摘要

伴随着信息高速公路的建设，数字技术、数据库技术迅猛发展，人类的数据库里积累了越来越多的历史数据，而从这些海量的数据里探索出实用的有价值的信息对人类社会的发展有着重要的指导意义，这便形成了近几年学术研究的热点，应运而生的就是“数据挖掘”学科。简单的解释就是通过数据库、机器学习、人工智能、统计学等领域的技术，从数据库或web中提取出隐含的，有应用价值的知识和模式，为人们的决策提供有意义的支持和指导。

数据挖掘技术已经逐渐应用到了银行、证券公司以及零售行业的领域中，并且取得了不错的业绩，深受研究人员和商业组织的青睐。当前随着我国税收工作的不断完善，税控系统的应用将会越来越广泛。它的主要核心技术在于，通过嵌入在销售企业pos终端软件中，时时的采集企业的销售数据，并将企业完整的销售记录及时地储存起来，以便于税务机关随时进行核查并进行合理的收税，对消除企业的逃税、漏税起着积极的重要作用。

然而，现在销售行业的规模越来越大，企业的销售数据已越来越庞大，如何在这些海量的销售数据中挖掘出有意义的，对国家税收有帮助的，并对企业的经营策略有价值的知识便成为了现在一个重要的研究课题，而数据挖掘技术正是从这一点出发，利用它本身的各种挖掘技术，从中探索出那些鲜为人知的知识，从而有效地解决了以上问题。

本文主要从税收管理分析的角度来讨论数据挖掘技术。首先介绍了数据挖掘的概念和一些算法以及商业智能的应用，然后针对税控数据源进行分析处理：大量数据迁移、数据预处理，以及建立税源检测数据模型。接着重点研究了聚类算法。并对k-均值算法进行了有效的学习和改进，将其良好的整合到第三方开源挖掘工具—Weka。Weka的全名是怀卡托智能分析环境，已将大量的数据挖掘和机器学习算法嵌入其中，并且为我们提供了算法融入接口。最后通过对其进行算法的改进和界面的更新，达到用数据描述现状、预测趋势的目的，使困绕税务部门的零税申报、低税申报、虚假申报、发票违章等难题得到进一步解决。

关键词 数据挖掘；税收管理；算法；Weka

Abstract

With the construction of information highway, digital and database technology has been greatly developed, and our database has stored more and more historical data. How to explorer Valuable information from the mass data has an important guiding significance to the development of human society, which came into being the hot focus data mining, of academic research in recent years. The Simple explanation is that with the technology of database, machine learning artificial intelligence and so on, pick up the implied and valuable knowledge and pattern from database or web, so as to provide people with Strong support and guidance to make decision.

Data mining technology has been gradually applied to banks, securities companies, as well as the area of the retail industry and achieved good results, which fascinates the researchers and commercial organizations. At present, with the continuous improvement of our tax work, tax-control system will be used more and more widely. Its main core technology is to collect sales data from time to time through the monitoring software embedded in the terminal pos, and will complete the sale of corporate records stored in a timely manner, so that tax authorities verify at any time and tax reasonably. It plays an important role to the elimination of corporate tax evasion.

But now, the scale of sales industry becomes bigger and bigger, the sales quantity becomes larger and larger, and how to excavate significantly and instrumental to the state's tax valuable knowledge for the Business strategy has come into been an important research topic. Data mining based on the above requirement, use its own a variety of mining algorithms to probe those little-known knowledge, as a result effectively solving the above problems.

This paper aims to discuss data mining technology in the Analysis of tax administration aspect. Firstly, introduce the concept of data mining, algorithms and Business Intelligence Application. Secondly, do the analysis and process, for example, data migration, data reprocessing, and constitution of data check model and so on. Thirdly, focus on the clustering algorithm, especially improve on k-means algorithm then integrate it into a open-source mining tool, Weka, named Waikato Environment for Knowledge Analysis, which has a large number of data mining and machine learning algorithms embedded, and provide us with interface for algorithm access. At last, use the graphical interface to complete the purpose of describe present situation

and forecast the trend, in order to solve the tax department plagued conundrums of zero-tax reporting, low-tax reporting, false reporting and invoices anti-regulations.

Keywords Data Mining; tax management; algorithm; Weka

独 创 性 声 明

本人声明所呈交的论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得北京工业大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

签名：王[illegible] 日期：2009.6

关于论文使用授权的说明

本人完全了解北京工业大学有关保留、使用学位论文的规定，即：学校有权保留送交论文的复印件，允许论文被查阅和借阅；学校可以公布论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存论文。

（保密的论文在解密后应遵守此规定）

签名：王[illegible] 导师签名：[illegible] 日期：2009.6.3

第 1 章 绪论

1.1 研究背景与意义

在信息时代，数据与信息同时存在，相互依赖。随着时代的发展，我们身边充满了各式各样的数据，只有将这些杂乱无章的海量数据进行甄别、挑选、分析，转化为信息和知识，才能帮助我们做出明智的选择。随着从数据到智慧这种层次的出现，数据挖掘技术便应运而生。

在商业中，数据挖掘被定义为一种新的商业信息处理技术。其主要特点是对商业数据库中的大量业务数据进行抽取、转换、分析和其他模型化处理，从中提取辅助商业决策的关键性数据[12][16][26]。

当前，随着我国市场经济的不断完善和税收改革的不断深入，纳税根据和方式日益重要，并渐渐出现在我国的税收管理工作中。以销售额为根据纳税已经逐渐成为是现代税收的主要方式，它是一种通过利用税控系统等一系列信息化平台，对企业的各项销售数据和销售信息进行采集、加工和处理，从而对企业征收合理税费并对纳税人纳税情况进行评价的新的管理模式[9][27]。

通过对现有数据的分析和利用，可以为税务部门决策提供科学的参考依据，强化税收收入分析预测，规范数据源头，提高数据质量。由此税务部门可以更好的掌握企业的销售情况，推测企业有无逃税漏税等问题，并对将来税率的改革起到极其重要的指导作用[29][30]。

作为占市场主导的企业来讲，通过对销售数据的挖掘和处理，可以更准确地了解市场行情，市场的销售趋势，从而编制出一套有效地更适应市场发展的经营战略，更好的推动市场的扩大和企业的攀升[24]。

但是目前，我国的纳税分析系统还没有一套完整的可行方案，评估纳税人员只能依据一些常用的指标以及一些历史经验做出推算，缺乏科学依据，而且费时费力，不能进行全面有效的评估考核，得出来的结果，往往与实际差异甚大。另外，由于外界因素的影响，各种指标和数据维度的建立也会有波动变化，依靠人工维护非常困难，常常导致数据的分析结果严重偏离实际[14]。

基于此的数据挖掘技术恰恰能够很好的解决上述各种问题。利用税控系统这个网络信息化平台，可以通过数据挖掘找出隐藏在大量历史数据中的有用数据模式，来辅助税收部门进行科学地科学的判断和合理的收税，使得税收工作更加客观公正，大幅度地提高税收管理效能。

