模型评估 **1**

从表 3 中可以看出，模型测试准确度为 97.6%，误判 2.4%。预建模的目的是为了从整体判断现有变量与目标变量的相关性，以便可以根据经验预估最终的效果和可操作性。

??????????

通常来说有三种办法：

**第一， **增加新数据，以便引入更多重要的影响因素；

**第二， **尝试其他模型，以便找到更适合的模型；

**□□□ **□□□

□□□

表 4: 模型评估 2

从表 4 中可看出，模型精度从原来的 80%提升到了 100%。当然实际中如果遇到预测精度为 100%的情况一定就是错的。下面我们再来解读一下得到的决策树模型。如下图所示:

从整体来看，得此类病的人有大约一半的人（ 45%）选择服用了 Y 药物

- ○ 如果病人血液中钠和钾的比例大于 14.8285，则选择服用 Y 药物，准确率为 100%;
  - ○ 如果病人血液中钠和钾的比例不大于 14.8285，若全部判成 X 药物准确率只有 47.5%，因此再看病人的另外一个指标血压;
    1. 如果血压（其值有 1、2 和 3）为 2，则全部判为 X 药物，准确率为 100%
    2. 如果血压为 1，全部判为 C 药物，准确率仅 50%。再看另外一个指标胆固醇
       1. 若胆固醇的值为 1，全部判为 X 药物，则准确率为 100%。
       2. 若胆固醇的值为 2，全部判为 C 药物，则准确率为 100%
    3. 如果血压为 3，全部判为 A 药物，则准确率仅 56.8%。再看另外一个指标年龄
       1. 若年龄大于 50.5 岁，判为 B 药物，则准确率 100%
       2. 若年龄不大于 50.5 岁，判为 A 药物，准确率为

100%。

??

第一，在建模中，应该注意过拟合问题。在商业中建模的目的是为了商业应用，因此不仅要保证模型的准确性，更要保证模型的稳定性。

第二，模型应用的核心是模型的风险控制。只有可以控制风险的模型才可以使用，否则不管理论模型建的多么漂亮都是没有实际意义的。

第三，数据挖掘的成果并不只是数学模型，数据挖掘也不等价于高级模型（如决策树、神经网络等）。数据挖掘最重要的不仅是得到一个模型，还有分析得出最合理的输入变量以及给出数据质量管理的建议。

第四，数据挖掘的一个核心工作就是图形化探索，貌似很简单却又最为重要。因为这是你数据挖掘思路的源泉。

??????

我们将通过一个信用评分的案例来一起学习基于数据挖掘技术来解决信用风险评估问题的基本思路和方法。

□□□

□□□

第三步，以违约为目标字段，使用 C4.5 决策树算法，可以使用默认配置，节点连接如上图所示。

第四步，使用评估节点评估模型。模型评估如下图所示，模型测试整体的准确度为 83%，测试用的个案一共是 700 个，训练模型对这些个案判断正确 581 个，错判 119 个。

??

第一步，连接之前的选择节点的第二个输出端口筛选出待判的 150 人。

第二步，使用概率 C4.5 决策树（预测）节点基于训练模型预测这 150 人。节点配置如下：

预测结果如下：

第三步，使用过滤节点对系统生成的字段重命名。上表中的 表示用户不违约的概率，

为用户违约的概率。因此对这两个字段重命名如下：

第四步，使用 Java 代码段节点（也可以使用派生节点）生成风险得分字段。配置如下：

第五步，获取高风险用户名单。

使用排序节点按风险得分降序排列。配置如下：

使用选择节点选择风险得分大于 60 的用户。配置如下：

使用表格节点查看高风险用户名单。如下：

类似方法，可以获取高信用用户名单。

??

首先，通过该案例我们掌握了如何建立信用评分模型，当然此处使用的分类预测算法并不局限于 C4.5 决策树算法，只要是适用于目标变量为字符型的分类预测算法都可以（如神经网络、支持向量机、贝叶斯网络、KNN、Logistic 回归等）。

其次，从该案例中我们还可以体会到有模型预测的概率比预测结果更重要，在该案例中我们基于违约概率和不违约概率分别得到了信用得分和风险得分的计算方法。

??????

客户挽留在很多行业都是一个备受关注的问题，比如电信、银行、保险、零售等。要做客户挽留就需要对客户流失进行预警、客户流失原因分析、客户满意度或忠诚度研究、客户生命周期研究等相关问题进行深入而全面的分析。例如，对客户的行为特征进行分析，可以了解有多少客户流失，客户是什么时候流失的，以及客户是如何流失的等问题，从而监控客户流失、实现客户关怀。

应用数据挖掘技术可以根据过去拥有的客户流失数据建立客户属性、服务属性和客户消费数据与客户流失可能性关联的数学模型，找出客户属性、服务属性和客户消费数据与流失的关系，给出明确的数学公式或规则，从而计算出客户流失的可能性。

□□□

业务理解：该案例所用的数据是一份关于电信客户流失的数据，如下：

表 **1**: 数据视图

其中，该数据表中一共有 42
个字段，而流失字段记录了用户是否已经流失。业务目标：建立交互式存量维系营销系统，系统会定期输出流失风险较大的

用户名单以及这批用户的特征，为用户的维系提供决策支持。

数据挖掘目标：建立现存用户的流失风险评估模型，该模型以用户的信息指标为输入，以流失为目标，建立预测评估模型，该模型可以根据输入指标的值，计算预测值（流失）及流失概率。

????

