普通高等教育“十一五”国家级规划教材

教育部“高等学校教学质量与教学改革工程”立项项目

张兴会 等编著

数据仓库与数据挖掘工程实例

计算机科学与技术专业实践系列教材

清华大学出版社

普通高等教育“十一五”国家级规划教材

计算机科学与技术专业实践系列教材

教育部“高等学校教学质量与教学改革工程”立项项目

数据仓库与数据挖掘工程实例

张兴会 等编著

清华大学出版社

内 容 简 介

数据仓库与数据挖掘是与计算机、信息类等相关专业的核心课程。本书采用提出问题、分析问题、解决问题的思路，通过工程实例介绍了 SQL Server 2005 和 Weka 软件的使用方法以及联机分析处理技术、关联规则方法、决策树方法、贝叶斯方法、人工神经网络方法、聚类分析方法、线性回归方法等数据仓库与数据挖掘技术。

本书结构严谨，条理清晰，语言浅显易懂，循序渐进地表达了知识内容；坚持理论与实际相结合，知识理论与具体实现方法相结合，使技术实现具体化、生动化、可操作化；工程实例的实现过程建立在 SQL Server 2005 和 Weka 软件的基础上，以帮助读者在学习后达到学以致用的效果。本书可以和《数据仓库与数据挖掘技术》教材配合使用，旨在帮助读者在学习数据仓库与数据挖掘理论知识的基础上，通过学习工程实例分析，较好地掌握数据挖掘与数据仓库构建模型的操作过程，进一步提高对信息管理和利用能力。

本书可以作为计算机、信息类等专业本科生数据挖掘课程的教材，也可以作为其他专业技术人员的自学参考书。

图书在版编目(CIP)数据

数据仓库与数据挖掘工程实例/张兴会等编著. —北京：清华大学出版社，2014
计算机科学与技术专业实践系列教材
ISBN 978-7-302-35541-0

Ⅰ. ①数… Ⅱ. ①张… Ⅲ. ①数据库系统－高等学校－教材 ②数据采集－高等学校－教材
Ⅳ. ①TP311.13 ②TP27

中国版本图书馆 CIP 数据核字(2014)第 034937 号

责任编辑：汪汉友　赵晓宁
封面设计：傅瑞学
责任校对：李建庄
责任印制：沈　露

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　邮　　编：100084
社 总 机：010-62770175　　邮　　购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
课件下载：http://www.tup.com.cn，010-62795954
印 装 者：北京国马印刷厂
经　　销：全国新华书店
开　　本：185mm×260mm　　印　　张：8.5　　字　　数：212 千字
版　　次：2014 年 9 月第 1 版　　印　　次：2014 年 9 月第 1 次印刷
印　　数：1～1500
定　　价：23.00 元

产品编号：034552-01

前　言

数据挖掘技术在科学研究和日常生活中具有广泛的应用，被列为21世纪最具潜力的应用技术之一。现在数据挖掘技术已经成为信息系统、应用数学等专业学生的必修教学内容。

为此，本书在编写时力求突出以下特色：

(1) 引入数据挖掘研究的热点问题以及最新研究成果，保证教材的先进性。

(2) 强化目标驱动观点，使读者学习有的放矢。

(3) 每章后面都详细讲解了在SQL Server 2005或Weka环境下相关理论的具体实现技术，使得读者可以理论联系实际，培养解决实际问题的能力。

(4) 在文字表达方面争取语言更通俗、易懂、易读。

本书具体内容如下：

实例1～实例10，详细介绍了基于联机分析处理技术、关联规则方法、决策树方法、贝叶斯方法、人工神经网络方法、聚类分析方法、线性回归方法等方法的10个工程实例的具体实现。

附录A和附录B，分别介绍了SQL Server 2005和Weka软件的任务描述和实现方法。

本书的案例来源于不同的专业领域和最新的工程实践，新颖独特，具有代表性和很强的实际借鉴价值。读者通过学习，可以了解和掌握数据挖掘技术的理论和算法，熟悉在各个领域应用的流程和分析方法，从而为以后的数据分析工作夯实基础。

为了能更好地将工程实例与相关理论知识相结合，将基本概念与具体的方法、工具相结合，达到学以致用的效果，读者可参考笔者所编著的《数据仓库与数据挖掘技术》进行学习。

本书由张兴会统稿，王明春、郑晓艳、刘玲、刘新钰参加了本书的编写、图表绘制、模型构建、软件调试等工作。在本书编写过程中，安淑芝教授提出了宝贵的修改意见。另外，本书还参阅和引用了许多专家和学者的文献资料，在此表示衷心的感谢。

由于笔者水平和能力有限，新技术的发展和更新较快，书中难免有不妥之处，欢迎读者批评指正。笔者邮箱为xhzhang@tute.edu.cn。

编　者

2014年8月

目　录

实例 1　基于联机分析处理技术的税务审计分析 …… 1
1.1　任务描述 …… 1
1.2　技术原理 …… 1
1.2.1　联机分析处理的定义 …… 1
1.2.2　联机分析处理的一些具体操作 …… 1
1.3　具体实现 …… 4
1.3.1　建立数据库 …… 4
1.3.2　新建数据源 …… 10
1.3.3　新建数据源视图 …… 15
1.3.4　浏览数据 …… 17
1.3.5　数据分析 …… 20
1.4　案例总结 …… 23