1.2 研究现状

随着我国税收信息化工作的不断深入，在各级税务机关的信息系统内部已经积累了大量的基础数据。这些数据本应该为税务部门即决策者提供大量有用的信息，但是因为没有得到合理的分析利用，造成了大量的数据流失。但是在西方发达国家，数据挖掘与分析技术已经在政府税务部门应用的非常广泛并趋于成熟。澳大利亚税部门将数据挖掘技术应用于税务行业,系统经过九年的稳定运行，投人回报率达到1：15。美国加利福尼亚州特许税务委员会应用数据仓库解决方案，使得征税效率和政府税收收人大幅提高，并因此荣获2002年数据仓库协会(TDWI)最佳实践奖。

相比而言，我国在同类产品的研究和应用与西方国家还存在一定的差距。尽管数据挖掘已经不是一项新的研究领域，但在我国，应用在税收管理分析中却刚刚起步。还存在以下不足和缺陷：

(1) 数据异构、分散

从纳税单位申报上来的税收数据主要集中在地市级单位，并没有统一集中到省级或者总局。在实现了数据集成的大部门当中，大都把数据集中到低级的税务部门，没有建立真正意义的数据仓库。在数据库使用方面，存在Sybase、Oracle、SQL Server 等非单一软件，其中有些单位甚至使用了两种以上的数据库软件。这样就不利于税务部门进行数据监控与分析。

(2) 未监控真正的数据指标

这也是当前税收管理中最为严重的问题现在的税收分析只是针对各企业各单位上报的税收数据加以分析利用，而高层部门并没有真正的得到企业的销售数据以及盈利情况等。简而言之，就是一种盲目的征管，没有真正到达以票控税的目的。

另外还有很多问题，比如挖掘什么、税务部门关心哪些有用的指标等还没有明确的限定。这就需要我们认真把握数据挖掘与纳税评估的关系，数据挖掘与数据分析的关系，数据挖掘在税收工作中的地位等。总之，要明确数据挖掘在税收管理分析中的目标与任务，才更能突出它的作用与意义。

从技术角度上看，数据挖掘是直接服务于数据分析工作的技术手段，它不仅是税收数据分析的有力工具，而且代表着税收分析的发展方向，即智能化与自动化的决策支持；从政策角度上看，税收数据挖掘不仅是技术实现的过程，同时也是业务精细化与科学化的体现。数据挖掘是税收工作的侦听器，它发现税收征管的薄弱环节，成为税收分析、纳税评估、税务稽查、税收监控这一良性互动机制的发动机和触发器，直接关系到互动机制的运行质量，从而解决税务部门税收工作的盲目性和不合理性，为他们下一步的工作提供科学合理的决策支持，同时也

为企业的市场规划和市场管理提供了有力的证据和保障。

1.3 本课题的主要研究内容

商业税收是国家财政的重要组成部分，为了对商业销项税进行管理，掌控销项数据，国家先后出台了税控收款机标准 GB 18240.1~GB 18240.6 等六个部分，建立了以商业收款机+税控器+税控卡+税控收款机管理系统的基本工作模式，明确了以票控税的基本原则，有利地推动了商业税收工作的有序进行[27]。

本课题来源于国标 GB18240.7—商业自动化管理。GB18240.7 的标准制定面向大、中型商业流通企业，在其企业内已有的商业管理信息系统(MIS)基础上进行税控功能的改造。通过驻留在商用收款机操作系统核心层中的软件模块，时时监控企业的销售数据源，依照发票的使用情况缴纳税款，满足税务机关进行税收监管的工作需要。

面对大量的商场销售数据，如何构建数据挖掘平台是本课题的关键。数据挖掘系统由算法所支撑，然而各种算法都是有一定的针对性，针对特定的数据集寻找高效的算法变的尤为重要。学习算法要用到各种不同的参数，需要合适的参数值，选择适当的参数可以使获得的结果得到显著的改善。面对海量的销售数据，要高效的一次性处理是非常困难的。而计算机的内存是有限的，如何合理的分配内存，使之能够高效率的分析和处理数据也是本课题的重要研究内容。运用良好的数据挖掘平台挖掘出模式、知识，为税收监管部门提供科学的决策支持才是本文的最终目标。

本文共分为五章，每章的主要内容如下：

第一章为绪论，主要介绍了本课题的背景与研究意义、数据挖掘在税收管理分析中的作用以及当前国内外的研究现状，最后阐述了课题的来源以及本文的组织结构。

第二章介绍了数据挖掘技术，讲解了常用的一些挖掘算法。并就当前炙手可热的商业智能领域进行了一番论述。最后总结了两者密不可分的关系。

第三章开始进入我们数据挖掘的前半部分，数据的继承与预处理。针对课题国标 GB18240.7 中异构数据库集成技术进行了研究，并将其中的大量监控数据集中到同一的数据源，建立数据仓库，为后续的挖掘阶段打好坚实的基础。

第四章针对挖掘算法做文章，选择 Weka 作为数据挖掘平台（平台搭建），将改进的聚类算法整合于其中。

第五章应用平台，展开分析与挖掘

第 2 章 数据挖掘与商业智能

2.1 数据挖掘简介

一提到数据挖掘(Data Mining)，我们并不陌生。现实生活中信息数据堆积现象越来越普遍，并趋于严重化。大量信息在给人们带来方便的同时也带来了一大堆问题：一是信息过量，难以消化；二是信息真假难以辨识。人们开始慢慢学会“抛弃信息”，但是由此而引发海量数据中隐藏的知识常常被我们忽视而当作垃圾丢弃，这就迫切需要。人们逐渐开始考虑，如何才能从大量数据中及时发现有用的知识，提高信息利用率？数据挖掘和知识发现技术便应运而生。

数据挖掘是一个多学科交叉研究领域，如下图 2-1.它融合了数据库、人工智能、机器学习、统计学和面向对象方法等最新技术研究成果，并且正在以一种全新的概念改变着人类利用数据的方式。

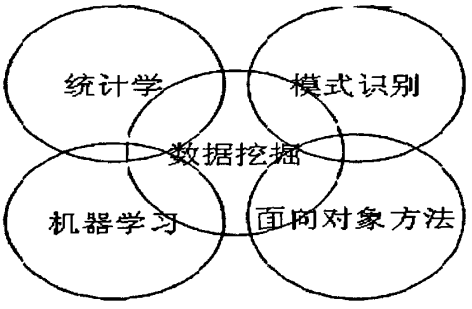

图 2-1 数据挖掘研究领域

Figure 2-1 research areas of data mining

2.1.1 数据挖掘的概念

数据挖掘，又称为数据中的知识发现(Knowledge Discovery in Database, KDD)和模式探索（Pattern Explore），就是通过一定的高效率算法，从大量数据中获取有效的、新颖的、潜在有用的、最终被人们可理解的模式的非平凡过程。简而言之，数据挖掘就是从大量数据中获取知识的过程[3]。