这里主要介绍如何优化输入。

第一步，使用 数据库 导入节点读取数据。配置如下：

第二步，优化输入。使用特征选择节点过滤不重要的字段或者数据质量不好的字段。

过滤之后原表中的数据剩余 11 列。如下所示：

??

这一部分的目标是建一个可靠的预测模型。该模型可以评估客户流失的概率。

第一步，使用类型转换节点将流失字段从数值型转化为字符型。节点配置如下：

第二步，使用分区节点建立训练集和测试集。节点配置如下：

第三步，以流失字段为目标变量，使用类型节点将流失字段的角色设为目标。配置如下：

第四步，使用逻辑回归节点训练模型。配置如下：

第四步，使用分类评估节点评估模型。

整体评估结果如下表所示，测试的整体准确率近 77.5%。在商业中，这样的准确度已经很不错了。

对保持用户预测能力评估如下表所示，测试准确率 89.9%，高于整体的准确率。

对流失用户预测能力评估如下表所示，准确率只有 44.8%

，远低于整体准确率，而实际中更关注模型对流失用户的预测能力，因此，从这里看该模型效果并不很满意。

综上所述，该模型并不十分理想。在实际中，一定要再做优化，此处就不再做优化了。

??????

如果为了挽留流失用户，则要生成流失用户名单。如果是为了奖励忠诚用户，则要生成忠诚用户名单。

[illegible] [illegible] [illegible]

[illegible]

[illegible]

[illegible]

[illegible] 70

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible] 80

[illegible]

[illegible]

[illegible]

再使用交互表格节点查询名单，预览如下：

??

通过该案例，首先，我们掌握建立客户流失模型的方法，以及如何输出高流

失风险的客户名单。其次，我们掌握了在输入变量较多的情况下，应该如何去筛选出重要的变量，以保证模型的效果。最后，我们还从该案例中体会到评估模型的预测能力，并不只看模型的总体的准确度，有时我们更关注模型对目标变量中某个类别的预测能力，比如在该案例中我们更关注的是模型对流失客户的预测能力，这种情况下我们就不能只关注模型整体的准确度，更关心模型对流失客户的预测准确度。

要想成功建立流失模型，其四个关键是：定义流失的商业含义、理解流失结果的使用、明确流失建模对数据的要求，以及设计对时间稳健的流失模型。当然，客户流失分析系统必须针对各种不同的种类分别定义业务问题，分别操作运行。如果运营商能够把握好这些要点，就非常容易将模型推广和应用到实际的营销当中，并且以概率为流失度量，还可以对营销数量进行灵活的控制。

二、应用案例

????

?????

购物篮指的是超级市场内供顾客购物时使用的装商品的篮子，当顾客付款时这些购物篮内的商品被营业人员通过收款机一一登记结算并记录。所谓的购物篮分析 (Market Basket Analysis)就是通过这些购物篮子所显示的信息来研究顾客的购买行为。主要的目的在于找出什么样的东西应该放在一起。藉由顾客的购买行为来了解是什么样的顾客以及这些顾客为什么买这些产品，找出相关规则，企业藉由这些规则的挖掘获得利益与建立竞争优势。举例来说，零售店可藉由此分析改变置物架上的商品排列或是设计吸引客户的商业套餐等等。

购物篮分析最主要的目的在于找出什么样的东西应该放在一起? 商业上的

应用在由顾客的购买行为来了解是什么样的顾客以及这些顾客为什么买这些产品，找出相关的联想（association）规则，企业由这些规则的挖掘获得利益与建立竞争优势。举例来说，零售店可由此分析改变置物架上的商品排列或是设计吸引客户的商业套餐等等。

??????????????????

1. 选择正确的品项：这里所指的正确乃是针对企业体而言，必须要在数以百计、千计品项中选择出真正有用的品项出来。
2. 经由对共同发生矩阵（co-occurrence matrix）的探讨挖掘出联想规则。
3. 克服实际上的限制：所选择的品项愈多，计算所耗费的资源与时间愈久（呈现指数递增），此时必须运用一些技术以降低资源与时间的损耗。

??????????????????

