CRISP-DM 1.0

Cross Industry Standard Process-Data Mining 1.0

数据挖掘指导手册

本书是一部有关跨行业的数据挖掘标准程序（以下简称CRISP-DM）模型的书籍，主要包括以下几个部分：CRISP-DM方法论，CRISP-DM参考模型，CRISP-DM用户指导，CRISP-DM报告的书写以及相关帮助信息的附录部分。

前言

1996末，在当时尚为年轻和不成熟的数据挖掘市场中，三位市场上的“老战士”设想、构思了CRISP-DM 。DaimlerChrysler公司 (后名为 Daimler-Benz)先于各个工商业组织,早已经在其商业运作中成功地运用了数据挖掘。SPSS 公司(后名为 ISL)早在1990年就开始提供基于数据挖掘的服务，并于1994年开发了第一个数据挖掘的工作平台——Clementine. NCR公司建立了包括数据挖掘咨询顾问以及技术专家在内的团队，为客户提供咨询服务，并把它作为旨在为其使用Teradata数据集的客户提供增值服务的一个组成部分。

那个时候，对数据挖掘爆炸式的广泛理解表明了当时的市场对数据挖掘的初步兴趣。这既令人兴奋，又让人有所顾虑。我们按照我们的方式逐步的去理解和开发数据挖掘。然而我们做得是否正确？是否每一个数据挖掘的初学者都要像我们那样经过不断的尝试和失败去学习它？从一个提供者的角度而言，我们怎么向一个有所预期的顾客展示：数据挖掘已经足够成熟，可以作为他们商业操作中的一个关键因素？

我们推断，一个非私有的、公开的标准程序模型，无论是对我们、还是所有从业者而言，都会涉及到上面提到的问题。

一年后，我们成立了委员会，创建了这个以CRoss-Industry Standard Process for Data Mining 首字母缩写的名字，并获得了欧洲委员会提供的基金，开始实施我们最初的想法。我们旨在使CRISP-DM成为一个在工业运用、工具性以及应用方面都没有偏颇的中立性模型，因此我们不得不获得来自尽可能宽泛领域内的从业者们的帮助（例如，数据集的提供者和管理顾问），同时他们还要对数据挖掘具有一定的兴趣。为了获得这些帮助，我们成立了CRISP-DM的专门兴趣小组（正如人们所知道的“The SIG”）。我们通过邀请那些对数据挖掘感兴趣的人参加我们在阿姆斯特丹的一天工作小组，从而建立了这个兴趣小组。在那个工作小组中，我们阐释了我们的想法，并邀请他们讲述他们自己的想法，公开的讨论怎样使得CRISP-DM获得进步和发展。

在组织工作小组的那一天，委员会中的成员都怀着一颗忐忑的心。会不会没有人对此有足够感兴趣，以至于他们不愿意发表意见？或者也许他们表明了自己的想法，但却告诉我们他们认为这一基本程序不会有什么足够吸引人的需要？再或者我们的想法如此超前，以至于任何标准化的想法都会成为一个不切实际的幻想？

然而，工作小组超出我们的预期。主要表现为以下三点：

①有两倍于我们起初预期的人出现在现场。

②与会者有一个近乎完全一致的想法：企业需要一个标准化的程序，而且现在就需要。

③由于每一位参与者都从他们自身的企业实践经验角度阐释了他们对于数据挖掘的见解，所以有关这一标准程序的看法已经很明晰：尽管仍有些表面的差异——主要表现在阶段的划分和术语上——但有关数据挖掘程序的理解，参与者有着惊人的一致。

到工作小组结束时，我们在听取了SIG成员的意见和批评之后，已有相当的自信认为可以发布一个标准程序模型来维持这个数据挖掘社团。

在接下来的两年半时间里，我们着手于CRISP-DM的进一步开发和研制工作，并在Mercedes-Benz公司和我们的保险部门合伙人——OHRA公司的大规模数据挖掘实践项目中，进行试验。此外，我们还进行了CRISP-DM与商

业数据挖掘工具的整合工作。The SIG的成立具有无限的价值，其成员数量已超过200人，伦敦、纽约和布鲁塞尔也已成立了工作小组。

1999年中期，也就是欧洲委员会资助的那部分项目结束的时候，我们自认为已经起草了一个相当好的程序模型的草稿。那些熟悉草稿的人会发现经过一年的时间，CRISP-DM 1.0绝对有了根本的不同，尽管现在它更加全面和完善。但是我们也清晰地意识到，在项目进行的过程中，程序模型仍然是一个需要不断改进的模型：CRISP-DM还仅仅在一个相当窄的领域内有效。在过去的一年中，DaimlerChrysler 有机会把CRISP-DM应用到更广泛的领域当中去。SPSS和NCR公司的专业服务组已经采纳了CRISP-DM，并在大量的涉及许多工商业问题的消费者应用中，成功了运用了CRISP-DM。在这段时间内，我们注意到，非协会成员的服务提供商们采用了CRISP-DM ；分析师们已把它作为一个行业标准，不断的参考这一模型；同时消费者们也逐渐意识到了CRISP-DM的重要性（目前CRISP-DM经常在RFP文件中被提到）。我们相信我们最初的想法已经被彻底的证实，进一步的扩展和改善虽然是必需的，但我们认为CRISP-DM 1.0已足可以继出版、发行。

从技术原理上来讲，CRISP-DM还未能以一个学术、理论的形式来构建，它也不是一些权威委员会的精英们闭门思过的结果。过去我们也曾尝试过这些方法，旨在构建CRISP-DM的方法论，但这些方法很少能够建立一个实践性的、成功的以及被广泛采纳的标准。CRISP-DM之所以成功，就在于它建立在人们进行数据挖掘项目的实践的和真实的经验的基础之上。基于这一点，我们要非常感谢那些在项目中努力并提供想法的从业者们。

CRISP-DM 委员会

2000年8月

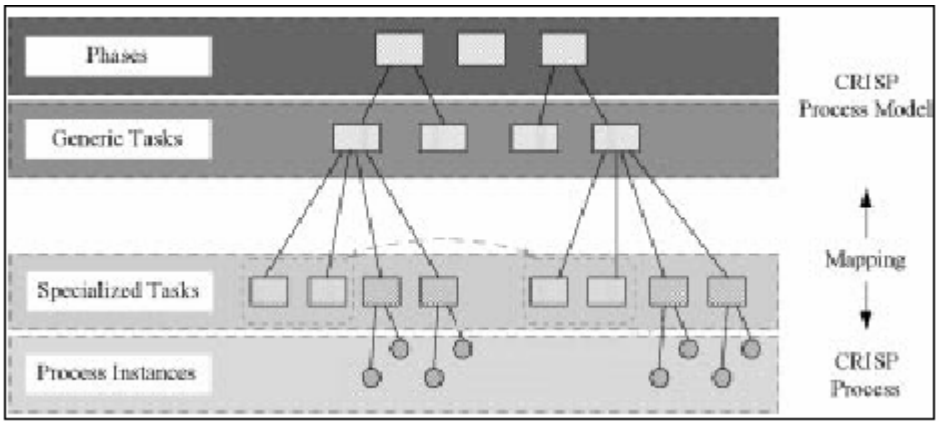
目录

Ⅰ 导言

1 CRISP-DM 方法论

1.1 分级细目

本章中，我们将以一种分层的程序模型来介绍CRISP-DM数据挖掘的方法，主要包括提炼出来的四个层次（即由一般到具体）：阶段描述、一般性任务描述、具体任务描述以及实施过程（process instance）。（详见图形1）。



图形1：CRISP-DM方法论的四层分级细目

第一层中，数据挖掘程序被分成几个阶段：每一个阶段都由几个二层一般性任务组成。

第二层之所以被称为一般性任务，是为了能够具有足够的概括性和一般性，以便这一层能够尽可能包括所有的数据挖掘情景。此外，一般性任务层要尽量全面和稳定。全面意味着它要既能够涵盖整个数据挖掘的过程，同时还要涉及到所有的数据挖掘的应用软件。稳定是指对于无法预见到的发展，例如新的建模技术,模型能够依然有效。

第三层，具体任务层，是指在某种具体情况下，对于一般性任务，要进行哪些具体的活动。例如，一般性任务层中有一项任务为清理数据，那么第三层就会介绍在不同情况下，清理数据会有哪些不同，例如，清理数值型数值还是字符型数值，或者处理数据的方法是用聚类还是预测模型。

我们以一种特定的顺序来分别介绍数据挖掘的阶段以及各阶段的任务，表现了一种理想的事物发展顺序。然而在实践中，许多任务往往以不同的顺序进行，而且重复不断地返回去做同一个任务、或重复某项活动有时候也是必要的。我们的程序模型并不企图能够获得在数据挖掘过程中可能出现的所有程序路线，因为那将会产生一个极为复杂的程序模型。

第四层，process instance，记录了数据挖掘实施过程中的具体活动、决策以及结果。每一步process instance都是根据上一级的任务组织进行的，但却呈现了具体实施过程中，而不是一般性任务中，经历的实际过程。

1.2 参考模型和用户指南

CRISP-DM在同一层次上将参考模型和用户指南加以区分。参考模型概括描述了数据挖掘过程中的各个阶段、任务以及结果，并对一个数据挖掘项目需要做什么进行了介绍。而用户指南则对数据挖掘的每一个阶段以及每个阶段内的任务给出了更加详细的信息和线索，同时描述了怎样做一个数据挖掘的项目。

本文既包括了参考模型，又涵盖了通用层面上的用户指南。

2 根据通用模型策划专用模型

2.1 数据挖掘文本

数据挖掘文本在CRISP-DM的一般和具体中相互转化。通常，我们把数据挖掘的文本分成四个不同的维度：

①**应用领域：**数据挖掘具体适用的范围

②**数据挖掘的问题类型：**数据挖掘处理数据的方法分类**（详见附录V.2）**

③**技术问题：**数据挖掘过程中出现的各种技术问题。

④**工具和技巧**：数据挖掘过程中所使用的各种数据挖掘工具和技巧。

下面的表1汇总了数据挖掘文本的四个维度，并分别给出了具体的例子。

	数据挖掘文本			
维度	**应用领域**	**问题类型**	**技术问题**	**工具和技巧**
举例	**反应模型**	**描述和概要分析**	**缺失值**	Climentine
	流失预测	**分割分析**	??	MineSet
		概念描述分析		**决策树**
		分类分析		
		预测分析		
		依赖性分析		

表 1：数据挖掘文本的维度及其示例

一个明确的数据挖掘文本，对于一个或更多的维度而言都有一个具体的值。例如，一个数据挖掘程序在流失预测中处理分类问题时，就组成了一个具体的文本。对于不同维度的文本，数值越多，这个数据挖掘也就越具体。

2.2 根据文本策划模型

在CRISP-DM程序中，本书将通用和专用两个层面加以区分，制定了两种不同类型的模型。

当前策划：

如果使用通用程序模型，只是为了实施一个专门的数据挖掘项目，并根据需要尝试策划一般性任务和具体的活动，那么我们就来讨论一套只做一种用途的专门的模型策划。.

未来策划：

如果是为了根据预先定义的文本，系统地将通用模型具体化的话（或者类似的，为了系统地分析和巩固一套专门的程序模型的项目经验，以便将来在相似的文本中使用），那么我们在这里就明确地讨论以CRISP-DM的形式，书写一套专门的程序模型。

上面哪一种策划对你而言是适合的，取决于你的具体的数据挖掘文本和组织的需要。

2.3 策划的策略

将通用模型具体化的基本策略，对于上面两种类型的策划模式而言都是相同的：

①分析具体的文本。

②剔除任何不适用于文本的细节。

③增加任何使文本更加明确的细节。

④根据文本的具体特性，使一般性概念更加明确（或者举例说明）。

⑤为了清晰，有可能的话，将某些一般性概念重新命名，从而提供更加清楚的含义、信息。

3 章节介绍

3．1 内容

CRISP-DM程序模型(即这篇文章)主要由以下五部分组成：

①第一部分（I）：主要CRISP-DM方法论的介绍以及根据通用模型制定专用模型的一般性指导。

②第二部分（II）：描述CRISP-DM参考模型，CRISP-DM的四个阶段，一般性任务及其输出结果。

③第三部分（III）：介绍CRISP-DM用户指南，这一章远远不是对CRISP-DM各阶段、一般性任务以及结果的抽象描述，包含了如何实施数据挖掘程序的更加详细的建议。

④ 第四部分（IV）：集中介绍数据挖掘过程中及之后项目报告的撰写，并提出了这些报告的轮廓和要点。同时本章还交叉介绍了数据挖掘过程中的任务及其结果。

⑤第五部分（V）：本章主要是附录部分，主要是一个涵盖了重要术语的术语表和数据挖掘的问题类型。

3．2 意图

读者和使用者请注意以下关于本书的用法说明：

①如果您是第一次看到本书，请您从导言部分开始看起。这样你可以了解CRISP-DM的方法论，及其所涉及到的概念和这些不同概念之间的关系。在进一步的阅读中，您可以跳过本章。只有当您需要进一步明确某些概念时，再重新温习本章即可。

②如果您想很快对CRISP-DM模型有一个总体的把握，从而能够迅速得开始一个数据挖掘项目或者是为了更好得理解CRISP-DM用户指南，那么请您参看第二章——CRISP-DM参考模型。

③如果您需要如何实施数据挖掘项目的详细的建议，那么第三章——CRISP-DM用户指南，将是本书中最为有价值的一章。 注意：如果您没有读过导言部分或者是参考模型，请返回去重新阅读这两部分。

④如果您正在实施一个数据挖掘项目，并开始撰写项目报告，那么请您直接跳到第五章。若您喜欢在报告中对项目实施的具体过程有所描述，请您根据需要来回参考第三章到第五章的内容。