实例 2　基于关联规则方法的网上交易服务质量评价分析 …… 24
2.1　任务描述 …… 24
2.2　技术原理 …… 25
2.2.1　关联规则的概念 …… 25
2.2.2　Apriori 算法 …… 25
2.3　具体实现 …… 26
2.4　案例小结 …… 32

实例 3　基于 Weka KnowledgFlow 模块的大学生专业方向预测分析 …… 33
3.1　任务描述 …… 33
3.2　技术原理 …… 33
3.2.1　数据收集和准备 …… 33
3.2.2　模型选择 …… 33
3.3　具体实现 …… 33
3.3.1　数据预处理 …… 33
3.3.2　建立和使用知识流 …… 35
3.4　案例小结 …… 39

实例 4　基于决策树方法的网球运动天气状况评价分析 …… 41
4.1　任务描述 …… 41
4.2　技术原理 …… 41
4.2.1　决策树的概念 …… 41
4.2.2　信息论的基本概念 …… 42

4.2.3 ID3 建树算法 …… 42
4.3 具体实现 …… 42
4.4 案例小结 …… 48
实例 5 基于 Weka Experimenter 模块的人力资源管理挖掘模型选择分析 …… 49
5.1 任务描述 …… 49
5.2 技术原理 …… 49
5.2.1 挖掘类型确定 …… 49
5.2.2 数据收集和准备 …… 49
5.3 具体实现 …… 50
5.3.1 数据预处理 …… 50
5.3.2 模型比较和选择 …… 51
5.4 案例小结 …… 55
实例 6 基于贝叶斯方法的证券客户流失预警分析 …… 56
6.1 任务描述 …… 56
6.2 技术原理 …… 57
6.2.1 朴素贝叶斯分类算法 …… 57
6.2.2 朴素贝叶斯分类举例 …… 58
6.3 具体实现 …… 59
6.4 案例小结 …… 63
实例 7 基于人工神经网络方法的信贷数据分析 …… 64
7.1 任务描述 …… 64
7.2 技术原理 …… 64
7.2.1 BP 神经网络结构 …… 64
7.2.2 BP 神经网络学习算法 …… 65
7.3 具体实现 …… 67
7.3.1 数据准备 …… 67
7.3.2 挖掘流程 …… 70
7.4 案例小结 …… 78
实例 8 基于 K-means 方法的栀子花聚类分析 …… 79
8.1 任务描述 …… 79
8.2 技术原理 …… 79
8.3 具体实现 …… 80
8.4 案例小结 …… 87
实例 9 基于线性回归方法的汽车油耗预测分析 …… 88
9.1 任务描述 …… 88
9.2 技术原理 …… 88
9.3 具体实现 …… 89

9.4 案例小结 …… 95
实例 10 基于决策树方法的中文文本自动分类分析 …… 96
10.1 任务描述 …… 96
10.2 技术原理 …… 96
10.2.1 文本挖掘的概念 …… 96
10.2.2 文本分词技术 …… 96
10.2.3 文本特征表示 …… 97
10.3 具体实现 …… 97
10.4 案例小结 …… 105
附录 A SQL Server 2005 的安装 …… 106
A1 任务描述 …… 106
A2 具体实现 …… 106
附录 B Weka 软件的安装和数据转换 …… 114
B1 任务描述 …… 114
B2 具体实现 …… 114
参考文献 …… 128

实例1　基于联机分析处理技术的税务审计分析

1.1　任务描述

需要对某市国税局的延期纳税审批情况进行审计，资料来源于某市国税局延期纳税数据库，此数据库中共有三个数据表。

(1) 延期纳税批件表：在此表中共有1568条记录，记录着税务局批准企业纳税的基本信息，例如征收项目种类、税款所属期、税额、纳税人名称等。

(2) 税务机关代码表：记录该市所属各区县的税务机关代码及名称。

(3) 征收项目代码表：记录各征收项目税种的代码及名称。

在审计时，面临诸多的困难，如时间跨度大(从2002年1月至2004年2月)、所属区县多、审批金额大等。对税务局审批延期纳税的合法合规性分析离不开对纳税企业的延伸。如何在这些浩如烟海的电子资料中找到需要的信息是这次审计的核心问题。

本例将通过分析数据库中国税局给企业批准延期纳税的大量数据，简要介绍审计过程中如何应用多维数据分析工具在统揽全局、把握总体的基础上对大量的电子数据进行筛选、分析，快速找出审计重点，准确定位延伸分析的对象。通过对某市国税局的深入了解，在审计时需要掌握以下情况：2002—2004年全市共审核批准了多少延期纳税税款？哪年审批的金额比较大？审计的都有什么税种？各税种占的比例有多大？各个区县分别审批了多少税款？哪个区县审批的金额较多？审批时间集中在什么时候？审计人员应从哪里进行突破？

在这个案例中，通过建立多维数据集，在把握总体、统揽全局的基础上观察趋势，选择重点；最后进行有针对性的延伸取证，顺利完成了整个审计过程。

1.2　技术原理

1.2.1　联机分析处理的定义

联机分析处理委员会对联机分析处理(OLAP)的定义为：使分析、管理或执行人员能够从多种角度对从原始数据中转化出来、能够真正为用户所理解、并真实反映企业维特性的信息进行快速、一致、交互地存取，从而获得对数据更深入了解的一类软件技术。