并非所有的从信息中发现任务都被视为数据挖掘。比如，使用数据库管理系统查找符合条件的记录，或通过因特网的搜索引擎查找特定的 Web 页面，则是信息检索（information retrieval）领域的任务。虽然这些任务是重要的，可能涉及使用复杂的算法和数据结构，但是它们主要依赖传统的计算机科学技术和数

据的明显特征来创建索引结构，从而有效地组织和检索信息。尽管如此，数据挖掘技术也已用来增强信息检索系统的能力。

2.1.2 数据挖掘的方法

数据挖掘的目的是发现隐藏的有价值的知识，而知识要通过一定的模式才能表现出来，数据挖掘中有许多知识表示模式及其所采用的方法，比如关联知识挖掘、类知识挖掘等。

要挖掘必须要有挖掘的对象。数据库作为常用的挖掘对象已屡见不鲜。数据库技术作为一种基本的信息存储和管理方式，仍然以联机事务处理（OLTP）为核心应用，缺少对决策、分析、预测等高级功能的支持（Decision Support）机制。随着数据库容量的膨胀，特别是数据仓库（Data Warehouse）以及Web、文本等新型数据源的日益普及，联机分析处理（OLAP）、分类、聚类方法等复杂应用逐渐成为必然，大量的挖掘方式和方法也渐渐地进入研究领域。通过对挖掘中知识模式以及挖掘方法的研究，我们可以更清楚地了解数据挖掘的本质和特点。
下面简要介绍一下常用的模式表示方式和方法[2][16][26]：

1．关联分析

关联关系表达的是一个事件和另一个事件之间的依赖关系。关联分析，即利用关联规则找出数据之间联系的方法。它主要是指集中在数据库对象之间的关联程度的刻画。人们提出了多种关联规则的挖掘算法，如STEM、AIS、DHP 等算法。最为著名的是Agrawal 等提出的Apriori 及其改进算法，它表示了一组项目关联在一起的需要满足的最低联系程度。关联规则的研究是数据挖掘中比较常用的方法并日渐趋于成熟。

2．分类（Classification）挖掘

分类分析是数据挖掘中一个重要的目标和任务，目前应用在商业中比较多。分类的目标是构造一个分类的模型，该模型可以将数据库中的数据按照指定的规则映射到给定的类别当中去。依照此规则，数据库中的所有信息总体以几大特征（几类）来最终呈现。要构造此分类器，必须要抽出一个数据样本作为原始输入源，然后对源数据进行过滤、抽取、以及概念提取等。构造分类器的方法大体有以下几种：

（1） 决策树：经常使用分治策略来处理决策树问题，但是要慎重考虑训练数据过渡拟合的情况，特别是推广到独立的训练集上。ID3 算法是最典型的决策树分类算法，之后ID4，ID5，C4.5 等都对其做了进一步改进，但是他们的缺点就是对大训练样本集很难适应

（2） 贝叶斯分类：来源于概率统计学，并且在机器学习中被很好的应用。

贝叶斯分类器的分类原理是通过某对象的先验概率，利用贝叶斯公式计算出其后验概率，即该对象属于某一类的概率，选择具有最大后验概率的类作为该对象所属的类。目前研究较多的贝叶斯分类器主要有 Naive Bayes、TAN、BAN 和 GBN。

（3） 人工神经网络：神经网络技术是一个独立的研究分支。由于需要较长的时间和其可解性较差，为它的应用带来了苦难。但是，人工神经网络通过对局部情况的对照比较(而这些比较是基于不同情况下的自动学习和要实际解决的问题的复杂性所决定的)，它能够推理产生一个可以自动识别的系统，具有较强的干扰力。

（4） 遗传算法：它是一类可用于复杂系统优化的具有鲁棒性的搜索算法，是一种基于进化理论的机器学习算法。由于与传统的优化算法相比，它具有以决策变量的编码作为运算对象、以适应度作为搜索信息、使用多个点的搜索信息以及使用概率搜索技术等特点，在函数优化、组合优化等研究领域等到了很好的应用。

分类规则是知识发现中应用最为广泛的数据挖掘技术。例如，金融业中可以通过客户分类构造一个分类模型来对银行贷款进行风险评估；当前的市场营销中很重要的一个特点是强调客户细分。客户类别分析的功能也在于此，采用分类技术，可以将客户分成不同的几大类别，比如呼叫中心设计时可以分为呼叫频繁的客户、偶然大量呼叫的客户、稳定呼叫的客户等，帮助呼叫中心寻找出这些不同种类客户之间的特征，这样的分类模型可以让用户了解不同行为类别客户的分布特征；另外在设计一个电子商店时，要涉及到商品分类的原则；安全领域有基于分类技术的入侵检测等。总之在数据挖掘和机器学习领域、分类规则起着不可替代的作用。

3. 聚类（Cluster）分析

聚类是以统计学、机器学习等为依托，把一组个体按照相似性规则归成若干个类的方法，目的是使的属于同一类别的个体之间的差别尽可能的小，而不同类别上的个体间的差别尽可能的大。聚类分析是由若干模式组成的。通常，模式是一个度量的向量，或者是多维空间中的一个点。聚类分析以相似性为基础，在一个聚类中的模式之间比不在同一聚类中的模式之间具有更多的相似性。

聚类的用途是很广泛的。在商业上，聚类可以帮助市场分析人员从商业 MIS 数据库中区分出不同的消费群体来，并且概括出每一类消费者的消费模式或消费观念，可以帮助税务部门更好的了解企业或个人的消费行为，这也是本文介绍的重点内容。它作为数据挖掘中的一个模块，可以作为一个单独的工具以发现数据库中分布的一些深层的信息，并且概括出每一类的特点，或者把注意力放在某一个特定的类上以作进一步的分析；并且，聚类分析也可以作为数据挖掘算法中其他分析算法的一个预处理步骤。

2000年，Han等研究者归纳了基于分类、层次、密度、网格和模型等五大聚类算法，它们在目前的应用中具有典型的代表性：

（1）分裂法(partitioning methods)：给定一个有N个元组或者纪录的数据集，分裂法将构造K个分组，每一个分组就代表一个聚类，K<N。而且这K个分组满足下列条件：一是每一个分组至少包含一个数据纪录；二是每一个数据纪录属于且仅属于一个分组；对于给定的K，算法首先给出一个初始的分组方法，以后通过反复迭代的方法改变分组，使得每一次改进之后的分组方案都较前一次好，而所谓好的标准就是：同一分组中的记录越近越好，而不同分组中的纪录越远越好。使用这个基本思想的算法有：K-MEANS算法、K-MEDOIDS算法、CLARANS算法。