⑤最后，附录部分作为CRISP-DM和数据挖掘的背景信息，是非常有用的。如果您还不是这一领域的专家，请参照附录部分查找有关术语。

II CRISP-DM参考模型

当前这个数据挖掘的程序模型，为数据挖掘项目的生命周期提供了一个综合的描绘。它包括了一个数据挖掘项目所要经历的各个阶段，各阶段的任务以及这些任务之间的相互关系。从描绘的层面来看，是不可能鉴别出所有这些任务之间的关系的。但本质上看，这些任务之间是否存在关系，取决于使用者的目的，背景及其利益所在，与此同时，更重的还在于数据。

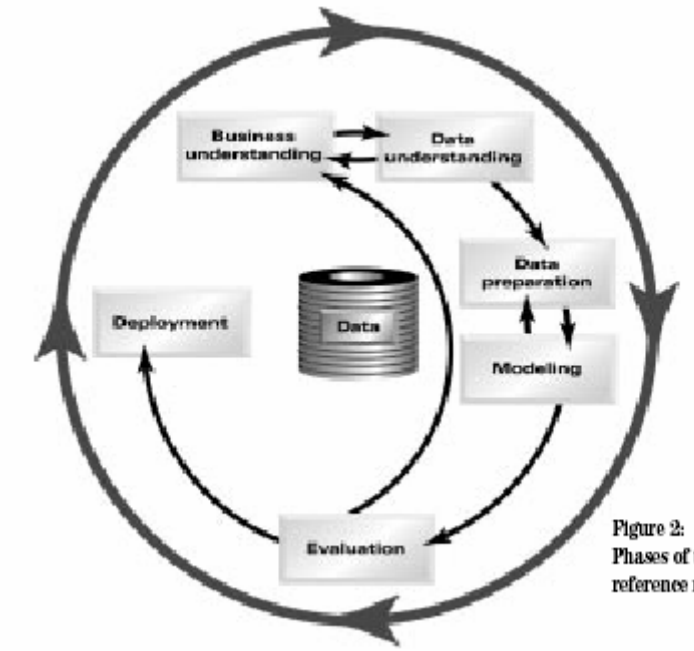


Figure 2:
Phases of the CRISP-DM reference model

数据挖掘项目的生命周期由六个阶段组成。图2展示了这一数据挖掘过程的各个阶段，这些阶段之间的顺序并不固定，在不同阶段之间来回流动往往是非常有必要的。究竟下一步要执行哪个阶段或者哪一个特定的任务，都取决于每一个阶段的结果。图中的箭头表明了阶段之间最重要和最频繁的依赖关系。

图2中最外层的这个循环表明了数据挖掘本身的循环性质。经过一个具体的数据挖掘项目得到了某项解决措施或办法并加以展开，并不代表数据挖掘本身已经结束。从这一数据挖掘过程以及解决措施展开的过程中所吸取的经验、教训，又引发了新的，通常是更加焦点的商业问题。接下来的数据挖掘过程将会从过去的项目经验中获利。

在接下来的内容中，我们将简要的勾勒一下每个阶段的轮廓：

商业理解

这一初始阶段主要集中在对项目目标的理解，以及从商业角度考虑，对客户需求的理解。进而把这些理解转化为一个数据挖掘的定义和为了达到目标的初步方案。

数据理解

数据理解阶段开始于数据的收集工作。接下来就是熟悉数据的工作，具体如：检测数据的质量，对数据有初步的理解，探测数据中比较有趣的数据子集，进而形成对潜在信息的假设。

数据准备

数据准备阶段涵盖了从原始粗糙数据中构建最终数据集（将作为建模工具的分析对象）的全部工作。数据准备工作有可能被实施多次，而且其实施顺序并不是预先规定好的。这一阶段的任务主要包括：制表，记录，数据变量的选择和转换，以及为适应建模工具而进行的数据清理等等。

建模

在这一阶段，各种各样的建模方法将被加以选择和使用，其参数将被校准为最为理想的值。比较典型的是，对于同一个数据挖掘的问题类型，可以有多种方法选择使用。一些建模方法对数据的形式有具体的要求，因此，在这一阶段，重新回到数据准备阶段执行某些任务有时是非常必要的。

评估

从数据分析的角度考虑，在这一阶段中，您已经建立了一个或多个高质量的模型。但在进行最终的模型部署之前，更加彻底的评估模型，回顾在构建模型过程中所执行的每一个步骤，是非常重要的，这样可以确保这些模型是否达到了企业的目标。一个关键的评价指标就是看，是否仍然有一些重要的企业问题还没有被充分地加以注意和考虑。在这一阶段结束之时，有关数据挖掘结果的使用应达成一致的决定。

部署

模型的创建并不是项目的最终目的。尽管建模是为了增加更多有关于数据的信息，但这些信息仍然需要以一种客户能够使用的方式被组织和呈现。这经常涉及到一个组织在处理某些决策过程中，如在决定有关网页的实时人员或者营销数据库的重复得分时，拥用一个“活”的模型。 It often involves applying “live” models within an organization’s decision making processes, for example in real-time personalization of Web pages or repeated scoring of marketing databases. 然而，根据需求的不同，部署阶段可以是仅仅像写一份报告那样简单，也可以像在企业中进行可重复的数据挖掘程序那样复杂。在许多案例中，往往是客户而不是数据分析师来执行部署阶段。然而，尽管数据分析师不需要处理部署阶段的工作，对于客户而言，预先了解需要执行的活动从而正确的使用已构建的模型是非常重要的。

商业理解	数据准备	数据理解	建模	评估	部署
确定商业目标 *背景* *商业目标* *商业成功的标准* *评估形势* *企业资源清单* *要求、设想和局限* *风险和意外* *术语* *成本和收益* **确定数据挖掘目标** **制定项目计划**	**收集原始数据** **描述数据** **探索数据** **检验数据质量**	**选择数据** **清理数据** **构造数据** **整合数据** **格式化数据**	**选择建模技术** **制作检验设计** **建造模型** **评估模型**	**评估结果** **回顾过程** **确定下一步方案**	**制定部署方案** **制定监控和维护方案** **书写最终报告** **回顾项目**

图3：CRISP-DM参考模型的一般性任务（粗体）和结果（斜体）

图3展示了各个阶段的大致轮廓，同时配有各个阶段的一般性任务（粗体）及其结果（斜体）。在接下来的部分中，我们将更加详细地介绍每一个一般性任务及其结果，但仍然集中在任务的综述及结果的摘要介绍。

1 商业理解

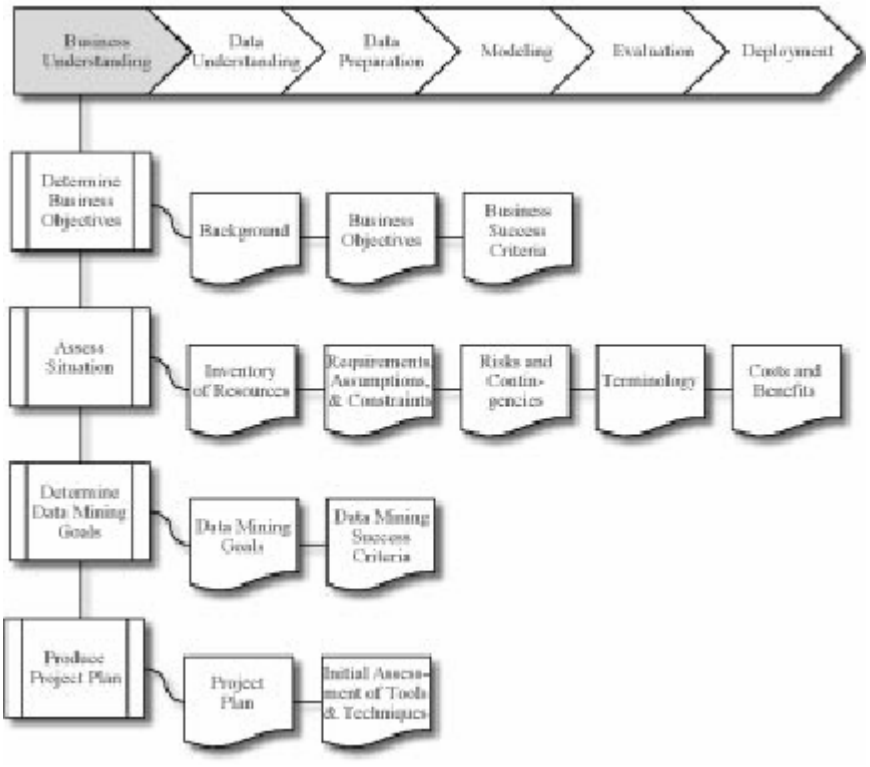


图4：商业理解

1.1 确定商业目标

任务 **确定商业目标**

数据分析师的第一个目标就是从商业的角度，彻底的了解客户想要实现什么？通常客户有许多竞争目标和需要正确加以衡量的局限。首先，数据分析师要去发现那些影响项目结果的重要因素。忽视这一步的可能后果是需要花大力气去寻找解决问题的正确答案。

结果 **背景**

项目之初，记录那些可获知的有关企业商业形势的信息。

商业目标

从商业角度描绘客户的首要目标。除此之外，还有其他一些与首要目标相关，由客户提出的典型商业问题。例如，企业的首要目标可能是通过预测其消费者何时转向其竞争对手那里，从而保持目前现有的消费者。相关的商业问题可能是“银行账户使用的第一信贷方式（如ATM，visit branch，网上交易）是如何影响他们的去留的？”或者“ATM的低收费是否会明显的减少高额存款账户离开的数量？”

商业成功的标准

从商业角度，描述项目结果是否成功或实用的标准。这些标准可能是相当详细和明确的，并且能够客观的加以测量，例如客户流失减少到一定的程度；也可能是非常概括和主观的，例如“对相互关系提出有益的见解”，但在后面的描述中要指出有由谁做出这一主观判断。

1.2 评估形势

任务 **评估形势**

这一任务涉及更加详细的事实资料——查找所有的资源，局限，设想以及在确定数据分析目标和项目方案时考虑到的各种其他的因素。在前一个任务中，你的目标是快速得到企业形势的症结所在。而现在，你需要充实细节。

结果 **企业资源清单**

列出所有对项目有用的资源，包括：人员（企业专家，数据专家，技术支持者，数据挖掘人员），数据（备好的数据摘录fixed extracts，获得数据库或者可操作数据的途径），数据处理资源（硬件平台）以及软件（数据挖掘工具，其他相关软件）。

要求，假设和局限

列出项目的所有要求，包括项目完成的进度表，项目结果的可理解性和质量，安全性以及相关的法律问题。作为这一工作的结果，确保使用数据的权利。

列出项目做出的所有假设。这些假设可能是有关数据的假设，可以在后面的数据挖掘过程中加以验证，但也可能包括这个项目所依赖的企业的一些无法验证的假设。如果后面这些假设可以构成决定项目结果合法性的条件，那么列出后面这一设想尤为重要。

列出项目的所有局限性。他们可能是资源可用性的局限，但也可能包括一些技术上的局限，比如数据集的大小，这将决定后面的建模。

风险和意外

列出所有的风险或者有可能推延项目进行甚至致使其失败的事件。列出相应的可能的计划；如果发生了意外，将采取什么样的行动应对。

术语

编辑一套与本项目相关的术语表。它可以包括以下两个部分：

(1) 一套相关的商业术语表，它将构成项目可利用的商业理解的一部分。构建这个术语表的过程本身就是一个“知识启发”和教育培训的过程。

(2) 一套数据挖掘术语的术语表，并用与商业问题有关的事例加以解释说明。

成本和收益

构建项目的成本-收益分析,把项目成本和项目成功后企业的获利相比较。这一比较要尽可能的详细和精确，例如使用商业环境下的货币测量方法。

1.3 确定数据挖掘目标

任务 **确定数据挖掘目标**

商业目标是以相关的商业术语表现出来的目标。而数据挖掘的目标则是以技术形式表现出来的目标。例如，一个商业目标可能是“增加现有消费者的销售类别catalog”。而数据挖掘的目标则可能是“根据消费者过去三年的购买情况、人口统计信息（如年龄，工资，居住低等等）以及商品的价格，预测他今年将要购买多少商品”。

结果 **数据挖掘的目标**

记录项目预计能够实现的达成商业目标的结果。

数据挖掘成功的标准

用相关的技术术语定义项目成功的标准，例如要达到一定水平要求的预测精确度，或者是购买倾向的一定程度的“提高”等等。至于商业成功标准，需要以主观的形式加以陈述，同时应明确指出做出这一主观判断的群体或个人。

1.4 制定项目计划

任务 **制定项目计划**

陈述为了实现数据挖掘目标进而实现商业目标的预计的计划。计划方案应尽可能详尽的陈述接下来的项目中所要进行的预期的一系列举措，包括数据挖掘的工具以及技术的初步选择等等。

结果 **项目计划**

陈列出项目过程中所要执行的各个阶段，包括各阶段持续的时间，所需要的资源，输入的数据，输出的结果及其附属物。明确数据挖掘过程中那些可能需要大规模重复的地方，例如建模和评估阶段的循环操作。

作为项目计划的一个组成部分，分析时间进度与风险之间的相互依赖关系也是非常重要的。项目计划中，应明确的标明这些分析的结果，比较理想的状态是能够配有风险发生时的解决方案和补救措施。

注意：项目计划应包含每一个阶段的详细计划。例如，确定评估阶段将要使用的评估策略等等。

项目计划是一个动态的文件。这也就是说在每一个阶段之末，研究者有必要对本阶段的进展和成果进行回顾，同时相应的更新项目计划。对于这些回顾的具体的着眼点也是本计划的组成部分之一。

工具和技术的初步评估

在第一阶段的末尾，项目组还应该进行一次工具和技术的初步评估工作。例如，在这儿，您应该选择一种数据挖掘工具来支持项目进程中不同阶段所使用的各种方法。项目进程的早期对工具和技术进行评估是非常重要的，因为工具和技术的选择很有可能影响到整个项目的进行。