OLAP的基本多维分析操作有钻取(Drill-up和Drill-down)、切片(Slice)和切块(Dice)以及旋转(Pivot)等。

1.2.2　联机分析处理的一些具体操作

1. 钻取

钻取是改变维的层次，变换分析的粒度。它包括向下钻取(Drill-down)和向上钻取(Drill-up)。向上钻取也称为上卷(Roll-up)，是在某一维上将低层次的细节数据概括到高层次的汇总数据，或减少维数；而Drill-down则相反，它从汇总数据深入到细节数据进行观察

或增加新维。例如，图 1-1 所示的数据立方体经过沿着分行维的概念层次上卷，由分行上升到城市，得到如图 1-2 所示的立方体；图 1-1 中的数据立方体经过沿时间维下钻，由年度下降到季度，得到如图 1-3 所示的数据立方体。

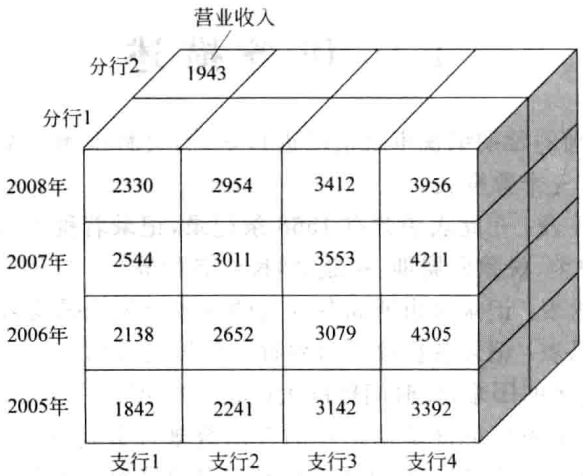

图 1-1　数据立方体示例

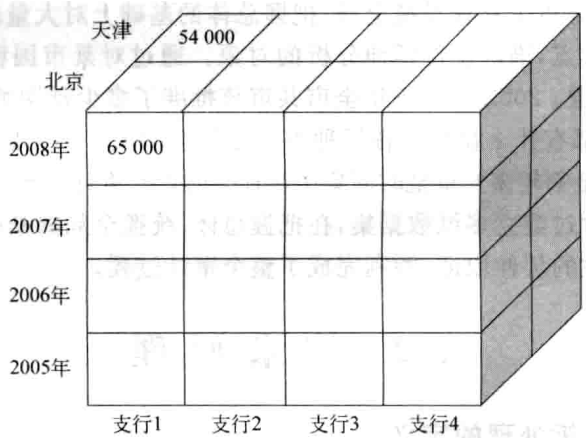

图 1-2　向上钻取后得到的数据立方体

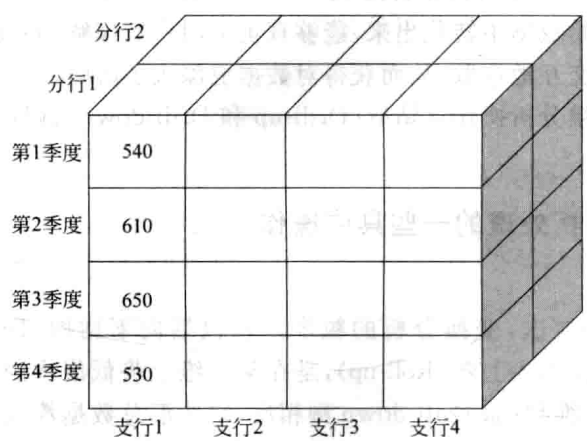

图 1-3　向下钻取后得到的数据立方体

2. 切片和切块

切片：在给定数据立方体的一个维上进行选择操作就是切片，切片的结果是得到一个二维平面数据。例如，对图 1-1 中数据立方体，使用条件：

```
"银行分行='分行 1'"
```

进行选择，就相当于在原来的立方体中切出一片，结果如图 1-4 所示。

切块：在给定数据立方体的两个或多个维上进行选择操作就是切块，切块的结果得到一个子立方体。例如，对图 1-1 所示数据立方体，使用条件：

```
(银行分行="分行 1"or"分行 2")
And (时间="2007 年"or"2008 年")
And (银行支行="支行 1"or"支行 2")
```

进行选择，就相当于在原立方体中切出一小块，结果如图 1-5 所示。

	支行1	支行2	支行3	支行4
2008年	2330	2954	3412	3956
2007年	2544	3011	3553	4211
2006年	2138	2652	3079	4305
2005年	1842	2241	3142	3392