（2）层次法(hierarchical methods)：这种方法对给定的数据集进行层次似的分解，直到某种条件满足为止。具体又可分为“自底向上”和“自顶向下”两种方案。例如在“自底向上”方案中，初始时每一个数据纪录都组成一个单独的组，在接下来的迭代中，它把那些相互邻近的组合并成一个组，直到所有的记录组成一个分组或者某个条件满足为止。代表算法有：BIRCH算法、CURE算法、CHAMELEON算法等；

（3）基于密度的方法(density-based methods)：基于密度的方法与其它方法的一个根本区别是：它不是基于各种各样的距离的，而是基于密度的。这样就能克服基于距离的算法只能发现“类圆形”的聚类的缺点。这个方法的知道思想就是，只要一个区域中的点的密度大过某个阀值，就把它加到与之相近的聚类中去。代表算法有：DBSCAN算法、OPTICS算法、DENCLUE算法等；

（4）基于网格的方法(grid-based methods)：这种方法首先将数据空间划分成为有限个单元（cell）的网格结构，所有的处理都是以单个的单元为对象的。这么处理的一个突出的优点就是处理速度很快，通常这是与目标数据库中记录的个数无关的，它只与把数据空间分为多少个单元有关。代表算法有：STING算法、CLIQUE算法、WAVE-CLUSTER算法；

（5）基于模型的方法(model-based methods)：基于模型的方法给每一个聚类假定一个模型，然后去寻找能很好的满足这个模型的数据集。这样一个模型可能是数据点在空间中的密度分布函数或者其它。它的一个潜在的假定就是：目标数据集是由一系列的概率分布所决定的。通常有两种尝试方向：统计的方案和神经网络的方案。

俗话说：“物以类聚，人以群分”。在自然科学和社会科学当中，存在着大量的分类聚类问题。分类分析与聚类分析相辅相成，它们之间既存在相同点也存在不同点。聚类分析是研究分类问题的一种统计分析方法，起源于分析方法学。他们的目标最终都是把特定的数据源归成几类，但聚类与分类不同。前者是通过对

数据的分析生成新的类标识，而后者是在特定的类标识下找出新元素的归属类。聚类没有训练事例和预先定义的类标识。在通常情况下，聚类分析形成一些概念，即一组数据可以用一个概念来概括，由此大量的源数据可以按照一定的算法归纳成几个类或簇，这样一来我们最终可以根据不同簇的特点性质得出不同的结论。在税收管理分析中，根据不同商品的销售情况得出不同的模式，采用不同的措施是我国税收征管的必然趋势。

4. 预测性分析

预测是数据挖掘中非常重要的任务之一。它指的是根据历史的和当前的实例数据总结出知识、模式，并能推测未来数据趋势走势的方法。预测型挖掘主要有两大方法：分类预测和时间序列预测。

（1）分类预测：首先对输入样本进行分析处理，得出数据的分类模型。这个过程可以利用分类技术的各种算法（决策数、遗传算法等），然后利用当前剩余的大量数据进行模型的验证并不断地对模型进行修正，最后对未来新的数据依照模型进行归类，达到预测的目的。

（2）统计预测：和分类预测有很大的不同，在统计学中的预测是指根据时间序列建立数学模型，然后对未来的某一时刻可能发生的情况进行预测。由于这类预测方法是以时间为关键属性的，所以可以称为时间序列分析。如对数据源中某一个变量 x(t)按照时间先后顺序进行观察和分析，在一段观察时刻 t1，t2，t3…tn(t 为时间的先后顺序)，会得到一组离散的数值。这就组成了一个时间序列集合。时间序列分析是对系统观测得到的时间序列数据建立模型的理论依据，一般采用曲线拟合、参数估计和回归预测方法。

预测型的数据挖掘是建立在统计学、神经网络和机器学习技术之上的。现在已经有了成熟的几类模式：

（1）趋势预测模式：主要针对那些具有时序属性的数据，如股票价格，蔬菜价格等。

（2）序列模式：主要是指在一段时间内根据某几个事件序列发生的次序以及出现的频繁程度来进行预测的模式。例如在商场销售中，很多顾客先买了油漆，然后买家具，接着买家电，那么在<油漆，家具，家电>就很有可能是一条序列模式。

（3）神经网络模式：通过对历史数据的分析建立神经网络模型，但是要学会基于时时数据不断地更新此网络模型。

预测分析一直是数据分析的目标，作为数据挖掘的一个有重要实际意义的分支，在商场应用中，它可以从顾客序列中挖掘出大多数人的连续购买模式，并且可以帮助税务部门确定大量交易数据中多种商品层次中隐含的鲜为人知的序列模式，对我国税收管理具有重要的现实意义。

2.2 商业智能

当商业智能（Business Intelligence）像旋风一样席卷国内时，BI 的概念就犹如满天飞絮一样，飘落在人们的脑海里。但是在当今的 IT 界，还是有相当多人对 BI 的认识和认知很浅薄。我们不得不承认商业智能的出现与数据挖掘是分不开的。伴随着经济的商业化，对商业数据分析和处理的需求越来越强烈，可以说商业智能为我们正确的了解和应用商务活动提供了一种解决方案。

2.2.1 商业智能的概念

商业智能的概念最早是 Gartner Group 于 1996 年提出来的。当时将商业智能定义为一类由数据仓库（或数据集市）、查询报表、联机分析、数据挖掘、数据备份和恢复等部分组成的，以帮助企业决策为目的的技术及其应用。而现在商业智能有了更新更深刻的含义：

1. 行业应用 BI 解决方案的价值已经逐步成为企业之间竞争的有力武器和目标追求

BI 应用作为近几年中国 IT 界增长最快的一个领域，在国外该系列产品可以卖到几十万甚至上百万人民币，国内产品也要卖到几万到几十万人民币。企业只要认知了这个平台价值，它就能给在当今市场竞争激励的 IT 企业带来新的利润增长点，创造高额的利润。其实，许多 IT 企业经过多年的运作，在不同的领域里都积累了不少的“行业经验”，例如在保险、电信、公安等等领域都有相当丰富的“行业经验”和资源，如果能充分认知 BI，并能借助 BI 工具的优势，搭建行业 BI 解决方案平台，将会使企业如虎添翼，充分发挥出“行业经验”的价值。

2. BI 中的统计报表与分析挖掘

首先现代商务中对报表的需求已逐渐加大。而 BI 中的信息处理包括查询和基本的统计分析，如使用交叉表、图表或者图进行报表的展示。分析处理支持基本的 OLAP 操作，如上钻、下钻、旋转、切片和切块等，其表现形式也大都以报表为主，并且数据源并不只包含传统报表的数据库数据源，而且融入了各种各样的跟企业运作相关的数据形式。其次引入了数据仓库的概念。因为数据仓库够大、够清楚、够全面，并且对统计分析需要的数据源支持得够好，这些优点是传统的数据库没有的，数据仓库的概念后面一章会有所介绍。