1. 针对信用卡购物，能够预测未来顾客可能购买什么。
2. 对于电信与金融服务业而言，经由购物篮分析能够设计不同的服务组合以扩大利润。
3. 保险业能藉由购物篮分析侦测出可能不寻常的投保组合并作预防。
4. 对病人而言，在疗程的组合上，购物篮分析能作为是否这些疗程组合会导致并发症的判断依据。

该案例通过一个简单的超市购物的数据来演示购物篮分析的方法和思想。工作流如下：

完成该案例需要以下几步：

第一步，读取数据 shoppingbasket。配置如下图：

第二步，剔除只有一种物品的购物篮子。

首先，使用统计节点统计 id 字段 (即为购物篮 id)值的个数。配置如下：

其次，使用选择节点筛选出计数 =1 的记录。配置如下：

过滤后的表一共有 174 行，这些都是购物篮中只有一种物品的情况。如下表所示：

第三步，生成购物篮。

首先，使用类型转换节点将 id 和卡号由数值型转化为字符型。

其次，使用汇总节点派生购物篮。注意，这里应该以 id 为购物篮 id ，而不是卡号，因为同一卡可能有多次购买。但是并不是不能以卡号为研究对象，如果您想把一个人作为一个购物篮对象，也可以用卡号，但是此处经分析重复购买的人并不多，因此以人为研究对象没有意义。节点配置如下：

再次，使用参考行过滤节点，从汇总表中过滤掉过虑表中的数据。配置如下：

再次过滤后的数据如下表所示：

第四步，使用 Apriori 节点生成训练规则。配置如下：

生成的规则预览如下：

第五步，派生规则质量字段。

首先，使用选择节点过滤前项为空的规则。配置如下：

然后，使用派生字段节点，派生规则质量字段。规则质量 ＝支持度 *规则置信度。配置如下：

最后，使用排序节点对数据按照规则质量和前项进行二重降序排列。排在前面的即为规则质量比较高的规则。

节点配置如下：

第六步，提取规则质量较高的规则，如下表所示：

规则解释如下：

如果客户购买啤酒和蔬菜罐头的话，他还有 87.4%的可能购买冻肉；如果客户购买啤酒和冻肉的话，他还有 85.9%的可能购买蔬菜罐头；如果客户购买蔬菜罐头和冻肉的话，他还有 84.4% 的可能购买啤酒；如果客户购买啤酒的话，他还有 62.7%的可能购买啤酒；

如果客户购买蔬菜罐头的话，他还有 60.5%的可能购买啤酒；如果客户购买鱼类的话，他还有 53.7% 的可能购买水果；

如果客户购买糖果的话，他还有 53.9%的可能购买啤酒；

……

**

小结：数据挖掘分为两种，一种是经验验证型，即您或者领导有个想法，然后让您验证一下是不是那样的，您分析之后的回答只有是或者不是，这属于验证型。还有一种是知识发现型的，即原来并不知道有这样的规律，分析之后发现的，即为意想不到的新发现。在这里，我们发现的规则中，

**

啤酒、蔬菜罐头、冻肉同时被购买的可能性较大，这一点可能是大家都知道的，这个结果只是验证了我们以前的想法，而鱼类和水果一起购买的可能性很大，我们可能就不知道了，因此

这就属于知识发现。因此，对于这家超市而言，应该把水果和鱼类排放的近一些，这样可以提高销量。还有一点要注意的是，数据挖掘的结果没有普遍适用性，对这家超市有这样的规律，并不代表其他超市也是这样。但是方法论是通用的，不同的超市可以使用同样的方法论进行知识发现和验证。

二、应用案例

????

????

聚类与分类的不同在于，聚类所要求划分的类是未知的。

聚类是将数据分类到不同的类或者簇这样的一个过程，所以同一个簇中的对象有很大的相似性，而不同簇间的对象有很大的相异性。

从统计学的观点看，聚类分析是通过数据建模简化数据的一种方法。传统的统计聚类分析方法包括系统聚类法、分解法、加入法、动态聚类法、有序样品聚类、有重叠聚类和模糊聚类等。

从机器学习的角度讲，簇相当于隐藏模式。聚类是搜索簇的无监督学习过程。与分类不同，无监督学习不依赖预先定义的类或带类标记的训练实例，需要由聚类学习算法自动确定标记，而分类学习的实例或数据对象有类别标记。聚类是观察式学习，而不是示例式的学习。

聚类分析是一种探索性的分析，在分类的过程中，人们不必事先给出一个分类的标准，聚类分析能够从样本数据出发，自动进行分类。聚类分析所使用方法的不同，常常会得到不同的结论。不同研究者对于同一组数据进行聚类分析，所得到的聚类数未必一致。

从实际应用的角度看，聚类分析是数据挖掘的主要任务之一。而且聚类能够

作为一个独立的工具获得数据的分布状况，观察每一簇数据的特征，集中对特定的聚簇集合作进一步地分析。聚类分析还可以作为其他算法（如分类和定性归纳算法）的预处理步骤。

聚类分析的核心思想就是物以类聚，人以群分。在市场细分领域，消费同一种类的商品或服务时，不同的客户有不同的消费特点，通过研究这些特点，企业可以制定出不同的营销组合，从而获取最大的消费者剩余，这就是客户细分的主要目的。在销售片区划分中，只有合理地将企业所拥有的子市场归成几个大的片区，才能有效地制定符合片区特点的市场营销战略和策略。金融领域，对基金或者股票进行分类，以选择分类投资风险。

下面以一个汽车销售的案例来介绍聚类分析在市场细分中的应用。

????

业务理解：数据名称：　　汽车销售

。该案例所用的数据是一份关于汽车的数据，该数据文件包含销售值、订价以及各种品牌和型号的车辆的物理规格。订价和物理规格可以从　　edmunds.com　和制造商处获得。定价为美国本土售价。如下:

表 **1**: 数据视图

**业务目标: **对市场进行准确定位，为汽车的设计和市场份额预测提供参考。 **数据挖掘目标: **
通过聚类的方式对现有的车型进行分类。

????