图 1-4　切片示例

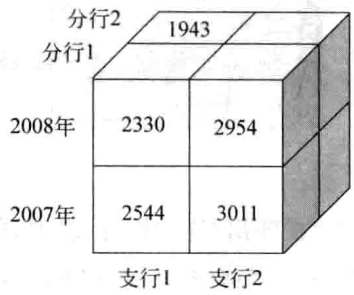

图 1-5　切块示例

3. 旋转

旋转是变换维的方向，即在表格中重新安排维的放置（如行列互换）。图 1-6 所示是图 1-1 中立方体通过旋转横纵坐标所得的立方体。

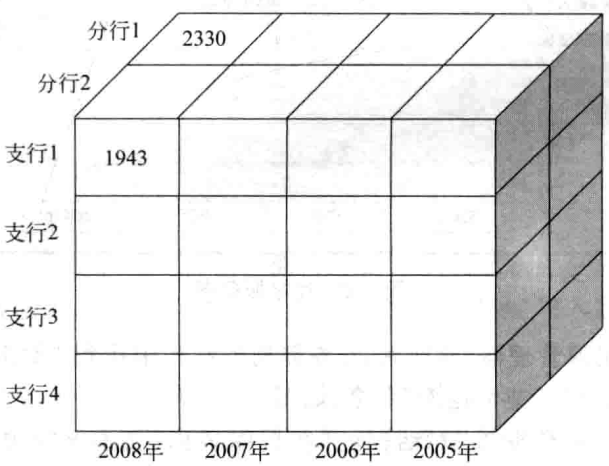

图 1-6　旋转后得到的数据立方体

1.3 具体实现

1.3.1 建立数据库

建立数据库的步骤如下：

(1) 依次执行“开始”→“程序”→Microsoft SQL Server 2005→SQL Server Management Studio 命令，如图 1-7 所示，打开 SQL Server 2005 数据库管理器。

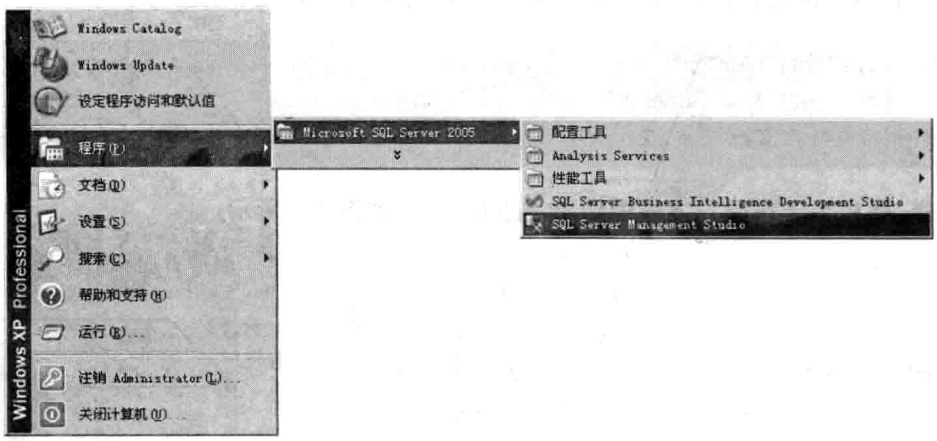

图 1-7 打开数据库管理器

(2) 在弹出“连接服务器”对话框中选择安装 SQL Server 2005 时所建立的命名实例名，在身份验证中选择“Windows 身份验证”项，单击“连接”按钮，如图 1-8 所示。