2.2.2 商业智能的工具与基本步骤

商业智能的实现包含了“数据—信息—知识—智慧—决策”这一过程所运用

的技术和方法。在国外已经具有广阔的应用前景，吸引了相当对的软件商为其提供解决方案。其中有 Microsoft、IBM、Oracle、Sybase 以及 SAS 公司等。一个完整的 BI 应用需要 ETL 工具、数据仓库工具、OLAP 工具、数据挖掘工具和报表查询工具。表 2-1 列出了这方面的几款主要产品[4]：

表 2-1 商业智能的几款常见产品

Table 2-1 some common business intelligence products

公司名称	ETL 工具	数据仓库管理工具	OLAP 工具	数据挖掘工具	报表工具
Microsoft	SSIS	SQL Server	SSAS	SSAS	SSRS
IBM	Warehouse Manager	Visual Warehouse	OLAP Server	Intelligent Miner	Insight
Oracle	ETL Server	Enterprise Manager	Express Server	Darwin	Express Analyser
Sybase	Replication Server PowerStage	Warehouse Studio	Warehouse Analyzer	SAS SPSS	Info Maker

Weka 作为研究和学习的开源挖掘工具也得到了良好的应用。其中集合了大量能承担数据挖掘任务的机器学习算法，包括对数据进行预处理，分类，回归、聚类、关联规则以及在新的交互式界面上的可视化。与以上工具相比，在非商业领域占有一席之地，本文后续会有详细的研究。

除了选取合适的工具以外，还要明确 BI 实施的目标任务，并且按照正确的逻辑步骤，才会取得良好的实效。以下是实施商业智能的主要步骤：

（1）需求分析：需求分析是商业智能运作的第一步，在其他活动开展之前必须明确的定义企业对商业智能的期望和需求。包括需要分析的主题，各主题可能查看的角度(维度)，需要发现企业那些方面的规律等。总之，用户的需求必须明确.

（2）数据仓库建模：通过对客户需求的分析，建立企业数据仓库的逻辑模型和物理模型,并规划好系统的应用架构,将企业各类数据按照分析主题进行组织和归类。

（3）数据抽取:数据仓库建立后需将数据从业务系统中抽取到数据仓库中，简称数据集中。在抽取的过程中还必须将数据进行预处理、转换、清洗，以适应后面分析的需要。

（4）挖掘平台搭建：商业智能的关键。根据需求选择合适的挖掘算法创建智能分析平台。基于此平台可以轻松的创建商业智能分析报表。用良好的界面形式呈现给客户。

（5）模式知识的发现：利用创建好的智能系统，通过对大量数据的分析处理，隐藏的、为决策者所关心的知识模式是不难发现的。

2.2.3 商业智能的作用与意义

下面 2-2 以图示的形式展示了商业智能带给我们的方便与快捷。

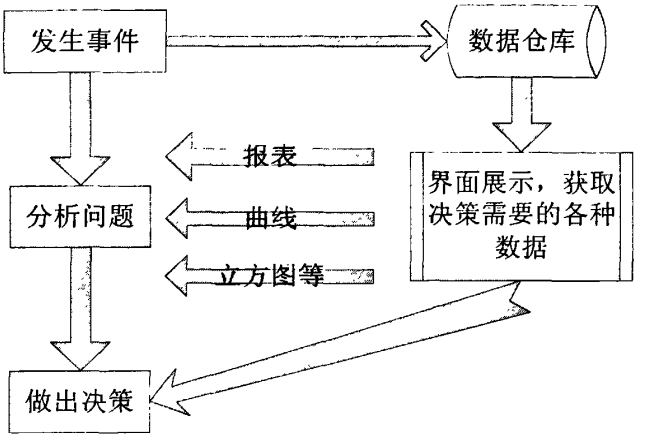

图 2-2 商业智能的作用

Figure 2-2 the role of business intelligence

由图可以看出运用商业智能后，企业内的信息都日常性地保存到企业的数据仓库中，以备决策者做决策时对信息访问的需要。决策者获得这些信息不需要再依赖于传统信息交换流程中的纸质报表、手工数据汇总的、落后耗时费力的信息传递方式，他通过非常简单的方式访问企业数据仓库，就可以访问到他在决策过程中需要的所有信息，而且这些信息的访问界面可以是为他的需要量身订做的。同时由于信息获取过程中完全的自动化和规范化，降低了由于人工传递信息而带来的无法避免的信息残缺和误差，使获取信息的准确性得到有力的保证，为企业决策与战略调整节省了大量的宝贵时间。

2.3 数据挖掘与商业智能的依存关系

提到数据挖掘，就不能不提商业智能。总的说来，数据挖掘是技术，商业智能是形态，两者是一个统一体，互为补充。商业智能技术并不是基础技术或者产品技术，它是数据仓库、联机分析处理 OLAP 和数据挖掘等相关技术走向商业应用后形成的一种应用技术。商业智能的本质就是将数据挖掘的智能计算技术应用于传统商业领域，从而提高数据分析能力，优化业务过程和规则，提高企业竞争力。虽然商业智能的普及仅仅是最近几年的事情，但已经渗透到金融，电信，零售，医药，制造，政府等各个行业和领域，成为大中型企业经营决策的重要组成部分。数据挖掘已经逐渐成为商业智能系统的高层应用，是不可或缺的重要部分。

数据挖掘是一项技术，由许许多多的算法构成，并且每种算法可以有多种实现方式。数据挖掘渗透到某些行业，产生了一些特定的应用，就形成了商业智能。

比如现在经常会听到的客户关系管理(CRM)。客户关系管理的概念出来已久，但现代的客户关系管理一般指以客户数据为处理对象的一类商业智能应用。通过挖掘客户信息，发现潜在的消费趋势或动向。比如电信公司通过分析用户通话模式(通话时间，时段，通话量等)，制订不同的计费方案，满足用户的同时也提高自己的利润。

随着数据挖掘技术的不断改进和日益成熟，它必将被更多的用户采用，使更多的管理者得到更多的商务智能。

2.4 本课题的应用需求

市场经济条件下，具有综合数据的能力并对数据进行快速和准确分析，从而做出更好的商业决策，可以为企业带来竞争优势。如何发现和使用这种优势，就是商业智能所研究的课题。

在日益竞争激烈的市场环境下，企业的压力自然会比较大，同时也会产生一些所谓“合理避税”的不规范现象。这些不规范的现象造成了市场管理的无序化，影响着国家的财政税收，也影响着企业的健康发展，任其发展会对国家的经济秩序产生不良的影响。