通过数据探索对数据的质量和字段的分布进行了解，并排除有问题的行或者列优化数据质量。

第一步，我们使用统计节点审核数据的质量，从审核结果中我们发现存在缺失的数据，如下图所示:

第二步，对缺失的数据进行处理，我们选择使用缺失填充节点删除这些记录。配置如下:

??

我们选择层次聚类进行分析，尝试根据各种汽车的销售量、价格、引擎、马力、轴距、车宽、车长、制动、排量、油耗等指标对其分类。

建模前，需要对数据的每个字段进行计算，或更新字段，包字段测量类型、值域和缺失等，这时需要 类型节点，配置如下:

因为 **K-means**不能自动确定分类数量，因此需要我们以自定义的方式规定最后聚类的类别数。 **K-means**节点配置如下（默认配置）:

可以使用交互表或者右击 K-means 节点查看聚类的结果，如下图所示:

查看聚类概述图，了解每个聚类簇的分布：

结果如下：

从图中可见，分成的五个类样本数差异太大，cluster_2 和 cluster_3 包含的样本数都小于 20，这样的分类是没有意义的，因此需要重新分类。

我们尝试在 K-means 节点的配置中，重新指定分类数。新的聚类样本数分布如下：

对数据初步确认好分类后，可以分类查看数据的各项指标，来评估分类是否达到最终目的。

如我们看一下每类的销售量分布情况，配置如下：

每类汽车的销售量分布：

??

通过这个案例，大家可以发现聚类分析确实很简单。进行聚类计算后，主要通过图形化探索的方式评估聚类合理性，以及在确定聚类后，分析每类的特征。

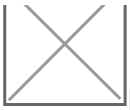
????

下面这个案例主要介绍如何对媒体进行分级。分级不同与分类，通过聚类分析将物或人分成几类，只是根据相似性定义的，没有哪类好哪类坏之分。而分级则是定义优劣级别。不过案例对媒体分级的方法使用的仍然是聚类算法，也只有在特定场景下，才可以通过聚类对对象进行分级。下面讲述的案例是关于一家专

门做网络游戏广告投放的广告公司。这家广告公司有

万多个网络广告位，帮助游戏公司投放广告，以往都是由他们的媒介专家（媒介经理或者媒介总监）选择满足

□□□



cluster_1 cluster_3 cluster_0 cluster_2 cluster_4

_1,3,0,2,4_ _A_ _B_ _C_ _D_ _E_

??

□□□

□□□

????????

□□□

第五步，使用直方图看偏移的分布趋势，直方图配置如下：

从执行结果可以看出，偏移字段有少数分布在尾部，这些即是偏移较大的对象，也就是可疑对象，如下图所示：

第六步，输出可疑名单。

首先，使用选择节点筛选出偏移大于 20 的记录。配置如下：

其次，使用交互表格节点查看可疑名单。如下：

第七步，分析偏移的分布，以及与申请类型的相关性。

在这里，我们使用直方图节点可视化发现，节点配置如下所示：

由执行后的结果可见，申请类型为 B 的偏移都在 20 以内（如下图所示），这说明这类申请类型应该不存在欺诈，由此可以再从业务中印证是否这种申请类型的规定没有漏洞，本身就无漏洞可钻，如果确实是，我们就没有必要再分析这种申请类型。

图 **2**：偏移直方图分布

????????

□□□

第六步，可视化分析。

首先，使用直方图节点可视化发现申请金额差异较集中

第七步，生成可疑名单。

首先，使用选择节点筛选出申请金额差异大于 20 的住户。节点配置如下:

其次，使用交互表格节点查询可疑名单，如下所示:

表 **4**: 可疑名单 **3**

???????

第一步，使用异常节点找出异常值。节点配置如下:

第二步，名义值过滤，配置如下:

第三步，生成可疑名单 4。

□□□

二、应用案例

????

时间序列预测法是一种历史数据的延伸预测，也称历史引伸预测法。是以时间数列所能反映的社会经济现象的发展过程和规律，进行引伸外推，预测其发展趋势的方法。如下图（用时间序列预测 2017 年 12 月销量）：

??????

下面我们结合时间序列数据演示基于历史数据挖掘规律并进行预测的流程。在数据云中，工作流程如下：

引入数据

数据准备

建模分析

该案例所用的数据是销售数据。这是一个包含销售日期、销量和利润（万元）三个指标的数据，具体内容在软件中打开显示如图 9.5 所示：

????

建立数据库节点

，导入要分析的历史数据（新建工作流，拖拽数据库进入工作流，右键配置数据库节点，配置数据路径，引入源数据，具体参考帮助手册）

通过配置数据库，导入元数据 —“销售数据 ”。在软件中打开数据，如下图：

????

任何数据都不是完美的，因此在建模分析之前，需要做数据清洗，数据清理好之后，我们需要了解数据的规律，以帮助我们建立预测模型。

二、应用案例

????