2 数据理解

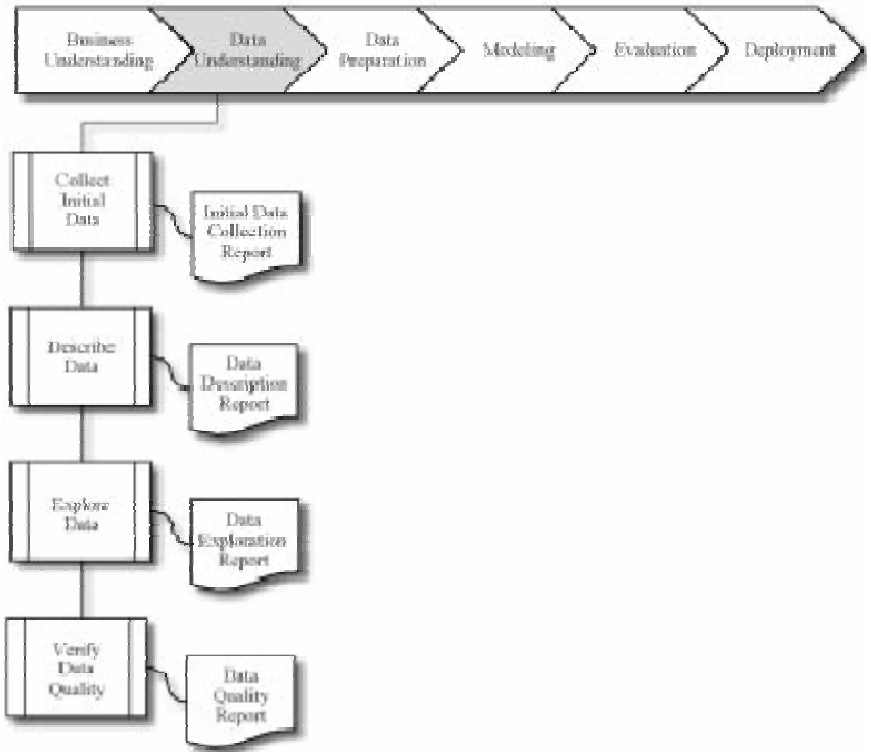


图5：数据理解

2.1 收集原始数据

任务 **收集原始数据**

获得项目资源中列出的项目数据（或者使用权）。有必要的话，还要进行数据的装载工作。例如，当你需要使用某一特定的工具进行数据理解时，将数据安装在这一工具之中是非常有意义的。同时，这一工作还有可能与初期的数据准备阶段相联接。

注意：如果你获得了多个数据资源，那么或者在这儿，或者在后面的数据准备阶段，数据整合将成为一个您额外需要处理的问题。

结果 **初步的数据收集报告**

列出所获得的数据（或数据集），以及这些数据在项目中需要出现的位置，获得的方法和所面对的问题。记录下数据收集中面对的问题以及处理这些问题的解决办法,将有助于今后执行同样的或者与之相类似的方案。

2.2 描绘数据

任务 **描绘数据**

检验所得数据的“总的”或者说“表面的”特征并报告其结果。

结果 **数据描绘报告**

描绘所得数据，包括：数据格式，数据性质，例如每一个数据表格中记录的条数和变量的数目，变量特征以及任何其他表面特征。数据是否已经满足一系列相关的需要？

2.3 探索数据

任务 **探索数据**

这一任务将处理数据挖掘的问题，这些问题可以以质疑的方式被提出，也可以是一些显而易见的问题或者是以报告的形式加以陈述。他们主要包括：关键变量的区分，例如预测任务中的目标变量；配对变量或几个变量之间的关系；简单聚合的结果；重要的潜在人群的特征；简单的统计分析等等。这些分析将直接涉及到数据挖掘的目标；同时也将有助于进一步完善数据描述，改进报告质量，并为进一步的数据分析提供了数据转换和

其他的数据准备。

结果 **数据探索报告**

描绘数据探索的结果，包括最初的发现或者初期的假设以及这些发现对余下项目的影响。可行的话，报告还应包括一些图表，用来展示数据的特征或者产生有趣的数据子集以作进一步的检测。

2.4 检验数据质量

任务 **检验数据质量**

检验数据质量，列举相关的问题，例如：数据是否完整（它是否覆盖了所需要的全部case）?数据是否正确或者数据是否包含错误？如有错误，是否常见？数据中是否存在缺失值？如存在，它们以何种方式出现，在哪里出现，是否常见？

结果 **数据质量报告**

列举数据质量检测的结果；如果存在数据质量问题，列出可能的解决办法。通常数据质量问题的解决办法主要取决于数据和商业知识。

3 数据准备

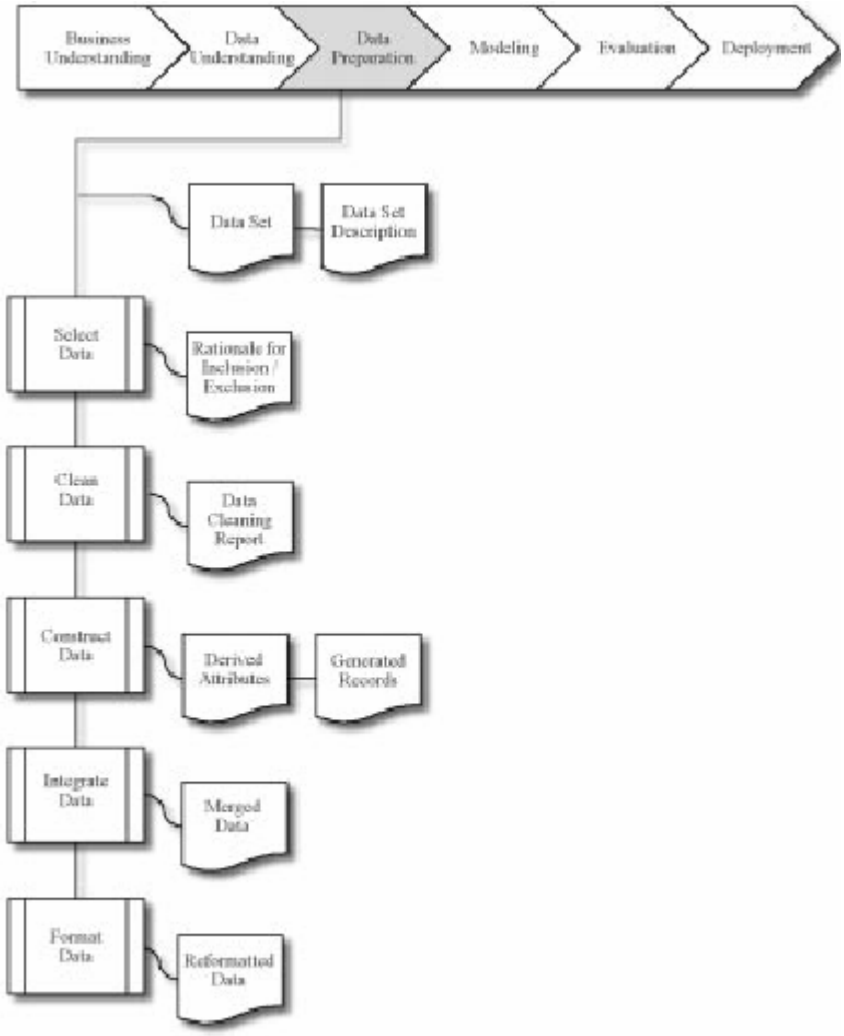


图6：数据准备

***结果* 数据集**

数据准备阶段产生数据集（或数据集合），以作建模或者项目的其他分析工作之用。

数据集描述

描绘用于建模或者项目主要分析工作的数据集。

3.1 选择数据

***任务* 选择数据**

确定作分析之用的数据。选择的标准包括与数据挖掘目标的相关性，数据质量以及技术限制，例如对数据大小或数据种类的要求等。注意数据选择既包括数据表各种数据记录（行）的选择，同时也包括数据变量（列）的选择。

***结果* 选择和剔除的基本原理**

列出被选择/剔除的数据及其原因。

3.2 清理数据

***任务* 清理数据**

根据所选分析技术的要求，将数据质量提高到相应的水平。这一工作可能涉及选择的干净的数据子集，插入适当的默认值或者使用更加出色的处理技术，例如通过建模估测缺失值。

***结果* 数据清理报告**

针对*数据理解*阶段中*检测数据质量*任务所报告的数据质量问题，描绘本阶段任务中所采取的决策和行动。同时为了清理数据，对数据所进行的转换以及这些转换对分析结果的可能的影响在此报告中都要有所涉及。

3.3 构造数据

***任务* 构造数据**

这一任务包括建设性的数据准备工作，例如衍生变量的产生，全新的记录的产生或者已存变量的值转换。

***结果* 衍生变量**

衍生变量是新生的变量，来自于同一条记录中的一个或多个变量。例如，面积=长*宽。

新生记录

描绘全新记录的产生。例如：创造一些去年没有购买过产品的消费者记录。在原始数据中是不可能有这样的记录的，但是为了建模的需要，明确地表现出一定顾客“0”消费的事实，是非常有意义的。

3.4 整合数据

***任务* 整合数据**

这是一些有关信息如何从多个数据表格或数据记录中汇总，进而产生新记录或新值的方法。

***结果* 合并的数据**

合并表格是指把那些有关同样对象，但有不同信息的两个或以上的表格合并起来。例如，一个零售连锁店有一个关于每个零售店总的信息的数据表格（例如，房屋面积，商场类型），另外一个是有关汇总的销售信息的数据表格（例如，利润，去年以来的销售百分比变化），还有一个是关于店铺周围人口信息的数据表格。这几个表格都包含有每一个零售店的记录。因此，这些表格可以合并为一个新的表格，一条记录就是一个零售店，并将几个表格中的变量栏汇总。

合并数据也包括聚合数据。聚合数据是指将多个记录或多个数据表中的信息汇总起来，计算得出的新的数值。例如，将一个有关消费者购买的数据表格转化为一个新的数据表格。在原来的表格中，消费者的每一次购买都是一条记录。而在新的表格中，通过把消费者的购买信息加以汇总计算，使的新表格中每一个消费者成为一条记录，相应的，每一个记录的信息为：购买数量，平均购买量，信用卡支付的比例，促销产品的购买比例等等。

3.5 格式化数据

***任务* 格式化数据**

格式化数据主要是指在不改变数据原意的基础上，根据建模工具的需要将数据进行句法的改变。

***结果* 格式化了的数据**

有些工具对数据变量的顺序有所要求，例如，每一条记录的第一列应为其唯一的标识或者最后一列应为模型预测的结果列。

有时候改变数据集中记录的顺序也是非常重要的，建模工具可能需要数据记录根据某一变量的值进行排序。通常，一个比较常见的情形是数据集中的记录本来是按照一定顺序进行排列的，而某一建模的计算法则却要求记录随机排列。例如，在使用神经网络时，数据记录最好是以随机方式排列，当然有些工具并不需要分析者的人为的外在处理，可以自动进行这一处理。此外，为了满足一些特殊的建模工具的要求，还要对数据进行纯粹的句法改变。例如，将那些以逗号区隔的数据文件中的文本栏中的逗号剔除，把所有的值整理为32个字符的最大值。Examples: removing commas from within text fields in comma-delimited data files, trimming all values to a maximum of 32 characters.

4 建模

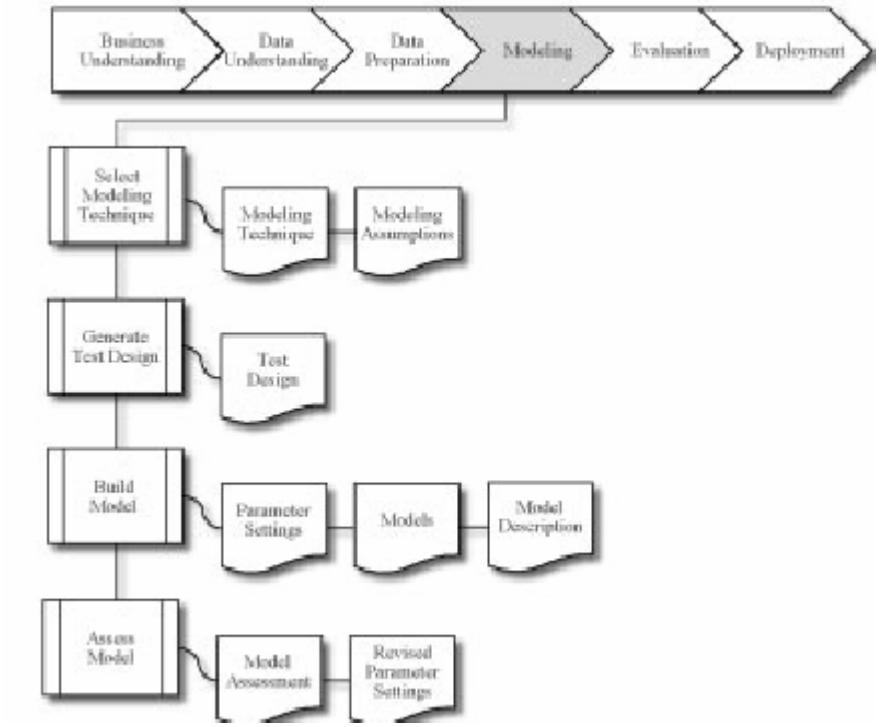


图7：建模

4.1 选择建模技术

***任务* 选择建模技术**

建模第一步：选择一个实实在在的建模技术来使用。在商业理解中，你可能已经选择了一个建模工具，但在这儿是指具体的建模技术，例如C4.5的决策树或者利用了back propagation的神经网络。如果有多种技术要使用，那么在这一任务中，对于每一个要使用的技术要分别对待。

***结果* 建模技术**

记录将要使用的实实在在的建模技术。

建模假定

许多建模技术对于数据均有明确的假定，例如，所有变量有一致的分布，不允许有缺失值，分类变量必须为符号型（symbolic）等等。记录所有这样的假定。

4.2 制作检验设计

任务 **制作检验设计**

在正式建立模型之前，我们需要制作一个程序或者机制来检验模型的质量和有效性。比如，在监控数据挖掘任务中，如分类预测，通常使用错误率（error rates）作为检验模型质量的方法。因此，比较有代表性的做法即把数据集分为训练集和检验集，通过训练集来建立模型，再通过分开的检验集来评估模型的质量。