图 1-8 连接服务器

(3) 进入“对象资源管理器”界面后，在左侧树形结构中找到“数据库”文件夹，右击，在弹出的快捷菜单中选择“新建数据库”命令，如图 1-9 所示。

(4) 在弹出的“新建数据库”对话框的“数据库名称”文本框中填写“延期纳税”，单击“确定”按钮，如图 1-10 所示。

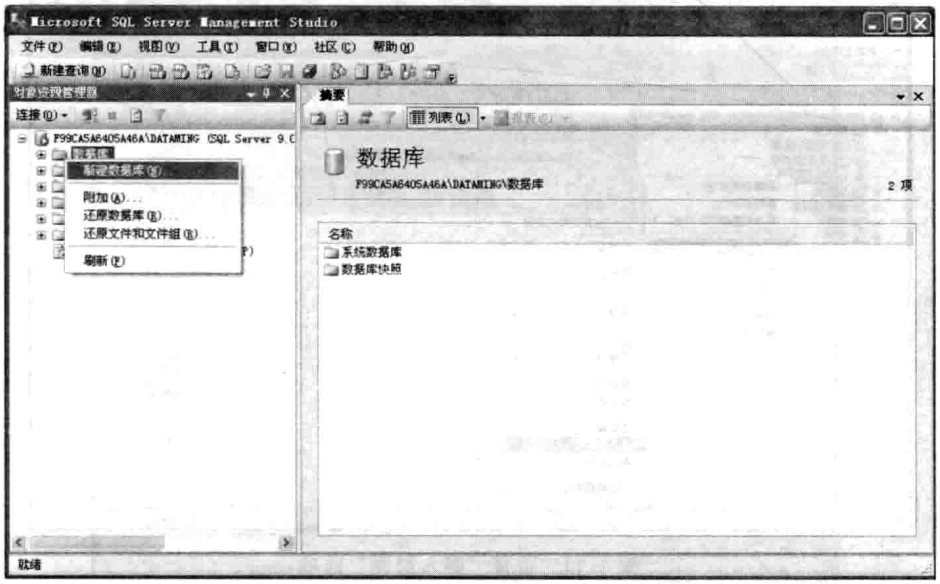

图 1-9　进入对象资源管理器

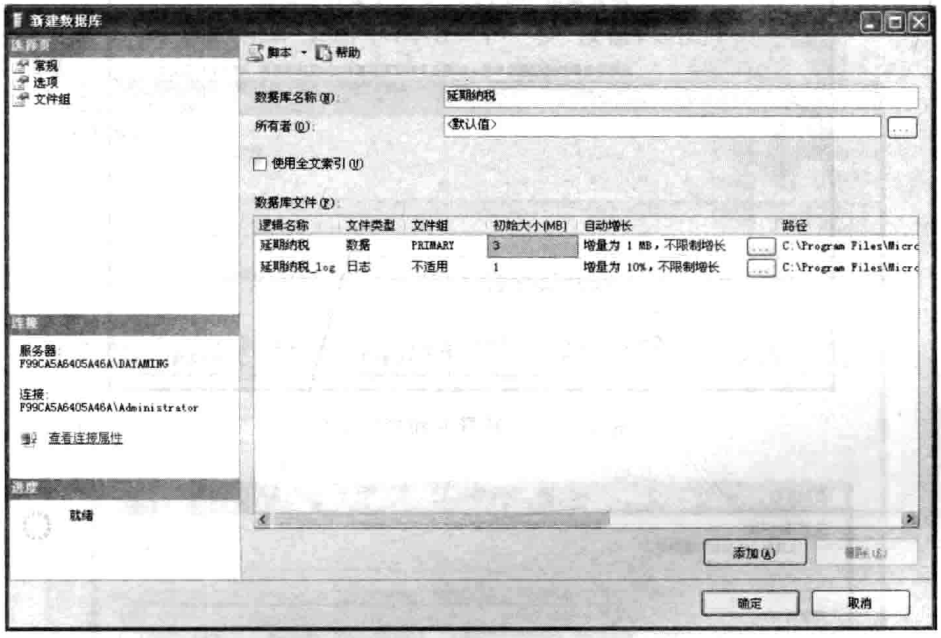

图 1-10　新建数据库

(5) 回到“对象资源管理器”界面，在左侧树形结构中找到新建立的数据库“延期纳税”项，右击“延期纳税”数据库，在弹出的快捷菜单中选择“任务”→“导入数据”命令，如图 1-11 所示。