数据清洗和处理，包括缺失值填充、异常值分析、筛选所需时间区间、按周期汇总数据、日期变量的派生等。

此案例，数据表中存在 2 个问题：

- ○ ○ 日期字段是字符型，需要将此字段转换成日期型数据。
  ○ 个别数据时间过早，对预测的数据指导意义不大，需要对数据筛选。

下面就根据案例介绍数据准备常用的节点和方法。转换日期数据的字段类型（如下图）：

- 数据准备 节点中选择 类型转换 ，并拖拽至工作流中，连接输入输出箭头
- 右键 配置， 选择 字符串到期

同样的方法，建立 时间选择 节点，选择 85 年 -95 年数据。

数据清洗的工作基本完成。但是在数据进行建模之前，还需要指定模型预测所需要的时间区间。

??????

使用 时间区间　　节点指定预测模型的时间间隔。

时间区间节点可以为时间序列建模节点或自相关图节点的时间序列数据指定时间间隔并生成标签，作为未来估计或预测的基础。目前支持的时间间隔包括年、季度、月、日和时、分、秒。

时间时区节点适用于以下两种情况：

1. 数据表中不含时间字段：

数据表虽然并不包含时间字段，但该序列的数据已按照时间顺序按照等间隔排列，则可通过时间区间节点为数据自动添加日期标签。如果某个时点的数据存在缺失，建议增加空行表示此次度量或提前使用缺失值替代的方法进行填充。这种情况下选择　　“从第一个记录开始附加标签　　”。

1. 数据表中已包含时间字段：

对于这样已经含有时间字段的数据，需要对数据按照指定时间间隔进行转换，例如，将小时数据整合成以天为间隔的数据。由于数据中已经存在时间字段，不等间隔的情况允许存在。这种情况下选择　　“从数据构建　　”。在进行整合时，需要对数据进行汇总，汇总值的操作根据具体的分析需要进行，以确保对各种字段类型按照指定间隔进行汇总的方法。

当然，　　1 和　2 也可以结合使用，先生成时间标签，然后按照用户所需要的时间间隔进行数据汇总。

表　9.3 时间区间节点配置

对话框选项	内容
时间区间	在选择框中指定用于构建或标注时间序列的间隔和 周期　（支持的时间间隔包括年、季度、月、日和时、分、秒　　）
从第一个记录开始附加标签	从用户指定的日期开始以指定的时间区间为间隔连 续为记录添加标签。例如，如果选择的时间区间是月，指定的日期 表示第一条记录是某年和某月，之后以一个月为间隔，为下一条记 录做标签，直到遍历整个数据表。输出表中将增加新生成的日期标 签。

| □□□□ | □□□□

□□□

对话框选项	内容
ACF 和 PACF 最大延迟数	用于指定计算 ACF（auto correlation function，自相关）和 PACF（partial auto correlation function，偏自相关）时最大的延迟期数。一般为序列周期的倍数。
时序周期	可以有用户指定时间序列的周期，或者由系统自动识别

时间序列有三个部分构成，平稳性，趋势性和季节性，根据预测时考虑到的时序的影响因素的不同，影响方式不同和平滑的次数不同，从简单指数平滑衍生出更多中模型：

指数平滑模型算法选择：

- 当数据比较稳定，没有较大波动，可以采用“自动”
- 如果数据存在线性趋势且没有季节性，建议“Holt”
- 如果存在显著季节性且趋势性不明显，建议“简单季节”
- 如果同时存在线性趋势和季节性，建议“Winter 加法”或者 Winter 乘法”

????

实际操作时，建立模型后，可以尝试运行不同的模型，对模型评进行估对比，从而选择最佳的模型。

指数平滑节点执行后，输出时间散点图，残差自相关图，模型评估和预测数据，同时也支持 PMML 模型查看和导出。

?????

时间散点图展示了使用指数平滑法的预测值与实际观测值的对比以及每个预测值的置信区间。主要是看模型拟合效果。

??????

关于残差自相关图的解读我们会在 XX 具体解释。

????

□□□

????

业务理解：该案例所用的数据是《中国纱年产量 .xls》、《奶牛月产奶量 .xls》和《北京市每年最高气温 .xls》，三组数据均包含年份 /月份、产量 /最高气温两个指标，如下：

??????

以中国纱年产量的工作流为例：

分析时间序列趋势是时间序列建模的初步工作。目的是为了分析时间序列的规律和序列特征，如序列是否平稳？是否有周期性或者趋势性？并由此分析哪些分析模型更适合。

操作如下：

第一步，设置时间区间。节点配置如下：

时间区间设为年，从第一个记录开始附加标签，因为数据本身就是年度数据，所以没有进行汇总设置。执行后生成的时间区间表如下：

第二步，使用　　趋势图　　节点分析序列的趋势，节点配置如下：

从下图可以看出，序列整体呈上升趋势，且没有周期性。

另外两组数据的节点配置与之类似，不再赘述。配置完成并运行后得到的奶牛月产奶量的趋势图如下，序列整体呈上升趋势，且具有周期性：

北京市每年最高气温的趋势图如下，序列整体没有明显的周期性和趋势性，围绕某个常数上下波动：

???????