结果 **检验设计**

记述训练、检验、评估模型的计划方案。本方案的一个主要内容是确定如何将有效的数据集分成训练数据、检验数据和确认数据。

4.3 建造模型

任务 **建造模型**

在备好的数据集中运行建模工具，建立一个或多个模型。

结果 **参数设置**

对于任意一个建模工具，均有大量可调整的参数。列出这些参数，被选值以及选择这些参数设置的理由。

模型

此结果是由建模工具得出的实实在在的模型，而不再是一个报告。

模型描绘

描绘模型，报告有关模型的解释并记录任何其含义理解上的困难。

4.4 评估模型

任务 **评估模型**

数据挖掘工程师要根据其专业领域的知识、数据挖掘成功标准以及需要的检验设计来解释模型，这将涉及到接下来的评估阶段的工作。尽管数据挖掘工程师可以从技术的角度来判断模型应用和技术发现的成功与否，但是他仍然需要与商业分析师和后面的领域专家接触、沟通，从而进一步探讨商业环境下的数据挖掘结果。而且，这一步任务仅仅涉及模型的评估，至于整个项目产生的所有其他的结果，只有到评估阶段才会通盘考虑。

数据挖掘工程师要尽力给模型评级。在对模型进行评估时，他既要参照评估标准，同时也要考虑到商业目标和商业成功的标准。在大多数的数据挖掘项目中，数据挖掘工程师要不止一次的应用某个特定的技术或者是利用不同的可选择的技术产生多种结果。因此在这一阶段的任务中，他也要根据评估标准比较所有不同的结果。

结果 **模型评估**

汇总这一阶段工作的结果，列出所有模型的质量（例如，以精确度的形式表现）并将他们进行比较排序。

经校订的参数设置

根据模型评估，校订、调整参数设置，为下一轮的**模型建造**工作做准备。重复进行模型的建造和评估工作，直到你相信已经找到了最好的模型为止。记录所有这些校订和评估的工作。

5 评估

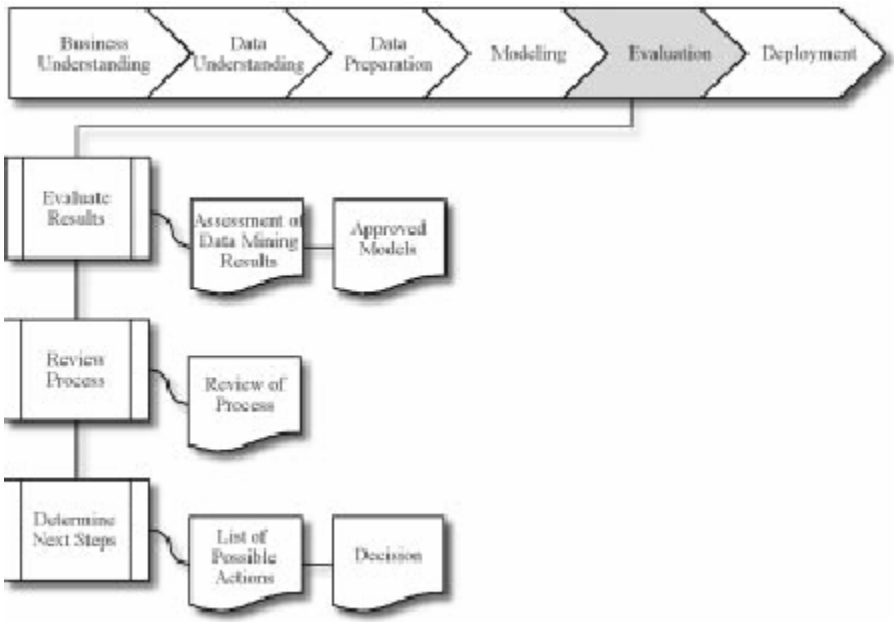


图8：评估

5.1 评估结果

任务 评估结果

前面的评估工作主要处理诸如模型精确度和普遍有效性（generality）等的因素，而本阶段则主要评估模型在多大程度上符合商业目标，寻找并确定模型不完善的原因。评估阶段的另一项可选的任务就是在时间和经费允许的条件下，将模型置于实践中加以检验。

此外，评估阶段也要评估数据挖掘产生的其他结果。数据挖掘结果既包括与原来商业目标有必然联系的模型，同时也包括其他的发现，这些发现可能与原来的商业目标没有必然的联系，但却有可能揭露出了额外的挑战以及指引将来发展方向的信息或者暗示。

***结果* 参照商业成功标准的数据挖掘结果评估**

汇总参照商业标准的数据挖掘评估结果，包括项目是否符合最初的商业目标的最终陈述。

核准模型approved models

经过参照商业成功标准的模型评估之后，符合选择标准的模型成为了核准模型。

5.2 回顾历程

***任务* 回顾历程**

在这里，模型将表现得令人满意，符合商业要求。

现在是时候对数据挖掘的实施过程作一次彻底的回顾了，以便发现并确定是否仍然有一些重要的步骤或者因素被忽略了。此外，这一回顾工作还涉及质量保证问题，例如，我们正确的建造了模型吗？我们只使用了那些我们可以得到的变量吗？这些变量在将来的使用中是否仍然适用？

***结果* 历程回顾**

汇总历程回顾的结果，突出显示那些已经被错过或者应该被重复的工作。

5.3 确定下一步方案

***任务* 确定下一步方案**

根据评估结果和历程回顾，确定接下来的任务。项目是应该结束项目本身而进行部署呢？还是开始进一步的重复工作或者建立新的数据挖掘项目。此外，这一任务还包括对影响决策的剩余资源和预算的分析。

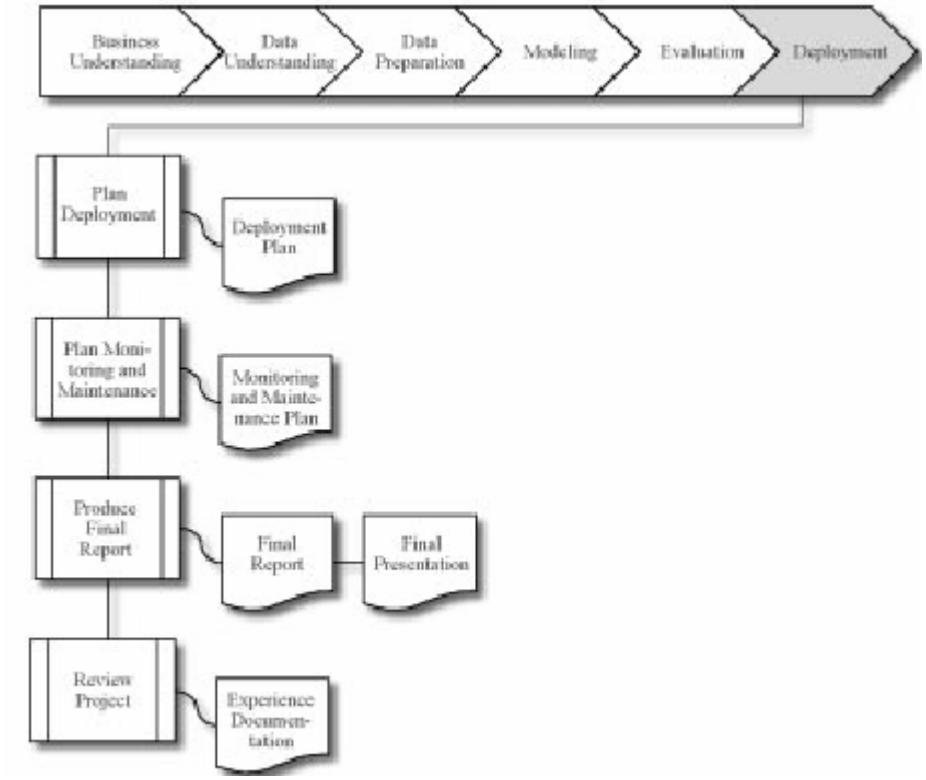
***结果* 可能的举措的列表**

列出所有可能的下一步的举措，以及支持和反对的理由。

Decision

陈述下一步方案的决策以及原因。

6 部署运用



6.1 制定部署运用方案

***任务* 制定部署方案**

为了把数据挖掘结果在商业中加以部署运用，此阶段的任务涉及评估结果和议定一个部署运用的策略。如果已经确定了一个创造相关模型的全面的程序，将此程序记录以作后面部署之用。

***结果* 部署方案**

汇总部署策略，包括一些必要的步骤以及相应的实施办法。

6.2 制定监控和维护方案

***任务* 自定监控和维护方案**

如果数据挖掘结果成为日常商业运作和商业环境中的一个组成部分，那么监控和维护将成为一个重要的议题。精心准备监控策略将避免今后对数据挖掘结果的长期误用。为了能够监控数据挖掘结果的部署工作，项目组需要制定一个详细的监控方案。此外，此方案还要考虑部署的类型。

***结果* 监控和维护方案**

汇总监控和维护策略，包括一些必要的步骤以及相应的实施办法。

6.3 书写最终报告

***任务* 书写最终报告**

项目之末，项目组领导人和他的团队要写最终的报告。依照部署方案，这一最终报告可能仅仅是项目的汇总以及相关经验（如果经验并没有随着项目的前进而被记录的话），也可能是对数据挖掘结果的最终的复杂的陈述。

***结果* 最终的报告**

这是数据挖掘实施的最终手写报告。它包括前面所有结果的提交、组织和综述。

最终展示

通常在项目结束之时，也会有一个会议，会上项目结果将会以口头的方式展示给客户。

6.4 回顾项目

***任务* 回顾项目**

评估项目进程的正确与错误，哪些做得好？哪些还需要改进？

***结果* 经验文档**

汇总项目进程中的经验。例如，缺陷，在某些相似情形下选择最合适的数据挖掘技术时，容易产生误解的地方或者暗示等等都可以记录下来。在一份理想的报告中，经验文档应包括项目每个阶段的参与者所写得有关其工作的报告。

III CRISP-DM用户指南

1 商业理解

1.1 确定商业目标

***任务* 确定商业目标**

数据分析师的第一个目标就是从商业的角度，彻底的了解客户想要实现什么？通常客户有许多竞争目标和需要加以正确衡量的局限。首先，数据分析师要去发现那些影响项目结果的重要因素。忽视这一步的可能后果是需要花大力气去寻找解决问题的正确答案。

***结果* 背景**

项目之初，比较那些可获知的有关企业商业形势的信息。这些细节信息不仅有助于进一步识别要解决的企业目标，同时还有助于识别人力和物力资源，这些资源在项目进行过程中可能会被需要和使用。

***业务活动* 组织**

① 制作一份组织图，标识出区域，部门和项目组，以及各个部门负责人的名字和职责。

② 识别企业中的关键人物及其职位。

③ 确定一位企业内部的负责人（财政负责人和主要使用者/领域专家）

④ 企业中是否有一个筹划指导委员会？它的成员都是谁？

⑤确定企业中能够受到数据挖掘项目影响的部门（例如，市场部，营销部，财政部）

问题领域

① 确定问题领域（如，行销，客户管理，商业发展等等）

② 概括描述这些问题。

③ 检查项目目前的状态（如，企业内部机构是否已经清楚了解我们正在实施一个数据挖掘项目或者我们是否需要把数据挖掘作为一个关键技术在企业内部进行推广？）

④ 清楚项目的先决条件（如，项目的动机是什么？企业是否已经使用数据挖掘？）

⑤ 如有必要，准备一份口头报告，向企业展示数据挖掘。

⑥ 确定项目结果的目标群体（如，我们是要为企业的高端提供一份书面报告还是仅仅为终端用户提供一个运作体系）。

⑦ 确认用户的需要和预期。

当前的解决方案

① 描绘当前使用的处理问题的解决办法。

② 列举当前解决方案的优缺点，以及使用者对此方案的接纳程度。

***结果*　商业目标**

从商业角度描绘数据挖掘中客户的首要目标。除此之外，还有相当大量的，客户提出来的与首要目标相关的典型商业问题。例如，企业的首要目标可能是通过预测其消费者何时转向到竞争对手那里，从而保持目前现有的消费者，而第二位的商业目标则可能是确定低收费是否只能影响一小部分的顾客。

业务活动

① 粗略的描述数据挖掘要解决的问题。

② 尽可能准确地详细说明企业问题。

③ 详细说明企业的任何其他要求（如，企业不想流失任何顾客）。

④ 详细说明企业的预期收益。

注意！ 谨防设置不可实现的目标——尽可能的现实。

***结果*　商业成功的标准**

从商业角度，描述项目结果成功或实用的标准。这些标准可能相当详细和明确，并且能够加以客观地测量，例如客户流失减少到一定的程度；也可能非常概括和主观，例如“对相互关系提出有益的见解”，对于此类标准，后面的描述中要指出由谁做出这一主观判断。

业务活动

① 详细说明企业成功标准（如，将邮购活动中的回应率提高10%，签收率sign-up rate提高20%）

② 明确由谁负责评估成功的标准。

切记！ 每一个成功标准要至少涉及一个专门的企业目标。

好主意！ 在形式评估开始之前，你可以参考这类问题过去的相关经验——内部使用过的CRISP-DM或者是外部使用的成套的处理办法。

1.2 评估形势

***任务*　评估形势**

这一任务涉及更加详细的事实资料——查找所有的资源，局限，设想以及在确定数据分析目标和项目方案时考虑到的各种其他的因素。在前一个任务中，你的目标是快速得到企业形势的症结所在。而现在，你需要充实细节。

***结果*　企业资源清单**

列出所有对项目有用的资源，包括：人员（企业专家，数据专家，技术支持者，数据挖掘人员），数据（备好的数据摘录fixed extracts，获得现成数据库或者操作数据的权力），数据处理资源（硬件平台）以及软件（数据挖掘工具，其他相关软件）。