增强国家税控(国家税控收款机标准第 7 部分：商业自动化管理)，保障合理有序竞争是建立国家税收制度的基本要求。如果能在影响力较大的大中型商场实现有力的税控制度则会影响到几亿群众的税收意识。由于技术条件的限制，以往税务机关难以实现及时的管理和稽查工作，而事后问题的查处又很难全面的展开，久而久之形成了纳税人的侥幸心理。这种心理普遍影响着每位纳税人。本标准的实施能够将电子信息技术的发展与目前大中型商场、超市、连锁店的商业形态相结合，可以保护税务机关在最快的时间内得到商业企业的销售数据，同时得到该企业的销项税额。这是一项利国利民的工作。

本课题主要从商业智能研究的角度出发，以国标 18240.7 为背景，以企业销售数据为源泉。每笔帐单的销售明细也都有记录。从中找出企业上报的销售数据和系统监控得到的数据之间的差距，以及在税务稽查期间给稽核人员建立一个友好的标准的交互界面模型，选择一个良好的交互方式是本文研究的重点内容[29][30]。

第 3 章 数据抽取与集成

3.1 数据仓库设计

要想进行数据的分析与挖掘，就必须提供一个有效的统一的数据源，而数据仓库作为大量数据的统一载体，以其数据规范性在商业智能领域发挥着越来越重要的作用。数据仓库系统是一个信息提供平台，他从业务处理系统获得数据，主要以星型模型和雪花模型进行数据组织，并为用户提供各种手段从数据中获取信息和知识。建立和利用数据仓库已经成为商业智能不可或缺的重要步骤之一，为OLAP、数据挖掘等深层次的数据分析提供统一的平台[37]。

3.1.1 数据仓库的概念

数据仓库之父Bill Inmon 1991 年出版的"Building the Data Warehouse"一书中所提出的定义被广泛接受—数据仓库（Data Warehouse）是一个面向主题的（Subject Oriented）、集成的（Integrated）、相对稳定的（Non-Volatile）、反映历史变化（Time Variant）的数据集合，用于支持管理决策(Decision Making Support)。

◆面向主题：通常情况下数据库的数据组织面向事务处理任务，目前较为流行的是关系数据库。在各个业务系统之间各自分离，而数据仓库中的数据是按照一定的主题域进行组织的。

◆集成的：数据仓库中的数据是在对原有分散的数据库数据抽取、清理的基础上经过系统加工、汇总和整理得到的，必须消除源数据中的不一致性，以保证数据仓库内的信息是关于整个企业的一致的全局信息。

◆相对稳定的：数据仓库的数据主要供企业决策分析之用，所涉及的数据操作主要是数据查询，一旦某个数据进入数据仓库以后，一般情况下将被长期保留，也就是数据仓库中一般有大量的查询操作，但修改和删除操作很少，通常只需要定期的加载、刷新。

◆反映历史变化：数据仓库中的数据通常包含历史信息，系统记录了企业从过去某一时点(如开始应用数据仓库的时点)到目前的各个阶段的信息，通过这些信息，可以对企业的发展历程和未来趋势做出定量分析和预测[38]。

也有人把数据仓库定义为"只提供具有统一模式的数据和大多数用户的概要数据，而不提供特定格式的数据和面向少量用户的细节数据"的数据集合。它要求我们数据仓库设计人员要明确哪些数据是用户关心的，针对特定的数据集建立数据仓库，确保仓库中数据的有效性规范性和避免数据的过于庞大和冗余。

综上所述，数据仓库是一种语义上一致的数据存储，它充当决策支持数据模型的物理实现，并存放企业战略决策所需信息。数据仓库也常常被视为一种体系结构，通过将异种数据源中的数据集成在一起而构成，支持结构化和专门的查询、分析报告和决策制定。

3.1.2 税控管理中数据仓库的构建

数据仓库中的数据源往往有两种：一是从事务数据库中周期性迁入的数据；二是企业购买的可以与内部数据相连通的外部数据。

数据仓库的特性就决定了它的设计不能采用同开发传统的 OLTP 数据库一样的设计方法。特别针对其随时间的变化的特点，我们必须从构建系统的简单的基本框架着手，不断丰富与完善整个仓库。具体实施步骤如下：

◆ 收集和分析业务需求

众所周知，数据仓库必须有一个或者几个明确的主题，只有确定了分析的主题，我们才能按部就班的从数据源头来收集需要的数据，这也是决定数据仓库本质的因素。

◆ 建立数据仓库的物理模型

数据仓库的物理模型是面向企业全局建立的，它为集成来自各个面向应用的数据库的数据提供了统一的概念视图。根据现有的需求，建立物理模型是整个数据仓库设计的关键。目前常用的物理模型主要有：

（1） 星形模型：当前最流行的也是最简洁的模型，它能够准确反映出个实体之间的关系。主要由事实表和维度表两部分构成。

（2） 雪花模型：是在星形模型的基础上发展起来的，它在事实表和维度表的基础上增加了详细类别表，用于对维度表进行描述。但是它仅限于一个事实表与多个维度表构成的一到多对应的关联关系，不能处理复杂的情况。

在物理模型设计中，必须对数据存储结构有一个良好的把握，应当对系统的存储时间、空间存储效率等做同一考虑。另外，由于数据仓库中数据量的庞大，因此添加适量的索引是必要，这样有利于提高运行效率。

◆ 从源数据库中抽取、净化、和转换数据到数据仓库

ETL 过程即数据抽取、转化和加载。在现代商务中，ETL 已被广泛应用于数据仓库与数据挖掘中。它是一种数据集成的过程，后续章节会有详细介绍和应用。

◆ 更新数据仓库

因为数据仓库特性之一是随时间变化的，企业的销售数据也是时时在增加变

化的这就需要我们不断地利用 ETL 技术注入最新的数据，这样才能准确的采集到时时数据，并进行分析与挖掘，制定出正确的决策支持。

我们采用自底向上的模式构建所需的数据仓库，见下图 3-1：

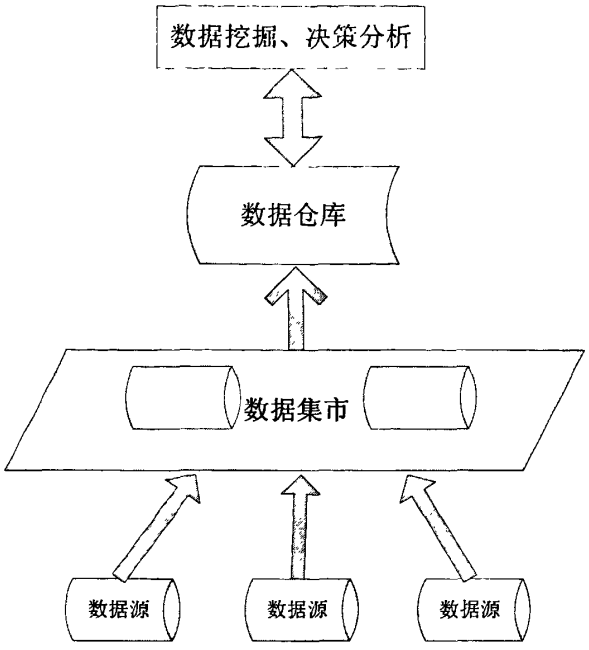

图 3-1 自底向上的数据仓库构建模式

Figure 3-1 bottom-up model of building a data warehouse

在商业自动收款机系统中，税控数据库服务器起着数据储存、连接网络税控器与发票终端、业务财务的桥梁作用。他主要用于收集网络税控器、业务服务器、财务服务器的税控数据，并完成税务管理部门税控功能要求的服务器。在其中，储存了大量的数据，如日常的 MIS 财务相关的业务数据、设备管理相关数据以及税务相关数据，以便适时进行报税和稽查，这也正是在现有的商场局域网络环境下加入税控服务器以达到以票控税的意义所在。如何对以上数据进行采集、集成和分析处理便是本文的研究点。