仍以中国纱年产量数据为例，工作流如下：

第一步，设置时间区间。具体操作参见上一节。

第二步，使用自相关节点做时间序列图检验，设置如下：

差分阶数默认为　　0，表示不进行差分。勾选　　ACF，表示绘制自相关图，没有勾选　　PACF，表示不绘制偏自相关图，也可以勾选。由如下相关图可以看出，自相关图既不是拖尾也不是截尾（截尾是指

在某阶之后，系数都为 0；拖尾则会

有一个衰减的趋势，但是不都为 0

），而是一个三角对称的形式，这种趋势是单调趋势的典型图形，所以原时间序列不平稳。

另外两组数据的自相关节点配置与之类似，不再赘述。配置完成并执行后分别得到如下的相关图。下图是奶牛月产奶量的自相关图和偏自相关图，自相关图既不是拖尾也不是截尾，而是正弦曲线的形式，这种趋势是周期性趋势 的典型图形；偏自相关图同样既不是拖尾也不是截尾， 所以原时间序列不是平稳序列。

下图是北京市每年最高气温的自相关图和偏自相关图，自相关图和偏自相关图都是拖尾，所以原时间序列是平稳序列，可以使用 ARMA 模型进行时间序列分析。

??????

该案例的工作流如下：

????

业务理解：该案例所用的数据是《销售数据 .csv

》，该数据包含销售日期、销量和利润（万元）三个指标，如下：

表 **1**：销售数据预览

业务目标：建立预测分析系统，分析利润在未来的预测值。

数据挖掘目标：建立时间序列预测模型，该模型以日期为输入，以利润为目标，建立预测模型。

????

使用统计节点计算字段的特征统计量，输出结果如下：

??????

分析时间序列趋势是时间序列建模的初步工作。目的是为了分析时间序列的规律和序列特征，如序列是否平稳？是否有周期性或者趋势性？并由此分析哪些分析模型更适合。

操作如下：

第一步，使用类型转换节点将字符型的日期型转化为日期型。节点配置如下：

第二步，使用时间选择节点选择选择 85 年 - 95 年的数据的进行分析。节点配置如下：

第三步，设置时间区间，节点配置如下：

第四步，使用时序图节点分析序列的趋势，节点配置如下：

从下图可以看出，序列整体呈下降趋势，且没有周期性。

??????

工作流如下：

接下来，我们选择一个　　指数平滑　　预测算法进行预测。步骤如下：

第一步，使用时间区间节点设置时间区间为　　“月 ”，汇总配置中按照月对数据进行汇总，配置如下：

该节点执行前数据预览如下：

执行后数据预览如下：

第二个，使用　　类型 节点，制定输入变量，目标变量，为建模做准备，配置如下：

第三步，使用指数平滑预测算法，进行建模。节点配置如下：

执行后预览数据如下：

查看残差相关图：

查看预测结果趋势图：

如上图所示，指数平滑算法对该时间序列的趋势有一定预测能力，该模型也总结一些规律，但是尚有一些规律并未总结出。这是一个趋势性非平稳序列，您还可以考虑使用 ARIMA 模型重新建模。

??

时间序列分析是数据挖掘最难的算法之一。当然它也有一个好处，就是建模之前的数据量一般都很小，一般只有几十条到几百条记录（如果您的预测粒度是月，那么每年才 12 条数据， 10 年才 120 条），因此模型计算很快。

时间序列分析的一般步骤是：首先，探索序列的趋势，看看序列是趋势特征还是周期性特征，平稳序列还是非平稳序列，是白噪声序列，还是非白噪声序列。根据这些信息我们可以把分析尝试的模型范围缩小，有方向的去优化模型。其次，开始使用工具去尝试模型，并解释模型的质量。最后，评估模型。以及考虑是否使用混合模型来提高模型的稳定性。

二、应用案例

????

????

文本数据挖掘 (Text Mining)
是指从文本数据中抽取有价值的信息和知识的计算机处理技术。顾名思义，文本数据挖掘是从文本中进行数据挖掘 (Data Mining)。从这个意义上讲，文本数据挖掘是数据挖掘的一个分支。

它在商业智能、信息检索、生物信息处理等方面都有广泛的应用，例如，客户关系管理，自动邮件回复，垃圾邮件过滤，自动简历评审，搜索引擎等等。

文本挖掘有三个独特的关键步骤：

- ○ ○ 分词，通过分词可以将非结构化数据转化为结构化数据
  - ○ 情感分析，定义情感强度，从文本中挖掘作者的情绪或者从评论中挖掘用户的满意度或观点
  - ○ 去噪，即从分出的众多词中剔除干扰的无用的词，提升数据质量。下面结合一个小的案例演示文本挖掘的实现过程。工作流：

????