***业务活动*　硬件资源**

① 明确基本的硬件资源。

② 确定基本的硬件资源对于数据挖掘项目的实用性。

③ 检查硬件维护时间是否与数据挖掘项目的使用相冲突。

④ （如果本阶段已经明确将要使用的数据挖掘工具）确认硬件资源是否适用于此数据挖掘工具。

数据和知识资源

① 确认数据资源。

② 确认数据资源的种类（在线资源，专家，书面资料等等）。

③ 确认知识资源。

④ 确认知识资源的种类（在线资源，专家，书面资料等等）。

⑤ 检验可利用的工具和技术。

⑥ 描绘相关的背景知识（正式和非正式）。

人力资源

① 确认项目负责人（如果不同于前面1.1.1中提到的企业内部负责人）

② 确认系统管理者，数据库管理者以及可以处理深层次问题的技术支持人员。

③ 确认市场分析师、数据挖掘专家和统计专家，并检查他们的可得性。

④ 确认后面各阶段领域专家的可得性。

切记！整个项目过程中，项目可能偶尔会需要相关的技术人员，如数据转化员。

结果 **要求，假设和局限**

列出项目的所有要求，包括项目完成的进度表，项目结果的可理解性和质量，安全性以及相关的法律问题。作为这一工作的结果，确保使用数据的权利。

列出项目做出的所有假设。这些假设可能是有关数据的假设，可以在后面的数据挖掘过程中加以验证，但也可能包括这个项目所依赖的企业的一些无法验证的假设。如果后面这些假设可以构成决定项目结果合法性的条件，那么列出后面这一设想尤为重要。

列出项目的所有局限性。他们可能是资源可用性的局限，但也可能包括一些技术上的局限，比如数据集的大小，这将决定后面的建模。

业务说动 **要求**

① 详述目标群体的外部特征。

② 得到有关时间进度的要求。

③ 得到有关数据挖掘项目及其结果的可理解性，精确性，可运用的能力，可维护性以及可重复性的要求规定。

④ 得到有关安全性，法律限制，私有性，报告以及项目进度的要求。

假设

① 详细阐述所有的假设（包括不明显的假设），并使其清晰明确（如，提出企业问题，一个最小数目的年龄在50岁以上的消费者也是必要的）。

② 列出有关数据质量的假设（如，精确性，可用性）。

③ 列出有关外部因素的假设（如，经济问题，竞争产品，技术进步）

④ 阐明所有估价的假设（如，一个具体的工具的价格预计不低于1000美元）。

⑤ 列出所有有关理解和描述或解释模型必要性的假设。（如，模型及其结果应以怎样的方式呈现给高层管理者/负责人。）。

局限

① 检察总的局限（如，法律问题，预算，时间表和资源）。

② 检查获得数据资源的权力（使用权的限制，得到的密码）

③ 检查数据的技术可操作性（操作系统，数据管理系统，文件或数据库的格式）

④ 检查相关知识的可得性。

⑤ 检查预算限制（固定成本，执行成本等等）

切记！ 假设列表也包括项目之初的所有假设，例如，项目的出发点。

Output Risks and contingencies

结果 **风险和意外**

列出风险，也就是说可能发生的事件，紧迫的时间表，成本或者结果等等。列出相应的处理意外的方案：具体采取什么样的措施来避免或减小影响，或者从可预见的风险事件中恢复到正常状态。

业务活动 **识别风险**

① 明确商业风险（如，竞争者首先产生好的结果）

② 明确组织风险（如，要求项目的部门没有了资金的支持）

③ 明确财政风险（如，进一步的资金支持有赖于最初的数据挖掘的结果）

④ 明确技术风险。

⑤ 明确依赖数据和数据资源的风险（如，很差的质量和覆盖）

制定意外处理方案

① 确定每一个风险可能发生的环境。

② 制定意外处理方案。

结果 **术语**

编辑一套与本项目相关的术语表。它至少应包括以下两个部分：

(1) 一套相关的商业术语表，它将构成项目可利用的商业理解的一部分。

(2) 一套有关数据挖掘术语的术语表，并经与商业问题有关的事例解释说明。

业务活动

① 检察原来的术语标的实用性，否则重新起草一分术语表。

② 同领域专家交流，了解他们的专业术语。

③ 熟悉商业术语。

结果 **成本和收益**

准备项目的成本-收益分析，将项目成本与项目成功后企业的获利进行比较。

好主意！成本收益的比较要尽可能的详尽，因为这样将使得这个商业案例更加成功。

业务活动

① 估算数据收集的成本。

② 估算设计并实施一个解决方案的成本。

③ 确定运用某一个解决方案后的收益（如，改进的顾客满意度，ROI和收益的增长）。

④ 估算运营成本。

注意！切记识别那些潜在的成本，例如，需要重复进行的数据提炼和准备工作，工作流程的改变以及培训的培训时间等等。

1.3 确定数据挖掘目标

任务 **确定数据挖掘目标**

商业目标是以相关的商业术语表现出来的目标。而数据挖掘的目标则是以技术形式表现出来的目标。例如，一个商业目标可能是“增加现有消费者的销售类别catalog”。而数据挖掘的目标则可能是“根据消费者过去三年的购买情况、人口统计信息以及商品的价格，预测他今年将要购买多少商品”。

***结果* 数据挖掘的目标**

描述项目预计能够实现的达成商业目标的结果。注意这些结果通常为*技术上*的结果。

业务活动

① 将企业问题转化为数据挖掘的目标（如，一个市场策略需要将消费者进行划分，从而确定此策略的实施目标；市场分割的层次和大小要详细说明）

② 详细说明数据挖掘问题的种类problem type（如，分类，描绘，预测和聚类）。 有关数据挖掘问题种类的细节描述，请祥见附录V.2。

好主意！ 重新定义问题将是明智之举。例如，把产品持久力作为模型的预测目标而不是消费者持久力，因为定位在消费者的持久力很可能太迟而不能影响结果。It may be wise to re-define the problem. For example, modeling product retention rather than customer retention since targeting customer retention may be too late to affect the outcome!

***结果* 数据挖掘成功的标准**

用相关的技术术语定义项目成功的标准，例如要达到一定水平要求的预测精确度，或者是购买倾向的一定程度的“提高”等等。至于商业成功标准，需要以主观的形式加以陈述，同时应明确指出做出这一主观判断的群体或个人。

业务活动

① 详细说明模型评估的标准（如，模型的精度，性能和复杂性）。

② 定义评估标准的基准。

③ 详细说明主观的评判标准（如，模型的解释能力，模型对数据和市场的洞察能力）。

注意！ 切记数据挖掘标准不同于前面定义的企业成功的标准。

切记从项目之初开始计划部署方案是明智之举。

1.4 制定项目计划

***任务* 制定项目计划**

陈述为了实现数据挖掘目标进而实现商业目标的预计的计划。计划方案应尽可能详尽的陈述接下来的项目中所要进行的预期的一系列举措，包括数据挖掘的工具以及技术的初步选择等等。

***结果* 项目计划**

陈列出项目过程中所要执行的各个阶段，包括各阶段持续的时间，所需要的资源，输入的数据，输出的结果及其附属物。明确数据挖掘过程中那些可能需要大规模重复的地方，例如建模和评估阶段的循环操作。作为项目计划的一个组成部分，分析时间进度与风险之间的相互依赖关系也是非常重要的。项目计划中，应明确的标明这些分析的结果，比较理想的状态是能够配有风险发生时的解决方案和补救措施。

记住：尽管这一任务仅仅是直接命名了**项目计划**，但在整个项目进程中需要不断的参考和回顾这一任务的结果，至少是在开始一项新的任务或者重新进行某项重复工作或活动时参考项目计划。

业务活动

① 定义最初的项目进程计划，并同所有的参与者讨论其可行性。

② 把所有已经确定的目标和所选择的技术结合起来，组成一套连贯的程序，进而解决商业问题，符合企业成功的标准。

③ 估计完成和部署解决方案的工作强度和资源（在估计数据挖掘的时间进度的同时，考虑到其他人的经验也是非常有益的。例如，对于通盘的数据挖掘进程而言，通常假定数据准备阶段需要全部工作时间和强度的50-70%，数据理解阶段占20-30%，而只有10-20%的时间和精力花在每一次的建模，评估和商业理解阶段，5-10%的工作时间和强度集中在部署运用阶段。）。

④ 确定关键步骤。

⑤ 标明决策点。

⑥ 标明回顾点。

⑦ 确定需要重复的主要点。

工具和技术的初步评估

在第一阶段的末尾，项目组还应该进行一次工具和技术的初步评估工作。例如，在这儿，您应该选择一种数据挖掘工具来支持项目进程中不同阶段所使用的各种方法。项目进程的早期对工具和技术进行评估是非常重要的，因为工具和技术的选择很有可能影响到整个项目的进行。

业务活动

① 制作一个工具和技术选择标准的列表（或者使用已有的可用列表）

② 选择可能的工具和技术。

③ 评估技术的适用性。

④ 根据可选方案的评估，检查并给予那些可应用的技术以优先权。

2 数据理解

2.1 收集原始数据

***任务* 收集原始数据**

获得项目资源列表中的项目数据（或者使用权）。有必要的话，还要进行数据的装载工作。例如，如果你打算使用某一特定的工具进行数据理解，就需要将数据安装在这一工具之中。

***结果* 初步的数据收集报告**

列出项目所需的所有数据以及对数据的更加详细信息的要求。此外，数据收集报告还要详细说明数据中哪些变量相对于其他变量而言更加重要。

由于数据资源之间的不一致性，因此经多个数据资源整合后所得到的数据可能存在着单个数据资源不存在的问题。故切记对所有数据的质量进行评估，不仅包括那些单个的数据资源，还要包括数据资源整合后的数据。

***业务活动* 有关数据要求的计划**

① 计划需要哪些信息？（如，仅仅是已有的变量，还是要额外的信息）。

② 检查是否所有的信息都确实可以利用来实现数据挖掘的目标。

选择标准

① 详细说明选择的标准（例如，对于具体的数据挖掘目标都需要哪些必要的变量？哪些变量与数据挖掘目标不相关？我们要用已选的技术处理多少变量？）

② 选择interest的数据表格/文件。Select tables/files of interest.

③ 选择一个数据表格/文件内的数据。Select data within a table/file.

④ 即使数据在使用期内一直都是有效的，但仍然要考虑使用数据的时间（例如，也许数据的可用期为18个月，但我们也许只需要使用12个月而已）。

注意！ 注意当来自不同数据源的数据进行整合时，有可能会带来数据质量问题（例如，地址文件和其相应的消费者基本资料合并时，可能会出现数据格式或者数据无效等方面的矛盾）

数据插入

① 如果数据中包含有自由free的文本记录，为了建模，我们需要对其编码吗？或者我们需要把一些具体的记录进行分组吗？

② 怎么样获得缺失的变量？

③ 描述如何提炼数据。

好主意！ 注意有些数据信息是非电子版的（如，人们people，印刷文本等）。

注意有时可能需要重新处理数据（如，时间序列的数据，加权的平均值等等）。

2.2 描绘数据

***任务* 描绘数据**

检验所得数据的“总的”或者说“表面的”特征并报告其结果。

***结果* 数据描绘报告**

描绘所得数据，包括：数据格式，数据性质（例如每一个数据表格中记录的条数和变量的数目），变量特征以及任何其他表面特征。数据是否已经满足一系列相关的需要？

***业务活动* 数据容量分析**

① 确认数据及其获取方法。

② 获得数据资源。

③ 使用合适的统计分析方法。

④ 报告数据表格以及它们之间的关系。

⑤ 检查数据的容量，数据集的数目以及复杂性。

⑥ 检查数据是否包含自由free的文本记录。

变量的种类和数值

① 检查变量的可得性和有效性。

② 检查变量的种类（数值型，字符型，定类变量等）

③ 检查变量值的范围。

④ 分析变量间的关系。

⑤ 从商业角度了解每一个变量及其值的含义。

⑥ 对每一个变量进行基本的统计分析（如，计算其分布，均值，最大值，最小值，标准差，方差，众数，偏斜率等）

⑦ 分析基本的统计量，并把其结果与商业含义联系起来。

⑧ 变量是否与具体的数据挖掘目标相关联？

⑨ 变量含义是否始终一致？

⑩ 询问领域专家对于变量关系的见解。

11 是否有必要对数据进行权衡balance？（有赖于选用的建模技术）

关键变量keys

① 分析关键值keys的关系。Analyze key relations.

② 检查数据表中关键值keys重复出现的个数。

回顾假设和目标

① 如有必要更新假设列表。

2.3 探索数据

***任务* 探索数据**

本任务将处理数据挖掘的问题，这些问题可以以质疑的方式被提出，也可以是一些显而易见的问题或者是以报告的形式加以陈述。这些分析将直接涉及到数据挖掘的目标；同时也将有助于进一步完善数据描述，改进报告质量，并为进一步的数据分析提供了数据转换和其他的数据准备。

***结果* 数据探索报告**

描绘数据探索的结果，包括最初的发现或者初期的假设以及这些发现对余下项目的影响。此外，报告也可以包括一些图表，用来展示数据的特征或者产生有趣的数据子集以作进一步的检测。

***业务活动* 数据探索**

① 详细分析引人注意的变量特征（如，对引人好奇的潜在人群进行基本的统计分析）

② 识别潜在人群的特征。

提出设想以作后面的分析

① 思考和评估数据描述报告中的信息和发现。

② 提出假设并确定方案。

③ 如可能将假设变成数据挖掘的目标。

④ 阐明数据挖掘的目标或这是他们更加明晰。盲目的研究是绝对不可取的，除非此研究有一个明确的方向直指企业目标。

⑤ 做一些基本的分析检验假设。

2.4 检验数据质量

***任务* 检验数据质量**

检验数据的质量，列举有关问题，例如：数据是否完整（它是否覆盖了所需要的全部case）?数据是否正确或者数据是否包含错误？如果有错误，是否常见？数据中是否有缺失值？如果有缺失值，它们以何种方式出现，在哪里出现，是否常见？

***结果* 数据质量报告**

列举数据质量检测的结果；如果存在数据质量问题，列出可能的解决办法。

业务活动

① 确认具体的数值并将其含义编目。

回顾关键值keys，变量

① 检验数值的范围（例如，检验所有可能的数值是否已经全面）。

② 检验关键值。keys.