与商场销售相关的日、月销售报表以及发票明细、发票日交易数据等都是税务部门关心的信息。通过业务服务器上传到服务器的商场的各项销售报表与通过网络税控器上传的电子发票存根数据进行比对，从中找出其相关一致性和离异点，为税收稽查提供相对准确的方向。通过对发票明细以及日交易数据的汇总、分析，可以发现商品销售的模糊关联关系，以及税种税目的营销情况。

本课题按照上述所讲的星形模型构建税务分析的数据仓库。由于在税控系统服务器中存在大量的表，本题只针对某些用户感兴趣的部分设计数据仓库。结构如下图 3-2：

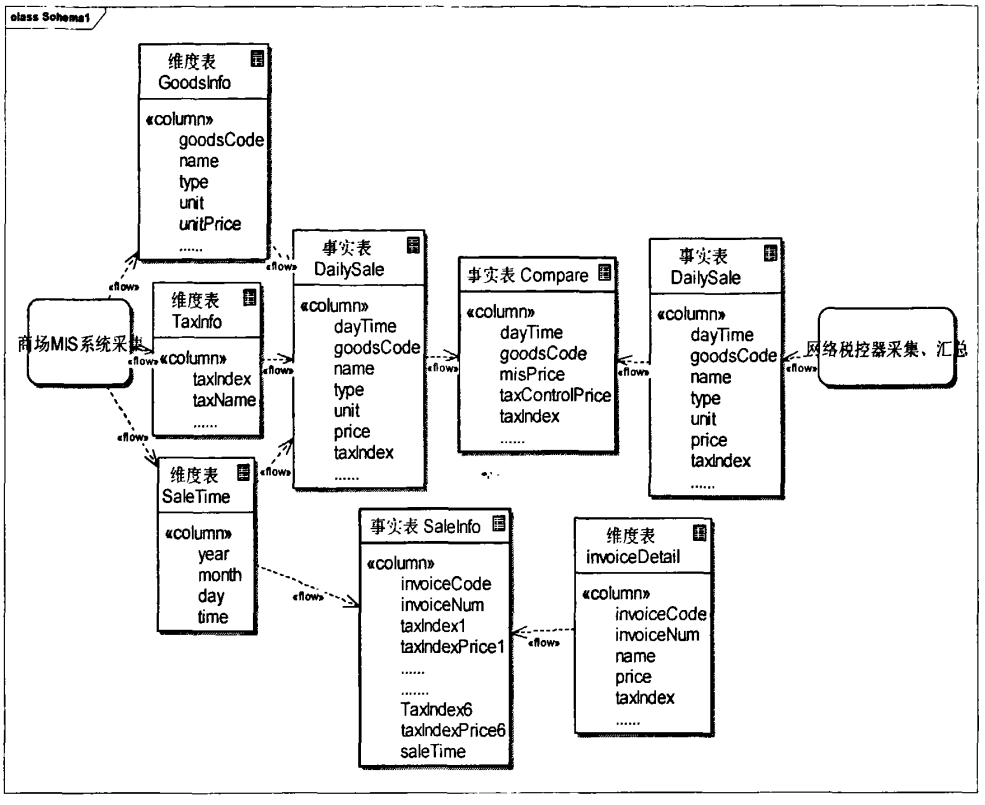

图 3-2 数据仓库部分结构设计

Figure 3-2 structural design of data warehouse

数据仓库设计以事实表 compare 和 SaleInfo 为主题中心，其中储存了大量数据待挖掘数据。事实表 Compare 连接商场 MIS 系统的日销售报表和网络税控器采集汇总的日销售报表，通过对比两者的差异可以找出商场是否存在漏报的嫌疑。而商场 MIS 的日报表是由商品基本信息维度表、税种税目信息维度表以及时间维度表来得到的。事实表 SaleInfo 定位到了每一张发票的使用情况，其中包括开票时间以及对于每种税目的销售情况等。根据国标 18240.7 的表设计结构，读者可以进行其余表的数据仓库设计，本题目的在于给予读者以设计模型与思路，并不涉及所有表的建模问题。

从图可以看出，事实表不仅是数据仓库的核心，还是构成数据仓库的表中体积较大的表。为了保证数据仓库的高效率，减少查询、备份等所需要的时间，必须注意数据的分割、粒度控制等环节。

3.2 异构数据库的集成

数据集成，简而言之就是将异构的数据源以及分布广泛的数据源利用一定的网络技术集成到统一的平台上来，建立大容量的数据仓库，从而解决企业信息孤

岛的问题。在国家标准即 GB18240.7 商业自动化管理中，业务和财务数据监控模块正是采用了这一数据集成技术实现了数据的监控。这两个模块是驻留在业务、财务服务器中的一个软件模快实现了数据的监控。这两个模块是驻留在业务、财务服务器中的一个软件模块，任务是向税控服务器提供及时的业务财务数据，税控服务器可以使用这些数据与通过网络税控器上传的电子发票存根与业务报表、财务报表进行比对，从中找出其相关一致性和离异点，为税收稽查提供相对准确的方向。比如当期现金流入量不能少于商品销售总额；日销售报表的销售月累计和月进销存报表的销售额应该与网络税控器提交的销售累计一致；网络税控器获取的销售数据和商品库存数据增减应基本一致等等。通过数据库映射技术可以集成各种异构数据源，从而在数据库层面上形成统一的访问接口，便于税收稽查人员进行有效的数据挖掘和分析。

3.2.1 XML 与数据库迁移技术

XML(Extensible Markup Language, 可扩展标记语言)是由W3C于1998年2月发布的一种标准规范。作为一种数据存储的载体，XML以一种开放的、自我描述方式定义了数据结构。在描述数据内容的同时能突出对结构的描述，从而体现出数据之间的关系。这样所组织的数据对于应用程序和用户都是友好的、可操作的。正是由于XML 所具有的自描述性、灵活性、强大的数据描述能力、数据交换能力以及可扩展性等优势，XML已经成为了当今重要的信息发布标准和表示技术之一，越来越多的应用之间通过XML来进行数据交换。