(6) 打开“SQL Server 导入和导出向导”对话框，如图 1-12 所示。

(7) 单击“下一步”按钮。在“数据源”下拉列表中选择 Microsoft Access 项，如图 1-13 所示。

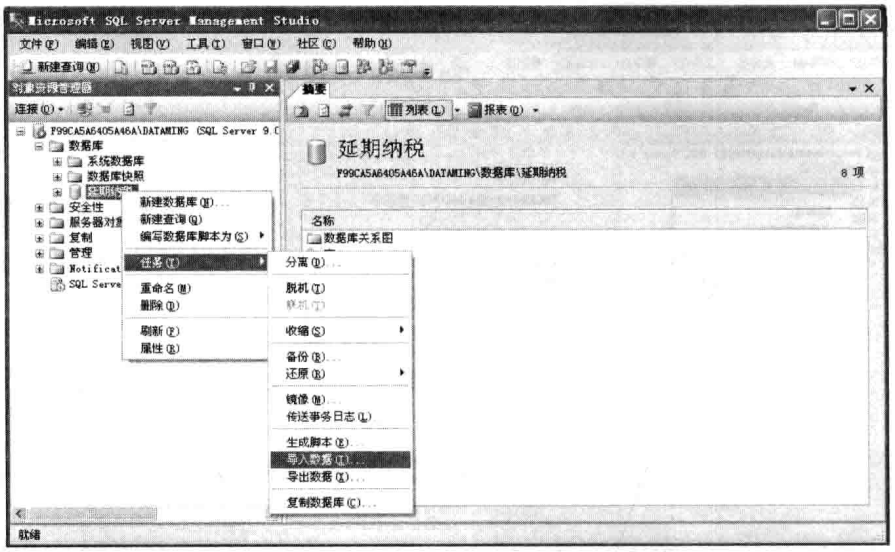

图 1-11　选择“导入数据”选项

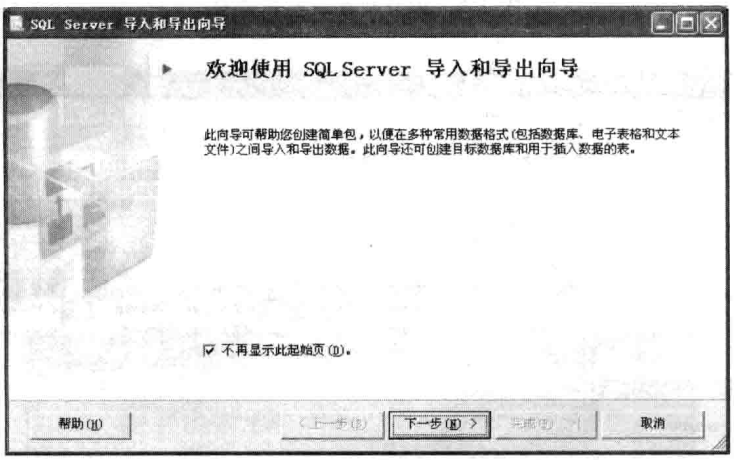

图 1-12　打开导入和导出向导

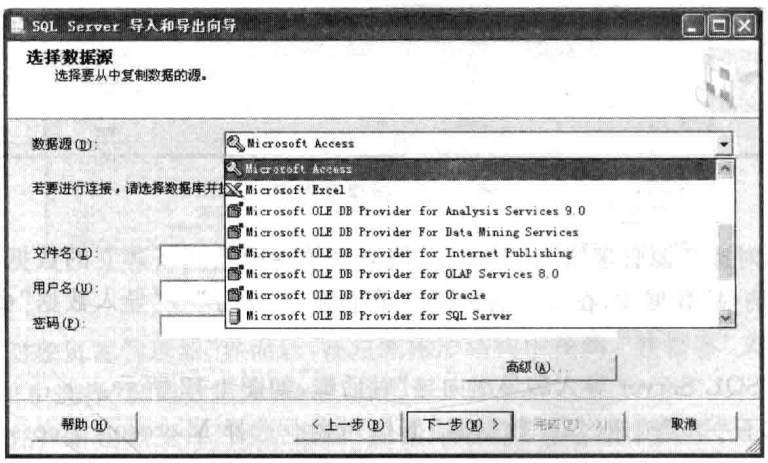

图 1-13　选择数据源类型

(8) 单击“下一步”按钮。选择需要导入的数据，单击“打开”按钮，如图 1-14 所示。

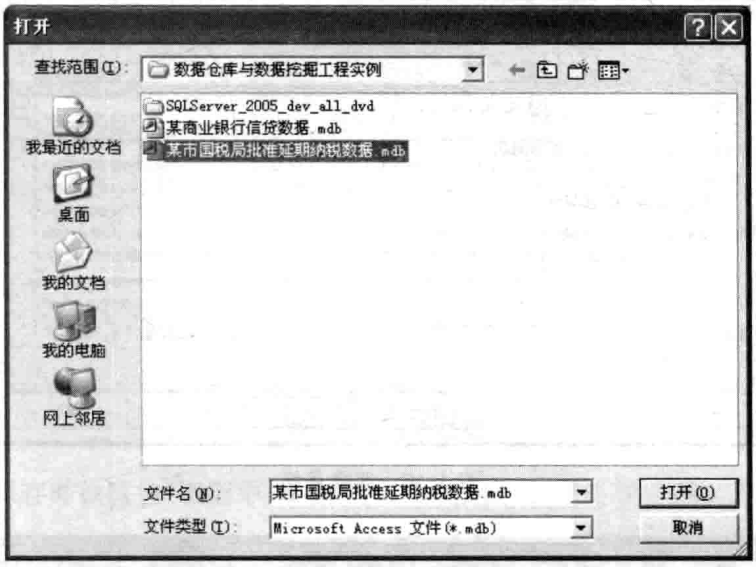

图 1-14　选择导入数据

(9) 在弹出的“选择数据源”页面中，单击“下一步”按钮，如图 1-15 所示。

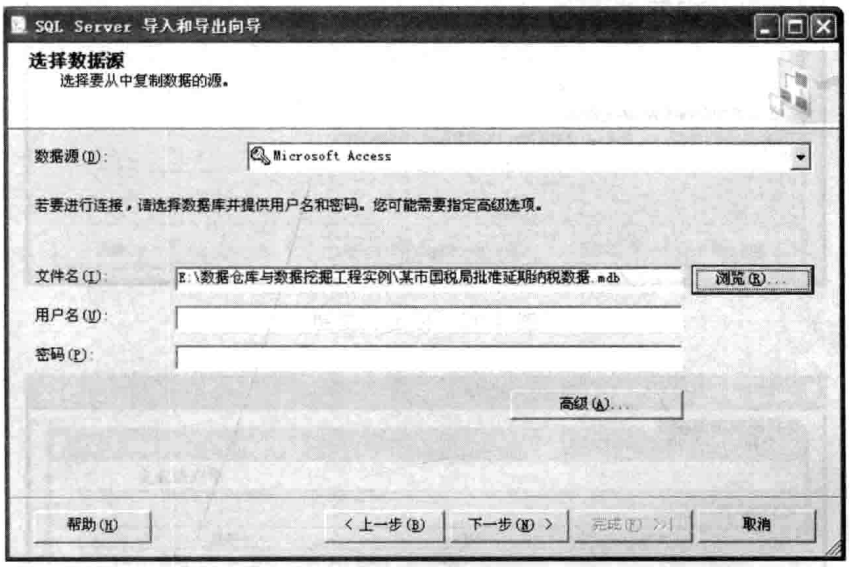

图 1-15　确定导入数据

(10) 在弹出的“选择目标”页面中，单击“下一步”按钮，如图 1-16 所示。