③ 变量含义与变量值是否一致fit together?

④ 确认缺失变量和空白栏。

⑤ 缺失值的含义。

⑥ 检验那些具有相似含义却有不同值的变量（如，低脂肪，节食）。

⑦ 检验数据值的拼写（如，相同的数据值有时以小写字母开头，而有时又以大写字母开头）。

⑧ 检验偏差，确认这一偏差是干扰信息还是预示了一个引人好奇的现象。

⑨ 检验数值的合理性，例如，当所有数据栏都有相同或相似的值时，需要检验其合理性。

好主意！ 重新检查那些给出了与常识相矛盾的答案的变量（如，拥有高收入的青少年）。通过视觉图，柱状图等来显示数据中的矛盾信息。

无格式文件中的数据质量

① 若数据存储在无格式文件中，确认其分隔符并检查数据中是否所有的变量都使用了这一分隔符。

② 若数据存储在无格式文件中，检验每一条记录的分栏数目。他们是否一致？

数据源之间的干扰与矛盾。

① 检查不同数据源之间的矛盾和冗余信息。

② 计划如何处理这些干扰信息。

③ 检测干扰信息的类型以及受到影响的变量。

好主意！ 有时有必要剔除一些数据，因为这些数据并未表现出任何正面或负面的行为（例如，检验消费者的贷款行为，剔除那些从未贷款的消费者，没有提供房屋贷款的消费者，或者抵押到期的消费者等等）。回顾所有的假设，检验数据是否还有效，或者未提供相关的当前信息或数据。

3 数据准备

结果 **数据集**

本阶段产生数据集（或数据集合），以作建模或者项目的其他分析工作之用。

数据集描述

描绘用于建模或者项目主要分析工作的数据集。

3.1 选择数据

任务 **选择数据**

确定作分析之用的数据。选择的标准包括与数据挖掘目标的相关性，数据质量以及技术限制，例如对数据大小或数据种类的要求等。

结果 **选择和剔除的基本原理**

列出被选择/剔除的数据及其原因。

业务活动

① 额外的收集适当的数据（来自于不同的数据源——既包括内部的，也包括外部的数据）。

② 进行显著性和相关性检验，确定要选择的变量栏。

③ 根据数据质量检验以及数据探索的经验，重新考虑“数据选择标准”（见任务2.1）（例如，可能希望选择/剔除其它数据集合。）

④ 根据建模经验重新考虑“数据选择标准”（见任务2.1）（例如，建模评估结果显示可能还需要其它数据集）。

⑤ 选择不同的数据子集（如，不同的变量，只符合某些特定条件的数据等等）。

⑥ 考虑抽样技术的使用（如，一个较快的解决方案可能需要减小检验数据集的大小，或者某一个工具不能处理整个数据集，需要将数据集分成检验子集和训练子集）。

将样本加权，赋予不同变量或同一变量的不同值以不同的权重，有时也许是有用的。.

⑦ 记录选择或剔除的原因。

⑧ 检验对数据抽样的可行的技术。

好主意！基于数据的选择标准，确定是否某一个变量或者某几个变量相对与其他变量更加重要，相应的给这些变量加权。并根据文本（如应用软件application，工具等等），确定如何处理加权。

3.2 清理数据

任务 **清理数据**

根据所选分析技术的要求，将数据质量提高到相应的水平。这一工作可能涉及选择的干净的数据子集，插入适当的默认值或者使用更加出色的处理技术，例如通过建模估测缺失值。

结果 **数据清理报告**

针对*检测数据质量*任务中报告的数据质量问题，描绘本阶段任务所采取的决策和行动。

此外，报告中还应指出那些仍然比较显著的数据质量问题。如果这些数据还要在数据挖掘中使用，指出它对结果可能产生的影响。

业务活动

① 重新考虑如何处理干扰的观测类型。

② 纠正，剔除或者忽视那些干扰信息。

③ 确定如何处理异常值及其含义。异常值可能会导致许多奇怪的结果产生，对此，研究者应仔细检查。例如，在一个调查中，当某些问题没有被问到，或者被访者未回答时，其结果中就会出现异常值，因为对于“无回答”数据，往往用“99”代替，而对于某些变量，“99”即为异常值，如婚姻状况，政治关系等。再有，有的时候数据被删节，也会如出现异常值。例如，对于年龄为“100”岁的人或者时速为“100,000”公里的汽车，其数值却被“00”代替，即为异常值。

④ 根据数据清理的结果，重新考虑“数据选择标准”（参见任务2.1）（如，可能希望选择或剔除其他的数据集）

好主意！ 某些变量栏和树据挖掘的目标无关，因此其干扰信息也就没有意义。但如果因此而忽略这些干扰信息的话，一定要将其完整的记录下来，以免后面情况会有所改变。

3.3 构造数据

任务 **构造数据**

这一任务包括建设性的数据准备工作，例如衍生属性的产生，全新的记录的产生或者已存属性的值转换。

业务活动

① 根据项目已有的工具列表，核查可用的数据构造方法。

② 从工具的全方面考虑，检验此种数据构造方法是否最好。（例如，有效性，准确性，可重复性）。

③ 参照构造数据的经验，重新考虑“数据选择标准”（参见任务2.1）（如，可能希望选择或剔除其他的数据集）

结果 **衍生变量**

衍生变量是新生的变量，来自于同一条记录中的一个或多个变量。例如，面积=长*宽。那么为什么我们要在数据挖掘研究中构造新的变量呢？显然并不是只有那些从数据库或者其他数据源中得来得数据才能用来建构模型，我们之所以要构造新的变量，是因为：

① 背景知识告诉我们，某些事实是非常重要的，尽管目前我们无法用变量将其表现出来，但是我们仍然需要将其以某种方式表现出来。

② 建模的运算法则只能处理某一特定类型的数据，例如，当我们正在使用线性回归处理数据时，却怀疑数据中存在着某些非线性因素并不包括在这一模型之中。

③ 建模阶段的结果显示，有某些因素并未包括在模型之中。

业务活动 **衍生变量**

① Decide if any attribute should be normalized (e.g. when using a clustering algorithm with age and income in lire, the income will dominate).

② Consider adding new information on the relevant importance of attributes by adding new attributes (for example, attribute weights, weighted normalization).

③ How can missing attributes be constructed or imputed? [Decide type of construction (e.g., aggregate, average, induction)].

④ Add new attributes to the accessed data.

Good idea! Before adding Derived Attributes, try to determine if and how they ease the model process or facilitate the modeling algorithm. Perhaps “income per head” is a better/easier attribute to use that “income per household.” Do not derive attributes simply to reduce the number of input attributes.

Another type of derived attribute is single-attribute transformations, usually performed to fit the needs of the modeling tools.

Activities Single-attribute transformations

① Specify necessary transformation steps in terms of available transformation facilities (for example. change a binning of a numeric attribute).

② Perform transformation steps.

Hint! Transformations may be necessary to transform ranges to symbolic fields (e.g. ages to age ranges) or symbolic fields (“definitely yes,” “yes,” “don’t know,” “no”) to numeric values. Modeling tools or algorithms often require them.

Output Generated records

Generated records are completely new records, which add new knowledge or represent new data that is not otherwise represented, e.g., having segmented the data, it may be useful to generate a record to represent the prototypical member of each segment for further processing.

Activities Check for available techniques if needed (e.g., mechanisms to construct prototypes for each segment of segmented data).

3.4 Integrate data

Task Integrate data

These are methods whereby information is combined from *multiple* tables or other information sources to create new records or values.

Output Merged data

Merging tables refers to joining together two or more tables that have different information about the same objects. At this stage it may also be advisable to generate new records. It may also be recommended to generate aggregate values.

Aggregation refers to operations where new values are computed by summarizing information from multiple records and/or tables.

Activities

① Check integration facilities if they are able to integrate the input sources as required.

② Integrate sources and store result.

③ Reconsider Data Selection Criteria (See Task 2.1) in light of experiences of data integration (i.e. may wish include/exclude other sets of data).

Good idea! Remember that some knowledge may be contained in non-electronic format.

3.5 Format data

Task Format data

Formatting transformations refer to primarily *syntactic* modifications made to the data that do not change its meaning, but might be required by the modeling tool.

Output Reformatted data

Some tools have requirements on the order of the attributes, such as the first field being a unique identifier for each record or the last field being the outcome field the model is to predict.

Activities Rearranging attributes

Some tools have requirements on the order of the attributes, such as the first field being a unique identifier for each record or the last field being the outcome field the model is to predict.

Reordering records

It might be important to change the order of the records in the dataset. Perhaps the modeling tool requires that the records be sorted according to the value of the outcome attribute.

Reformatted within-value

① These are purely syntactic changes made to satisfy the requirements of the specific modeling tool.

② Reconsider Data Selection Criteria (See Task 2.1) in light of experiences of data cleaning (i.e. may wish include/exclude other sets of data).

4 Modeling

4.1 Select modeling technique

Task **Select modeling technique**

As the first step in modeling, select the actual modeling technique that is to be used initially. If multiple techniques are applied, perform this task for each technique separately. It should not be forgotten that not all tools and techniques are applicable to each and every task. For certain problems, only some techniques are appropriate (See Appendix V.2 where techniques appropriate for certain data mining problem types are discussed in more detail). From among these tools and techniques there are "Political Requirements" and other constraints, which further limit the choice available to the miner.

It may be that only one tool or technique is available to solve the problem in hand - and even then the tool may not be the absolutely technical best for the problem in hand.

Figure 10:

Universe of Techniques

Output **Modeling technique**

Record the actual modeling technique that is used.

Activities Decide on appropriate technique for exercise bearing in mind the tool selected.

Output **Modeling assumptions**

Many modeling techniques make specific assumptions about the data, data quality or the data format.

Activities

① Define any built-in assumptions made by the technique about the data (e.g. quality, format, distribution).

② Compare these assumptions with those in the Data Description Report.

③ Make sure that these assumptions hold and step back to the Data Preparation Phase if necessary.

4.2 Generate test design

Task **Generate test design**

Prior to building a model, a procedure needs to be defined to test the model's quality and validity. For example, in supervised data mining tasks such as classification, it is common to use error rates as quality measures for data mining models. Therefore the test design specifies that the dataset should be separated into training and test set, the model is built on the training set and its quality estimated on the test set.

Output **Test design**

Describe the intended plan for training, testing and evaluating the models. A primary component of the plan is to decide how to divide the available dataset into training data, test data and validation test sets.

Activities

① Check existing test designs for each data mining goal separately.

② Decide on necessary steps (number of iterations, number of folds etc.).

③ Prepare data required for test.

4.3 Build model

Task Build model

Run the modeling tool on the prepared dataset to create one or more models.

Output Parameter settings

With any modeling tool, there are often a large number of parameters that can be adjusted. List the parameters and their chosen values, along with the rationale for the choice.

Activities

① Set initial parameters.

② Document reasons for choosing those values.

Output Models

Run the modeling tool on the prepared dataset to create one or more models.

Activities

① Run the selected technique on the input dataset to produce the model.

② Post-process data mining results (e.g. editing rules, display trees).

Output Model description

Describe the resultant model and assess its expected accuracy, robustness and possible shortcomings. Report on the interpretation of the models and any difficulties encountered.

Activities

① Describe any characteristics of the current model that may be useful for the future.

② Record parameter settings used to produce the model.

③ Give a detailed description of the model and any special features.

④ For rule-based models, list the rules produced plus any assessment of per-rule or overall model accuracy and coverage.

⑤ For opaque models, list any technical information about the model(such as neural network topology) and any behavioral descriptions produced by the modeling process (such as accuracy or sensitivity).

⑥ Describe the model's behavior and interpretation.

⑦ State conclusions regarding patterns in the data (if any); sometimes the model reveals important facts about the data without a separate assessment process (e.g. that the output or conclusion is duplicated in one of the inputs).

4.4 Assess model

Task Assess model

The model should now be assessed to ensure that it meets the data mining success criteria and the passes the desired test criteria. This is a purely technical assessment based on the outcome of the modeling tasks.

Output Model assessment

Summarize results of this task, list qualities of generated models (e.g. in terms of accuracy) and rank their quality in relation to each other.

Activities

Evaluate result with respect to evaluation criteria

Good idea! “Lift Tables” and “Gain Tables” can be constructed to determine how well the model is predicting.

① Test result according to a test strategy (e.g.: Train and Test, Crossvalidation, bootstrapping etc.).

② Compare evaluation results and interpretation.

③ Create ranking of results with respect to success and evaluation criteria

④ Select best models.

⑤ Interpret results in business terms (as far as possible at this stage).

⑥ Get comments on models by domain or data experts.

⑦ Check plausibility of model.

⑧ Check impacts for data mining goal.

⑨ Check model against given knowledge base to see if the discovered information is novel and useful.

10 Check reliability of result.

11 Analyze potentials for deployment of each result.

12 If there is a verbal description of the generated model (e.g. via rules), assess the rules; are they logical, are they feasible, are there too many or too few, do they offend common sense?

13 Assess results.

14 Get insights into why a certain modeling technique and certain parameter settings lead to good/bad results.

Output Revised parameter settings

According to the model assessment, revise parameter settings and tune them for the next run in task ‘Build Model.’ Iterate model building and assessment until you find the best model.

Activities

Adjust parameters to give better model.

5 Evaluation

Previous evaluation steps dealt with factors such as the accuracy and generality of the model. This step assesses the degree to which the model meets the business objectives and seeks to determine if there is some business reason why this model is deficient. It compares results with the evaluation criteria defined at the start of the project.