该案例有两个工作流，使用的是 textmining
文件夹的数据。上面一个工作流使用的是已经分过词的结构化数据，只做词云展示，数据名称为《大连旅游度假关键词频数 .csv
》。下面一个工作流使用的三个文本数据，展示了从读取到分词，再到词云展示，最后又做了情感分析，使用的数据名称为《 yantaidujia.txt》、

《 yantailvyou.txt》、《 yantaiyanglao.txt》。

?? 1

读取数据后，首先使用选择节点筛选高频词，配置如下：

其次，使用词云节点展示词云，配置如下：

展示结果如下：

从词云中，我们可以看出，词云中地名除了有大连之外，还有青岛、张家界、江南、桂林、哈尔滨、香港等。这说明提及大连的人中有很多都同时提到了这些地名，如此可以看出，大连在旅游方面竞争力不够强，同时被提到的地方主要都是竞品。

?? 2

首先，使用文本挖掘目录下的平面文件节点读取数据，配置如下：

读取的文本数据，如下图所示：

其次，使用中文分词节点对 Document 列的文本进行分词，配置如下：

可以通过词云查看分词后结果。

最后，使用情感分析节点分析情感强度，节点配置如下：

执行后，可以查看情感分析表。

??

通过这两个小示例，我们掌握了文本挖掘的一些基本方法，比如两种做词云的方式，以及如何进行中文分词和情感分析。

????

1. 数据理解

该案例使用的数据是 demo 文件夹下的《疾病信息 .xls》，包含日志和类型两列数据，数据如下：

??

该案例的工作流如下：

该工作流从读取 Excel 文件开始。首先使用分词对文档进行分词，并使用一系列的手段对词进行过滤。然后，从这些词中提取出最关键的词，并基于关键词生成标志型字段。这些字段将会作为针对文档进行分类预测的影响因素，之后以这些指标为输入，以文档类别为目标建立文档的分类预测模型。

读取数据后，首先使用分词节点对日志字段的记录进行分词，配置如下：

然后，对噪声词进行过滤，配置如下：

执行后，得到如下的过滤表：

之后进入数据准备阶段。使用一系列手段对词进行过滤，具体的工作流如下：

第一步，将所有字符都转换为小写，配置如下：

第二步，过滤不需要的字段，配置如下：

生成的选择表如下：

第三步，对分词字段进行汇总，配置如下：

第四步，选择计数大于 15 的记录，配置如下：

生成的选择表如下：

对于词云展示来说，数据准备进行到这一步就已足够，词云图如下：

□□□

二、应用案例

??????

????

通过该案例，我们将会学习如何对微博数据进行分析。将要学习和掌握的技术有文本挖掘技术、社会网络分析技术、集成 R 算法的能力、基于矩阵的层次聚类算法等。在学习该案例前首先需要了解以下基本概念。

????????

- 社会网络是指社会行动者及其间的关系的集合。也可以说，一个社会网络是由多个点（社会行动者）和各点之间的连线（行动者之间的关系）组成的集合。用点和线来表达网络，这个是社会网络的形式化界定。
- 这里的行动者不但指具体的个人，还可指一个群体、公司或其他集体性的社会单位。每个行动者在网络中的位置被称为“结点 (node)”。

???????????

- 亲属关系：父母、子女、夫妻关系等。
- 正式关系（权威关系）：正式角色也是关系性的，如老板/职员、教师/学生、医生/病人关系等。
- 个人之间的评价关系：喜欢、信任、尊重等。
- 行为上的互动关系：行动者之间的自然交往，如谈话、参加会议、拜访、提建议等。
- 隶属关系：如参加一项协会、属于某些俱乐部等。
- 物质资本的传递：商业往来、物资交流。
- 非物质资源的转换关系：行动者之间的交往、信息的交换等。
- 空间关联：城市之间的关系，迁入和迁出。
- 职位的升迁，地位的流动。

???????????

社群图：用于表示一个群体成员之间的关系，由点和线连成的图。根据不同的标准，社群图的种类也不同。

- 有向图、无向图：根据关系的方向进行划分。
- 二值图、符号图、赋值图：根据关系紧密的“程度”进行划分。
- 完备图、非完备图：根据网络中各个成员之间联系的紧密度进行划分。如果一个图中的任何两点之间都相连，则称为完备图，否则为非完备图。完备图很少见，一般图都是非完备图。
- 矩阵：矩阵中的行与列都代表“社会行动者”，即图中的各点。行与列对应的要素代表的就是各个行动者之间的“关系”。

????????

□□□

长度和距离:

一个途径的长度用构成该途径的线的条数来测量。如途径 1-4-5-6 的长度为 3。

两点之间的距离指的是连接这两点的最短途径的长度，如点 1 和 6 之间有很多途径，其最短途径 1-6 的长度为 1，故 1 与 6 之间的距离为 1。

距离矩阵:

??????????

与“关联性”有关的概念:

子图: 一个图 G 的子图 Gs 的定义是, Gs 中的点集（记作 Ns）是 G 的点集（N）的一个子集，并且 Gs 中的线集（Ls）也是 G 的线集（L）的一个子集, Gs 中的所有线也必须是在 G 中的所有点之间的线。

关联图和成分: 对于一个图来说，如果其中的任何两点之间都存在一个途径

（Path），则称这两点是相互可达的，称该图时关联图（connected graph）。也就是说，关联图中的任何两点之间都是可达的。如果一个图不是关联的，就称之为“不关联图”。一个“不关联图”，可以分为两个或者多个子图，我们称之为关联子图。一个图中的各个关联子图都叫做“成分”（components），它是最大的关联子图。也就是说，“成分”内部的任何点之间都存在途径。但是，成分内部的一点与任何外在于该成分的点之间都不存在任何途径。一个关联图仅仅由一个单一成分构成。其他图主要由一个或多个独立的成分以及一系列孤立点构成。

上图中共包含 4 个成分: 2 个关联成分和 2 个孤立成分。

切割点: 切割点是指如果去掉这个点，就会增加成分数目的点。没有切割点，子图就会分为两个或多个独立的子群，并且各个子群之间无关联。

?“??”?????