(11) 在弹出的“指定表复制或查询”页面中选择“复制一个或多个表或视图的数据”单选按钮并单击“下一步”按钮，如图 1-17 所示。

(12) 在弹出的“选择源表和源视图”页面中，单击“全选”按钮，如图 1-18 所示。所有需要导入的数据表全部被选中，单击“下一步”按钮。

(13) 单击“预览”对导入数据进行预览，并单击“确定”按钮，如图 1-19 所示。

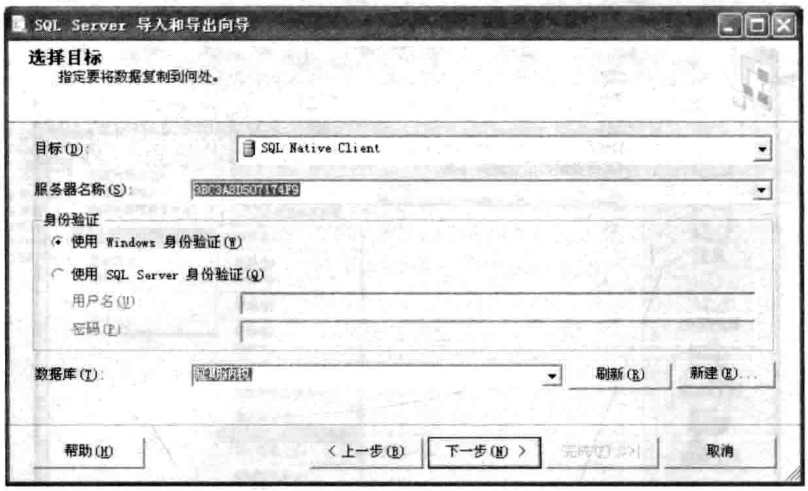

图 1-16　选择目标

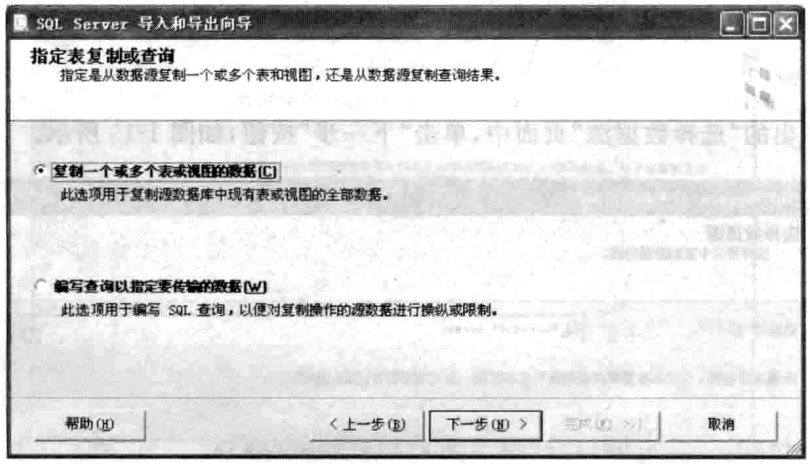

图 1-17　指定表复制

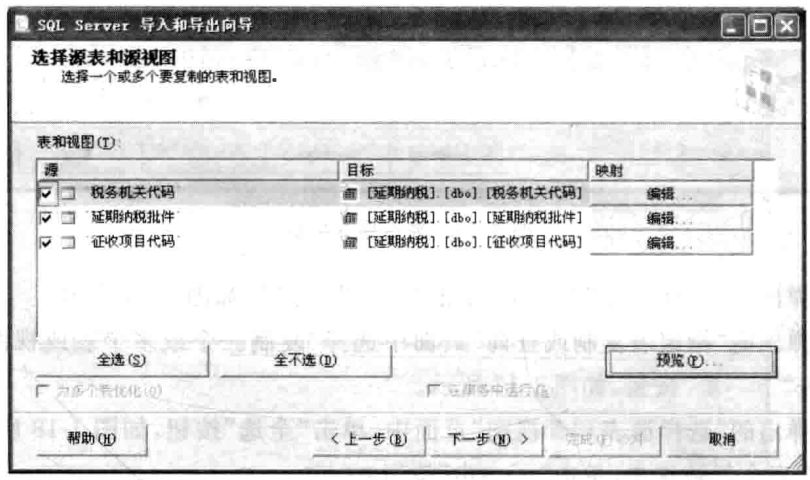

图 1-18　选择源表

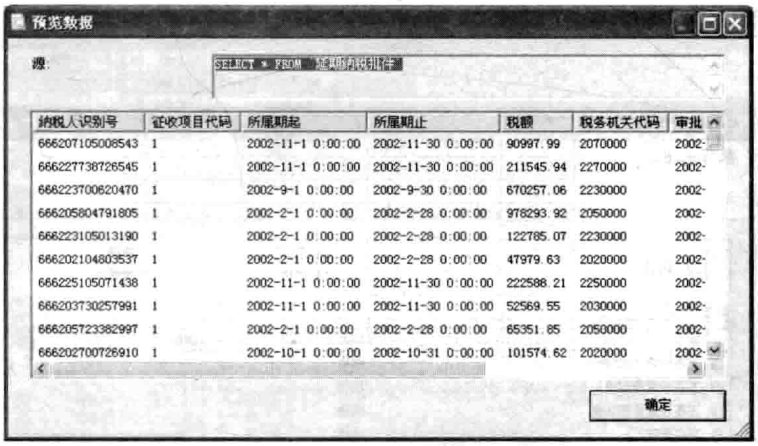

图 1-19　预览数据

(14) 在“保存并执行包”页面中,单击“下一步”按钮,如图 1-20 所示。

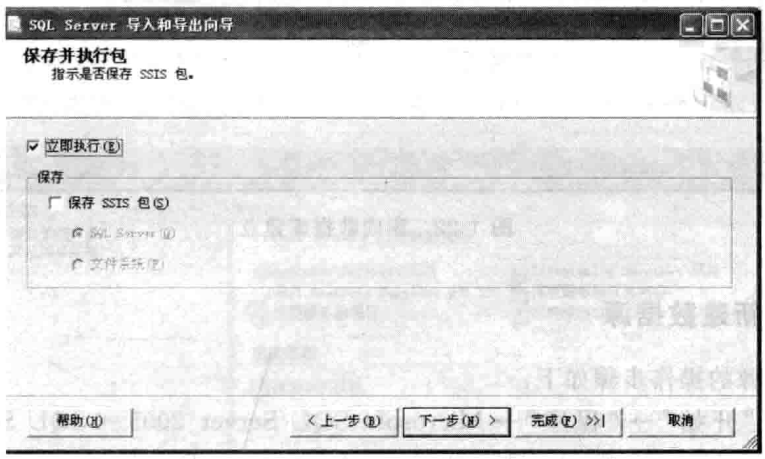