A good way of defining the total outputs of a data mining project is to use the equation: RESULTS = MODELS + FINDINGS

In this equation we are defining that the total output of the data mining project is not just the models (although they are, of course, important) but also findings which we define as anything *(apart from the model)* that is important in meeting objectives of the business (or important in leading to new questions, line of approach or side effects (e.g. data quality problems uncovered by the data mining exercise). Note: although the model is directly connected to the business questions, the findings need not be related to any questions or objective, but are important to the initiator of the project.

5.1 Evaluate results

Task Evaluate results

Previous evaluation steps dealt with factors such as the accuracy and generality of the model. This step assesses the degree to which the model meets the business objectives and seeks to determine if there is some business reason why this model is deficient. Another option of evaluation is to test the model(s) on test applications in the real application if time and budget constraints permit.

Moreover, evaluation also assesses other data mining results generated. Data mining results cover models which are necessarily related to the original business objectives and all other findings which are not necessarily related to the original business objectives but might also unveil additional challenges, information or hints for future directions.

Output Assessment of data mining results with respect to business

success criteria

Summarize assessment results in terms of business success criteria including a final statement whether the project already meets the initial business objectives.

Activities

① Understand the data mining result.

② Interpret the results in terms of the application.

③ Check impacts for data mining goal.

④ Check the data mining result against the given knowledge base to see if the discovered information is novel and useful.

⑤ Evaluate and assess result with respect to business success criteria i.e. has the project achieved the original Business Objectives?

⑥ Compare evaluation results and interpretation.

⑦ Create ranking of results with respect to business success criteria.

⑧ Check impacts of result for initial application goal.

⑨ Are there new business objectives to be addresses later in the project or in new projects?

⑩ States conclusions for future data mining projects.

Output Approved models

After model assessment with respect to business success criteria, you eventually get approved models if the generated models meet the selected criteria.

5.2 Review process

Task Review process

At this point the resultant model appears to be satisfactory and appears to satisfy business needs. It is now appropriate to make a more thorough review of the data mining engagement in order to determine if there is any important factor or task that has somehow been overlooked. At this stage of the Data Mining exercise, the Process Review takes on the form of a Quality Assurance Review.

Output Review of process

Summarize the process review and give hints for activities that have been missed and/or should be repeated.

Activities

① Give an overview of the data mining process used

② Analyze data mining process For each stage of the process:

③ Was it necessary in retrospect?

④ Was it executed optimally?

⑤ In what ways could it be improved?

⑥ Identify failures.

⑦ Identify misleading steps.

⑧ Identify possible alternative actions, unexpected paths in the process.

⑨ Review data mining results with respect to business success criteria.

5.3 Determine next steps

Task Determine next steps

According to the assessment results and the process review, the project decides how to proceed at this stage. The project needs to decide whether to finish this project and move onto deployment or whether to initiate further iterations or whether to set up new data mining projects.

Output List of possible actions

List possible further actions along with the reasons for and against each option.

Activities

① Analyze potential for deployment of each result.

② Estimate potential for improvement of current process.

③ Check remaining resources to determine if they allow additional process iterations (or whether additional resources can be made available).

④ Recommend alternative continuations.

⑤ Refine process plan.

Output Decision

Describe the decision as to how to proceed along with the rationale.

Activities

① Rank the possible actions.

② Select one of the possible actions.

③ Document reasons for the choice.

6 Deployment

6.1 Plan deployment

Task Plan deployment

This task takes the evaluation results and concludes a strategy for deployment of the data mining result(s) into the business.

Output Deployment plan

Summarize deployment strategy including necessary steps and how to perform them.

Activities

① Summarize deployable results.

② Develop and evaluate alternative plans for deployment.

③ Decide for each distinct knowledge or information result.

④ How will the knowledge or information be propagated to its users?

⑤ How will the use of the result be monitored or its benefits measured (where applicable)?

⑥ Decide for each deployable model or software result.

⑦ How will the model or software result be deployed within the organization's systems?

⑧ How will its use be monitored and its benefits measured (where applicable)?

⑨ Identify possible problems when deploying the data mining results (pitfalls of the deployment).

6.2 Plan monitoring and maintenance

Task Plan monitoring and maintenance

Monitoring and maintenance are important issues if the data mining result becomes part of the day-to-day business and its environment. A careful preparation of a maintenance strategy helps to avoid unnecessarily long periods of incorrect usage of data mining results. In order to monitor the deployment of the data mining result(s), the project needs a detailed plan on the monitoring process. This plan takes into account the specific type of deployment.

Output Monitoring and maintenance plan

Summarize monitoring and maintenance strategy including necessary steps and how to perform them.

Activities

① Check for dynamic aspects (i.e. what things could change in the environment?).

② How will accuracy be monitored?

③ When should the data mining result or model not be used any more? Identify criteria (validity, threshold of accuracy, new data, change in the application domain, etc.)? What should happen if the model or result could no longer be used? (Update model, set up new data mining project, etc.).

④ Will the business objectives of the use of the model change over time? Fully document the initial problem the model was attempting to solve.

⑤ Develop monitoring and maintenance plan.

6.3 Produce final report

Task Produce final report

At the end of the project, the project leader and his team write up a final report. It depends on the deployment plan, if this report is only a summary of the project and its experiences or if this report is a final presentation of the data mining result(s).

Output Final report

At the end of the project, there will be (at least one) final report where all the threads are brought together. As well as identifying the results obtained, the report should also describe the process, show which costs have been incurred, define any deviations from the original plan, describe implementation plans and make any recommendations for future work. The actual detailed content of the report depends very much on the audience for the particular report.

Activities

① Identify what reports are needed (slide presentation, management summary, detailed findings, explanation of models, etc.).

② Analyze how well initial data mining goals have been met.

③ Identify target groups for report.

④ Outline structure and contents of report(s).

⑤ Select findings to be included in the reports.

⑥ Write a report.

Output Final presentation

As well as a Final Report, it may be necessary to make a Final Presentation to summarize the project - maybe to the management sponsor, for example. The Presentation normally contains a subset of the information contained in the Final Report, but structured in a different way.

Activities

① Decide on target group for final presentation (will they already have received final report?).

② Select which items from final report should be included in final presentation.

6.4 Review project

Task Review project

Assess what went right and what went wrong, what was done well and what needs to be improved.

Output Experience documentation

Summarize important experiences made during the project. For example, pitfalls, misleading approaches or hints for selecting the best-suited data mining techniques in similar situations could be part of this documentation. In ideal projects, experience documentation covers also any reports that have been written by individual project members during the project phases and their tasks.

Activities

① Interview all significant people involved in the project and ask them about their experiences during the project.

② If end users in the business work with the data mining result(s), interview them: are they satisfied? What could have been done better? Do they need additional support?

③ Summarize feedback and write the experience documentation

④ Analyze the process (things that worked well, mistakes made, lessons learned, etc.).

⑤ Document the specific data mining process (How can the results and the experience of applying the model be fed back into the process?).

⑥ Abstract from details to make the experience useful for future projects.

IV The CRISP-DM outputs

This section contains brief descriptions of the purpose and the contents of the most important reports. Here, we focus on reports that are meant to communicate the results of a phase to people not involved in this phase (and possibly not involved in this project).

These are not necessarily identical to the outputs as described in the reference model and the user guide. The purpose of the outputs is mostly to document results *while performing* the project.

1 Business understanding

The results of the Business Understanding phase can be summarized in one report.

We suggest the following sections:

Background

The Background provides a basic overview of the project context. This lists what area the project is working in, what problems have been identified and why data mining appears to provide a solution.

Business objectives and success criteria

Business Objectives describe what the goals of the project are in business terms. For each

objective, Business Success Criteria, i.e. explicit measures for determining whether or not the project succeeded in its objectives, should be provided. This section should also list objectives that were considered but rejected. The rationale of the selection of objectives should be given.

Inventory of resources

The Inventory of Resources aims to identify personnel, data sources, technical facilities and other resources that may be useful in carrying out the project.

Requirements, assumptions and constraints

This output lists general requirements about how the project is executed, type of project results, assumptions made about the nature of the problem and the data being used and constraints imposed on the project.

Risks and contingencies

This output identifies problems that may occur in the project, describes the consequences and states what action can be taken to minimize the effect.

Terminology

The Terminology allows people unfamiliar with the problems being addressed by the project to become more familiar with them.

Costs and benefits

This describes the costs of the project and predicted business benefits if the project is successful (e.g. return on investment). Other less tangible benefits (e.g. customer satisfaction) should also be highlighted.

Data mining goals and success criteria

The data mining goals state the results of the project that enable the achievement of the business objectives. As well as listing the probable data mining approaches, the success criteria for the results should also be listed in data mining terms.

Project plan

This lists the stages to be executed in the project, together with duration, resources required, inputs, outputs and dependencies. Where possible it should make explicit the large-scale iterations in the data mining process, for example repetitions of the modeling and evaluation phases.

Initial assessment of tools and techniques

This section gives an initial view of what tools and techniques are likely to be used and how. It describes the requirements for tools and techniques, lists available tools and techniques and matches them to requirements.

2 Data understanding

The results of the Data Understanding phase are usually documented in several reports.

Ideally, these reports should the written while performing the respective tasks. The reports describe the datasets that are explored during data understanding. For the final report, a summary of the most relevant parts is sufficient.

Initial data collection report

This report describes how the different data sources identified in the inventory were captured and extracted.

Topics to be covered:

① Background of data.

② List of data sources with broad area of required data covered by each.

③ For each data source, method of acquisition or extraction.

④ Problems encountered in data acquisition or extraction.

Data description report

Each dataset acquired is described.

Topics to be covered:

① Each data source described in detail.

② List of tables (may be only one) or other database objects.

③ Description of each field including units, codes used, etc.

Data exploration report

Describes the data exploration and its results.

Topics to be covered:

① Background including broad goals of data exploration.

For each area of exploration undertaken:

① Expected regularities or patterns.

② Method of detection.

③ Regularities or patterns found, expected and unexpected.

④ Any other surprises.

⑤ Conclusions for data transformation, data cleaning and any other pre-processing.

⑥ Conclusions related to data mining goals or business objectives.

⑦⑧ Summary of conclusions.

Data quality report

This report describes the completeness and accuracy of the data.

Topics to be covered:

① Background including broad expectations about data quality.

For each dataset:

① Approach taken to assess data quality.

② Results of data quality assessment.

③ Summary of data quality conclusions.

3 Data preparation

The reports in the data preparation phase focus on the pre-processing steps that produce the data to be mined.

Dataset description report

This provides a description of the dataset (after pre-processing) and the process by which it was produced.

Topics to be covered:

① Background including broad goals and plan for pre-processing.

② Rationale for inclusion/exclusion of datasets.

For each included dataset:

① Description of the pre-processing, including the actions that were necessary to address any data quality issues.

② Detailed description of the resultant dataset, table by table and field by field.

③ Rationale for inclusion/exclusion of attributes.

④ Discoveries made during pre-processing and any implications for further work.

⑤ Summary and conclusions.

4 Modeling

The outputs produced during the Modeling phase can be combined into one report.

We suggest the following sections:

Modeling assumption

This section defines *explicitly* any assumptions made about the data and any assumptions that are implicit in the modeling technique to be used.

Test design

This section describes how the models are built, tested and evaluated.

Topics to be covered:

Background - outlines the modeling undertaken and its relation to the data

mining goals.

For each modeling task:

① Broad description of the type of model and the training data to be used.

② Explanation of how the model will be tested or assessed.

③ Description of any data required for testing.

④ Plan for production of test data if any.

⑤ Description of any planned examination of models by domain or data experts.

⑥ Summary of test plan.

Model description

This report describes the delivered models and overviews the process by which they were produced.

Topics to be covered:

① Overview of models produced.

For each model:

① Type of model and relation to data mining goals.

② Parameter settings used to produce the model.

③ Detailed description of the model and any special features.

For example:

① For rule-based models, list the rules produced plus any assessment of per-rule or overall model accuracy and coverage.

② For opaque models, list any technical information about the model (such as neural network topology) and any behavioral descriptions produced by the modeling process (such as accuracy or

sensitivity).

③ Description of Model’s behavior and interpretation.

④ Conclusions regarding patterns in the data (if any); sometimes the model will reveal important facts about the data without a separate assessment process (e.g. that the output or conclusion is duplicated in one of the inputs).

⑤ Summary of conclusions.

Model assessment

This section describes the results of testing the models according to the test design.

Topics to be covered:

① Overview of assessment process and results including any deviations from the test plan.

For each model:

② Detailed assessment of model including measurements such as accuracy

and interpretation of behavior.

③ Any comments on models by domain or data experts.

④ Summary assessment of model.

⑤ Insights into why a certain modeling technique and certain parameter settings led to good/bad results.

⑥ Summary assessment of complete model set.

5 Evaluation

Assessment of data mining results with respect to business success criteria

This report compares the data mining results with the business objectives and the business success criteria.

Topics to be covered:

n Review of Business Objectives and Business Success Criteria (which may have changed during and/or as a result of data mining).

For each Business Success Criterion:

① Detailed comparison between success criterion and data mining results.

② Conclusions about achievability of success criterion and suitability of data mining process.

③ Review of Project Success; has the project achieved the original Business Objectives?

④ Are there new business objectives to be addressed later in the project or in new projects?

⑤ Conclusions for future data mining projects.

Review of process

This section assesses the effectiveness of the project and identifies any factors that may have been overlooked that should be taken into consideration if the project is repeated.

List of possible actions

This section makes recommendations regarding the next steps in the project.

6 Deployment

Deployment plan

This section specifies the deployment of the data mining results.

Topics to be covered:

① Summary of deployable results (derived from Next Steps report).

② Description of deployment plan.

Monitoring and maintenance plan

The monitoring and maintenance plan specifies how the deployed results are to be maintained.

Topics to be covered:

n Overview of results deployment and indication of which results may require

updating (and why).