□□□

?“???”?????

“中心性　　”的研究意义：　　“权力　”

在社会学中是一个非常重要的概念。一个人之所以拥有权力，是因为他与他者存在关系，可以影响他人。在一个群体中，我们如何去界定某个人的权利大小？社会网络学者就从　　“关系　”的角度出发，用　　“中心性　　”来定量研究权力。人或者组织在社会网络中具有怎样的权力，或者说居于怎样的中心地位，这一思想是社会网络分析者最早探讨的内容之一。

点度中心度：刻画的是行动者的局部中心指数，测量网络中行动者自身的交易能力，没有考虑到能否控制他人

中间中心度：研究一个行动者在多大程度上居于其他两个行动者之间，因而是一种　　“控制能力　　”指数

接近中心度：考虑的是行动者在多大程度上不受其他行动者的控制

?“????”?????

大体上说，凝聚子群是满足如下条件的行动者子集合，即在此集合中的行动者之间具有相对较强的、直接的、紧密的、经常的或者积极的联系。

研究意义：通过对社会网络的凝聚子群的分析，可揭示社会结构，量化结构。派系：在一个图中，　　“派系　”指的是至少包含三个点的最大完备子图。

- 派系的成员至少包含三个点；
- 派系是　“完备　”的，即其中任何两点之间都是直接相关，都是邻接的；
- 派系是　“最大　”的，其含义是，我们不能向其中加入新的点，否则将改变　　“完备　”这个性质。

对社会网络分析的基本概念有所了解后，下面我们开始学习这个案例。工作流如下：

????

业务理解：数据名称：中国最强音微博。该案例所用的数据是中国最强音官方微博数据。其中，字段包括 Author、Weibo、Forward、Time_Weibo、Time_Search、Count_Reply。如下：

表　**1**：数据视图

业务目标：寻找意见领袖，为自媒体传播提供参考，另外还要对语义进行聚类，挖掘特征相似用户群体。

数据挖掘目标：建立社会网络图，挖掘意见领袖和微博传播路径，再使用聚类算法对语义进行聚类。

????

第一步，使用数据库节点读取数据，配置如下：

第一步，使用 R 代码段节点，借助 R 的编程能力利用关键字提取提及的微博用户。代码如下：

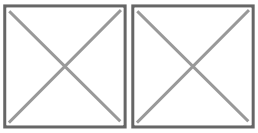
```
smartR.out<-smartR.in

library(data.table) library(stringr)

At<-str_extract_all(smartR.in$"Weibo","@\\b\\w*\\b")

textnumber = rep(1:nrow(summary(At)),as.numeric(summary(At)[,1])) smartR.out<-data.frame(
提及用户 =unlist(At),句子 =textnumber)
```

节点配置如下：

第三步，使用汇总节点，以句子为组，以提及用户为选项建立列表。节点配置如下：

第四步，使用 Apriori 节点通过关联分析算法训练规则。节点配置如下：

计算生成的结果如下：



?????

关联规则是形如 X→Y 的蕴涵式，其中， X 和 Y 分别称为前项和后项 。其中，关联规则 XY ，存在支持度和置信度。

假设

是 项 的集合。给定一个交易数据库 D，其中每个 事务 (Transaction)t 是 I 的非空子集，即，每一个交易都与一个唯一的标识符 TID(Transaction ID)对应。关联

规则在 D 中的 支持度 (support)是 D 中事务同时包含 X、Y 的百分比，即概率; 置信度 (confidence)是 D 中事务已经包含 X 的情况下，包含 Y 的百分比，即条件概率。如果满足 最小支持度阈值 和 最小置信度阈值 ，则认为关联规则是有趣的。这些阈值是根据挖掘需要人为设定。

第五步，派生规则质量字段。我们定义规则质量 =支持度 *规则置信度，配置如下:

第六步，对前项和规则质量字段按照降序进行二重排序。配置如下:

第七步，筛选规则质量在 20 以上的记录。配置如下:

第八步，使用 派生 节点派生边 id 字段。

经过以上八步处理结果如下表所示:

□□□

□□□

□□□

然后，使用参考行过滤节点将词性为名词和自定义的记录筛选出来。配置如下：

第四步，数据转换。

首先，使用过滤节点删除字段 Auther、计数。

其次，使用旋转节点生成以 MID 组，分词为支点，词性为选项的关联矩阵。节点配置如下：

然后，使用 过滤 节点修改字段名称。配置如下：

第六步，使用 K-means 聚类节点进行聚类。执行后，可右键该节点查看树状图。节点配置如下：

聚类结果和视图分别如下：

??

□□□