图 1-20　保存并执行包

(15) 在弹出的“完成该向导”页面中,单击“完成”按钮,如图 1-21 所示。

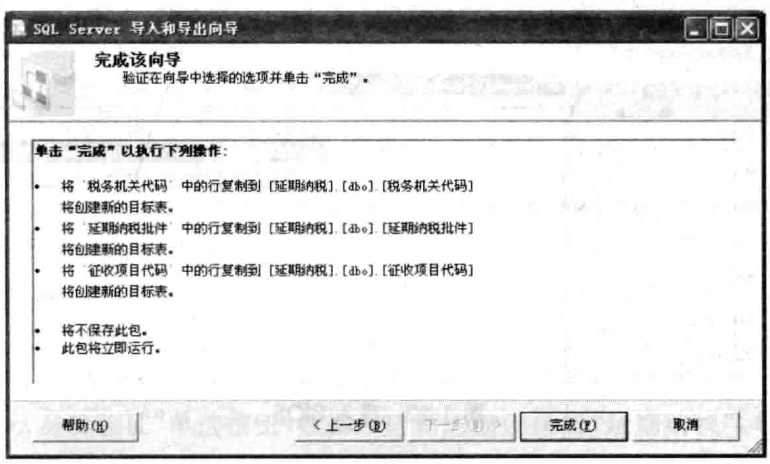

图 1-21　完成导入和导出向导

(16) 在弹出的“执行成功”对话框中，单击“关闭”按钮，完成数据库的建立，如图 1-22 所示。

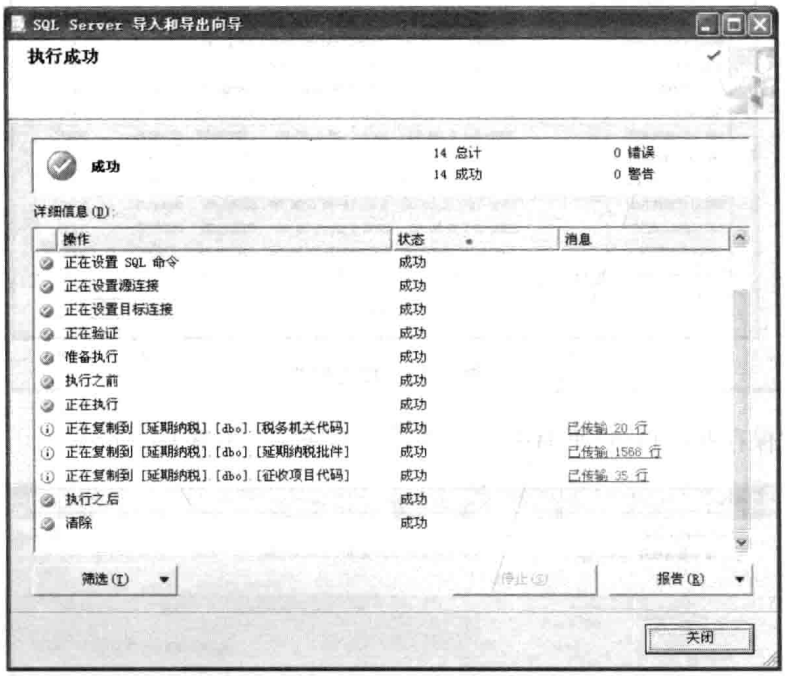

图 1-22 完成数据库建立

1.3.2 新建数据源

新建数据源的操作步骤如下：

(1) 选择“开始”→“程序”→Microsoft SQL Server 2005→SQL Server Business Intelligence Development Studio 进入 Business Intelligence Development Studio(BIDS)，如图 1-23 所示。

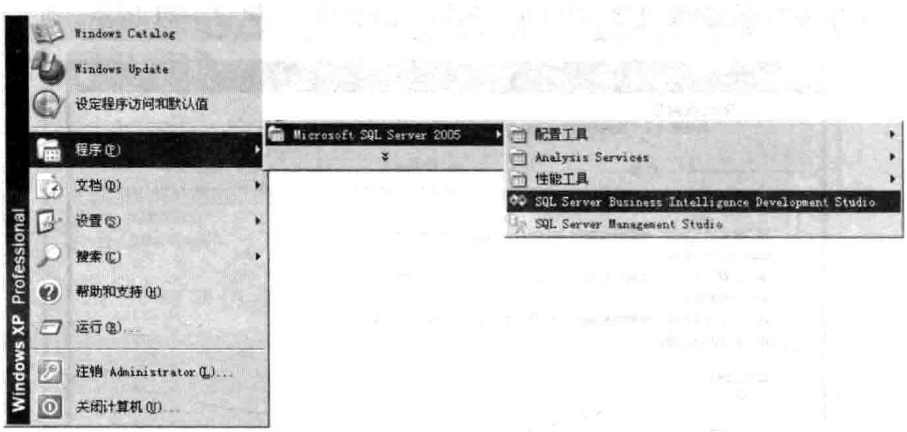

图 1-23 进入 BIOS

(2) 选择“文件”→“新建”→“项目”命令，如图 1-24 所示。