For each deployed result:

① Description of how updating will be triggered (regular updates, trigger event, performance monitoring).

② Description of how updating will be performed.

③ Summary of the results updating process.

Final report

The final report is used to summarize the project and its results.

Contents:

① Summary of Business Understanding: background, objectives and success criteria.

② Summary of data mining process.

③ Summary of data mining results.

④ Summary of results evaluation.

⑤ Summary of deployment and maintenance plans.

⑥ Cost/benefit analysis.

⑦ Conclusions for the business.

⑧ Conclusions for future data mining.

7 Summary of dependencies

The following table summarizes the main inputs to the deliverables. This does not mean that only the inputs listed should be considered - for example, the business objectives should be pervasive to all deliverables. However, the deliverables should address specific issues raised by their inputs.

V Appendix

1 Glossary/terminology

Activity

Part of a task in User Guide, describes actions to perform a task.

CRISP-DM methodology

The general term for all concepts developed and defined in CRISP-DM.

Data mining context

Set of constraints and assumptions such as problem type, techniques or tools, application domain.

Data mining problem type

Class of typical data mining problems such as data description and summarization, segmentation,

concept descriptions, classification, prediction, dependency analysis.

Generic

A task which holds across all possible data mining projects, as complete, i.e., cover both the whole data mining process and all possible data mining applications and stable, i.e., valid for yet unforeseen developments like new modeling techniques, as possible.

Model

Ability to apply to a dataset to predict a target attribute, executable.

Output

Tangible result of performing a task.

Phase

High-level term for part of the process model, consists of related tasks.

Process instance

A specific project described in terms of the process model.

Process model

Defines the structure of data mining projects and provides guidance for their execution, consists of reference model and user guide.

Reference model

Decomposition of data mining projects into phases, tasks and outputs.

Specialized

A task that makes specific assumptions in specific data mining contexts.

Task

Part of a phase, series of activities to produce one or more outputs.

User guide

Specific advice on how to perform data mining projects.

2 Data mining problem types

Usually, the data mining project involves a combination of different problem types, which together solve the business problem.

2.1 Data description and summarization

Data Description and Summarization aims at the concise description of characteristics of the data, typically in elementary and aggregated form. This gives the user an overview of the structure of the data. Sometimes, data description and summarization alone can be an objective of a data mining project. For instance, a retailer might be interested in the turnover of all outlets broken down by categories. Changes and differences to a previous period could be summarized and highlighted. This kind of problem would be at the lower end of the scale of data mining problems.

However, in almost all data mining projects data description and summarization is a sub goal in the process, typically in early stages. At the beginning of a data mining process, the user often knows neither the precise goal of the analysis nor the precise nature of the data. Initial exploratory data analysis can help to understand the nature of the data and to find potential hypotheses for hidden information. Simple descriptive statistical and visualization techniques provide first insights in the data. For example, the distribution of customer age and their living areas gives hints about which parts of a customer group need to be addressed by further marketing strategies.

Data description and summarization typically occurs in combination with other data mining problem

types. For instance, data description may lead to the postulation of interesting segments in the data. Once segments are identified and defined a description and summarization of these segments is useful. It is advisable to carry out data description and summarization before any other data mining problem type is addressed. In this document, this is reflected by the fact that data description and summarization is a task in the data understanding phase.

Summarization also plays an important role in the presentation of final results. The outcomes of the other data mining problem types (e.g., concept descriptions or prediction models) may also be considered summarizations of data, but on a higher conceptual level.

Many reporting systems, statistical packages, OLAP and EIS systems can cover data description and summarization but do usually not provide any methods to perform more advanced modeling. If data description and summarization is considered a stand alone problem type and no further modeling is required, these tools are also appropriate to carry out data mining engagements.

2.2 Segmentation

The data mining problem type *segmentation* aims at the separation of the data into interesting and meaningful subgroups or classes. All members of a subgroup share common characteristics. For instance, in shopping basket analysis one could define segments of baskets depending on the items they contain.

Segmentation can be performed manually or (semi-) automatically. The analyst can hypothesize certain subgroups as relevant for the business question based on prior knowledge or based on the outcome of data description and summarization. However, there are also automatic clustering techniques that can detect previously unsuspected and hidden structures in data that allow segmentation.

Segmentation can be a data mining problem type of its own. Then the detection of segments would be the main purpose of data mining. For example, all addresses in zip code areas with higher than average age and income might be selected for mailing advertisements on home nursing insurance.

Often, however, very often segmentation is a step towards solving other problem types. Then, the purpose can be to keep the size of the data manageable or to find homogeneous data subsets that are easier to analyze. Typically, in large datasets various influences overlay each other and obscure the interesting patterns. Then, appropriate segmentation makes the task easier. For instance, analyzing dependencies between items in millions of shopping baskets is very hard. It is much easier (and more meaningful, typically) to identify dependencies in interesting segments of shopping baskets, for instance highvalue baskets, baskets containing convenience goods or baskets from a particular day or time.

Note: In the literature there is a confusion of terms. Segmentation is sometimes called clustering or classification. The latter term is confusing because some people use it to refer to the creation of classes, while others mean the creation of models to predict known classes for previously unseen cases. In this document, we restrict the term classification to the latter meaning (see below) and use the term segmentation for the former meaning, though classification techniques can be used to elicit descriptions of the segments discovered.

Appropriate techniques:

① Clustering techniques.

② Neural nets.

③ Visualization.

Example:

A car company regularly collects information about its customers concerning their socio-economic characteristics like income, age, sex, profession, etc. Using cluster analysis, the company can divide its customers into more understandable subgroups and analyze the structure of each subgroup. Specific marketing strategies are deployed for each group separately.

2.3 Concept descriptions

Concept description aims at an *understandable* description of concepts or classes. The purpose is not to develop complete models with high prediction accuracy, but to gain insights. For instance, a company may be interested to learn more about their loyal and disloyal customers. From a concept description of these concepts (loyal and disloyal customers) the company might infer what could be done to keep customers loyal or to transform disloyal customers to loyal customers.

Concept description has a close connection to both segmentation and classification. Segmentation may lead to an enumeration of objects belonging to a concept or class without any understandable description. Typically, there is segmentation before concept description is performed. Some techniques, for example conceptual clustering techniques, perform segmentation and concept description at the same time.

Concept descriptions can also be used for classification purposes. On the other hand, some classification techniques produce understandable classification models, which can then be considered as concept descriptions. The important distinction is that classification aims to be complete in some sense. The classification model needs to apply to *all* cases in the selected population. On the other hand, concept descriptions need not be complete. It is sufficient if they describe important parts of the concepts or classes. In the example above, it may be sufficient to get concept descriptions of those customers who are clearly loyal.

Appropriate techniques:

① Rule induction methods.

② Conceptual clustering.

Example:

Using data about the buyers of new cars and using a rule induction technique, a car company could generate rules that describe its loyal and disloyal customers. Below are examples of the generated rules:

If SEX = male and AGE > 51 then CUSTOMER = loyal

If SEX = female and AGE > 21 then CUSTOMER = loyal

If PROFESSION = manager and AGE < 51 then CUSTOMER = disloyal

If FAMILY STATUS = bachelor and AGE < 51 then CUSTOMER = disloyal

2.4 Classification

Classification assumes that there is a set of objects – characterized by some attributes or features – which belong to different classes. The class label is a discrete (symbolic) value and is known for each object. The objective is to build classification models (sometimes called classifiers), which assign the correct class label to previously unseen and unlabeled objects. Classification models are mostly used for predictive modeling.

The class labels can be given in advance, for instance defined by the user or derived from segmentation. Classification is one of the most important data mining problem types that occurs in a wide range of various applications. Many data mining problems can be transformed to classification problems. For example, credit scoring tries to assess the credit risk of a new customer. This can be transformed to a classification problem by creating two classes, good and bad customers. A classification model can be generated from existing customer data and their credit behavior. This classification model can then be used to assign a new potential customer to one of the two classes and hence accept or reject him.

Classification has connections to almost all other problem types. Prediction problems can be transformed to classification problems by binning continuous class labels, since binning techniques allow transforming continuous ranges into discrete intervals. These discrete intervals are then used as class labels rather than the exact numerical values and hence lead to a classification problem.

Some classification techniques produce understandable class or concept descriptions. There is also a connection to dependency analysis because classification models typically exploit and elucidate dependencies between attributes.

Segmentation can either provide the class labels or restrict the dataset such that good classification models can be built.

It is useful to analyze deviations before a classification model is built. Deviations and outliers can obscure the patterns that would allow a good classification model. On the other hand, a classification model can also be used to identify deviations and other problems with the data.

Appropriate techniques:

① Discriminant analysis.

② Rule induction methods.

③ Decision tree learning.

④ Neural nets.

⑤ K Nearest Neighbor.

⑥ Case-based reasoning.

⑦ Genetic algorithms.

Example:

Banks generally have information on the payment behavior of their credit applicants. Combining this financial information with other information about the customers like sex, age, income, etc., it is possible to develop a system to classify new customers as good or bad customers, (i.e., the credit risk in acceptance of a customer is either low or high, respectively).

2.5 Prediction

Another important problem type that occurs in a wide range of applications is *prediction*. Prediction is very similar to classification. The only difference is that in prediction the target attribute (class) is not a qualitative discrete attribute but a continuous one. The aim of prediction is to find the numerical value of the target attribute for unseen objects. In the literature, this problem type is sometimes called regression. If prediction deals with time series data then it is often called forecasting.

Appropriate techniques:

① Regression analysis.

② Regression trees.

③ Neural nets.

④ K Nearest Neighbor.

⑤ Box-Jenkins methods.

⑥ Genetic algorithms.

Example:

The annual revenue of an international company is correlated with other attributes like advertisement, exchange rate, inflation rate etc. Having these values (or their reliable estimations for the next year) the company can predict its expected revenue for the next year.

2.6 Dependency analysis

Dependency analysis consists of finding a model that describes significant dependencies (or associations) between data items or events. Dependencies can be used to predict the value of a data

item given information on other data items. Although dependencies can be used for predictive modeling, they are mostly used for understanding. Dependencies can be strict or probabilistic.

Associations are a special case of dependencies, which have recently become very popular. Associations describe affinities of data items (i.e., data items or events which frequently occur together). A typical application scenario for associations is the analysis of shopping baskets. There, a rule like "in 30 percent of all purchases, beer and peanuts have been bought together" is a typical example for an association. Algorithms for detecting associations are very fast and produce many associations.

Selecting the most interesting ones is a challenge.

Dependency analysis has close connections to prediction and classification, where dependencies are implicitly used for the formulation of predictive models. There is also a connection to concept descriptions, which often highlight dependencies.

In applications, dependency analysis often co-occurs with segmentation. In large datasets, dependencies are seldom significant because many influences overlay each other. In such cases it is advisable to perform a dependency analysis on more homogeneous segments of the data.

Sequential patterns are a special kind of dependencies where the order of events are considered. In the shopping basket domain, associations describe dependencies between items at a given time. Sequential patterns describe shopping patterns of one particular customer or a group of customers over time.

Appropriate Techniques:

① Correlation analysis.

② Regression analysis.

③ Association rules.

④ Bayesian networks.

⑤ Inductive Logic Programming.

⑥ Visualization techniques.

Example 1:

Using regression analysis, a business analyst has found that there is a significant dependency between the total sales of a product and its price and the amount of the total expenditures for the advertisement. Once the analyst discovered this knowledge, he can reach the desired level of the sales by changing the price and/or the advertisement expenditure accordingly.

Example 2:

Applying association rule algorithms to data about car accessories, a car company has found that if a radio is ordered, an automatic gearbox is ordered as well in 95 percent of all cases. Based on this dependency, the car company decides to offer these accessories as a combination which leads to cost reduction.

To get more information, call your nearest SPSS office or visit our World Wide Web site at

www.spss.com

SPSS Inc. +1.312.651.3000

Toll-free +1.800.543.2185

SPSS Argentina +5411.4814.5030

SPSS Asia Pacific +65.245.9110

SPSS Australasia +61.2.9954.5660

Toll-free +1.800.024.836

SPSS Belgium +32.16.317070

SPSS Benelux +31.183.651.777

SPSS Brasil +55.11.5505.3644

SPSS Czech Republic +420.2.24813839

SPSS Danmark +45.45.46.02.00

SPSS East Africa +254.2.577.262

SPSS Federal Systems (U.S.) +1.703.527.6777

Toll-free +1.800.860.5762

SPSS Finland +358.9.4355.920

SPSS France +01.55.35.27.00

SPSS Germany +49.89.4890740

SPSS Hellas +30.1.72.51.925

SPSS Hispanoportuguesa +34.91.447.37.00

SPSS Hong Kong +852.2.811.9662

SPSS Ireland +353.1.415.0234

SPSS Israel +972.9.9526700

SPSS Italia +39.051.252573

SPSS Japan +81.3.5466.5511

SPSS Korea +82.2.3446.7651

SPSS Latin America +1.312.651.3539

SPSS Malaysia +603.7873.6477

SPSS Mexico +52.5.682.87.68

SPSS Miami +1.305.627.5700

SPSS Norway +47.22.40.20.60

SPSS Polska +48.12.6369680

SPSS Russia +7.095.125.0069

SPSS Schweiz +41.1.266.90.30

SPSS Singapore +65.324.5150

SPSS South Africa +27.11.807.3189

SPSS South Asia +91.80.2088069

SPSS Sweden +46.8.506.105.50

SPSS Taiwan +886.2.25771100

SPSS Thailand +66.2.260.7070

SPSS UK +44.1483.719200

SPSS Inc. enables organizations to develop more profitable customer relationships by providing analytical solutions that discover what customers want and predict what they will do. The company delivers analytical solutions at the intersection of customer relationship management and business intelligence. SPSS analytical solutions integrate and analyze market, customer and operational data and deliver results in key vertical markets worldwide including: telecommunications, health care, banking, finance, insurance, manufacturing, retail, consumer packaged goods, market research and the public sector. For more information, visit **www.spss.com.**