DTD（Documed Type Define，文档类型定义）就是要定义一门新的标记语言。将其作为数据存储的结构和全局数据模型给用户提供统一的用户视图。DTD文档可以在XML文档中被定义，也可以被独立定义在一个DTD（扩展名为.dtd）文件中，以便XML文档调用。XML文档的语法必须合乎DTD的定义，否则前者不是一个合法的XML文档，也不会被应用程序所解析。由于DTD存在历史悠久、数据类型有限以及与XML文档一对一的缺点，之后出现了Schema文档。Schema摒弃了DTD的缺点，遵循XML的语法要求，并且比DTD语法结构简单，采用了命名空间的机制，在代码的重用性和可扩展性方面要远远优于DTD，更易于我们学习与使用。

SAX（Simple API for XML）标准是一种操作XML数据文档的方法。SAX为应用程序操作XML文档提供了统一的接口。由于该技术是采取流的方式操作XML数据文件，因此对大量数据来说是反复的从硬盘读入内存和处理，这就避免一次性将大量数据装入内存，降低了对系统的要求，从而很好的避免了DOM标准将整个XML数据文件加载入内存的缺陷。

数据迁移是指将数据源中的数据迁移到目标数据库管理系统中来，并且可能要集成不同类型的数据，这就需要将一些非传统的数据类型转化成新的数据类型。利用数据库映射程序，从而使数据源中的数据能被其它的DBMS接收。使用这种方法处理的优点是简单经济，运行时效率高，它适合于对数据的实时性要求不高的场合，正符合GB18240.7标准。标准并不要求实时的传送数据，而是对一段时间内商场的销售数据与税控所采集到的数据做一个分析比对，找出其中的相通点和背离点，从而为税控稽查提供可靠的参考依据。

XML的跨平台性和强适应性符合数据迁移的要求，并因此而产生了大量的相关研究和相关产品。同时，随着XML技术影响的不断扩大，大量面向XML的应用层出不穷，客观上对基于XML的中间层服务和工具提出了大量需求，一些厂商和个人也纷纷投入到XML服务和工具的开发中来。XML 中间件(XML Middleware)应运而生。

3.2.2 异构数据源集成模型设计

3.2.2.1 XML 与关系数据库之间的映射方法

关系数据库是支持关系模型的数据库系统，一般采用关系作为单一的数据结构，数据的逻辑结构用二维表来表示。通常关系型数据到 XML 文档的映射规则为：表转为为元素，列转化为属性。而 XML 文档作为半结构化的数据文档，其依据是 XML Schema。本文研究的关键在于找到在结构化和半结构化的数据结构相互转化中的相通点，利用 XML 作为中间模式，将关系数据库表达的结构与约束通过标准 Schema 映射为 XML，并准确对 XML 数据文档进行解析，读入目标数据库。

根据映射关系建立方式的不同，可以分为两种数据映射方法：模板驱动的映射方法和模型驱动的映射方法。本文主要介绍前者。它主要应用于关系数据库与 XML 的数据传送。通过定义良好的文档标记语言，达到数据库与 XML 集成模式之间的双向映射。

3.2.2.2 XML 与数据库接口

XML 文档的数据源来自源数据库，XML 文档同样应用于目标数据库的更新，因此必须借助于数据库编程接口得以实现。通过对数据库的查询，分析，利用 ODBC，ADO 和 JDBC 等数据库访问接口，对数据进行添加、更新和删除等操作[42]。

对 XML 文档的操作可以分为生成和解析读入两个阶段。生成阶段借助数据源的各种数据结构形成统一的 XML Schema 形式，为源数据类型和结构提供统一的 XML 模式，在不改变源 DBMS 本身的基础上完成了各类数据源到集成模式的映射。然后借助于 XML 文档访问接口 SAX 生成符合标准的 XML 数据文档。

处理 XML 数据文档需要一个 XML 解释器。解释器为应用程序提供现成的读写接口，从而能够很好的从文档中识别标记、标记名称等。Xerces 就是一种标准 XML 解释器。由 Apache 开发组织的 Xerces 组维护。从 JDK1.5 以后，它便成为了 JDK 的默认实现。

JAXP 是 SUN 公司提出的一种 java 操作 XML 数据文档的标准，作用是在 Java 应用程序与解释器之间提供一个统一的编程接口，从而应用程序的可移植性。但其本身不是解释器，不能替代 SAX 接口的作用。其规范了应用程序获取 SAX 接口的方式[41]。下图 3-3 表示了 XML 解释器、SAX、JAXP 以及应用程序之间的关系。

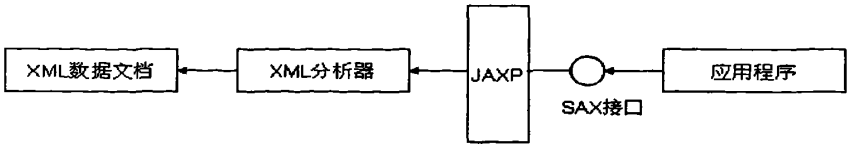

图 3-3 JAXP、解释器以及 SAX 关系图

Figure 3-3 relationship of JAXP, explainer and SAX

从图中可以看出，应用程序并不能直接访问 XML 文档。而是首先通过分析器解析 XML 文件，然后利用接口框架 JAXP 对数据源进行操作，完成了间接访问 XML 文档。

3.2.2.3 SAX 解析的基本原理

SAX 是为流文档准备的标准，处理过程上“读入一段数据，处理一段数据”。从而有效的处理了业务数据库数据庞大，不能一次读入内存的问题。其解析 XML 数据的基本机制是广播，SAX 解析其会将解析到的 XML 数据中的各种结构以事件的形式广播给特定的事件接收器和处理器。该机制主要存在以下几个关键点：

（1）SAX 事件接收器，接受 SAX 分析 XML 数据过程中的事件流。

（2）SAX 解析器，对 XML 文档文件进行分析的主要程序。

（3）XML 文档文件，其中存储了大量待分析数据。

3.2.2.4 HanldeBase 模式

作为用 SAX 处理关键的事件处理器由一些特定的接口组成。通过制定实现了接口的类，就可以监听 SAX 解析器广播出来的事件信。按照 SAX 处理 XML 数据文档方式的不同可以分为 XMLReader 模式、DefaultHandler 模式和 HandlerBase 模式。本文采用 HandlerBase 模式，它主要实现了以下几个核心接口：

（1）DocumentHandler 接口，负责文档的内容的处理。

（2）EntityResolver 接口，负责实体部分的解析。

（3）ErrorHandler 接口，负责解析错误的处理。

正是由于 XML 的平台无关性和易于扩展性，而且能方便的描述结构化或非结构化的数据,因此本文将其作为处理异构数据源的统一的数据模型，实现了

XML 数据集成技术。当用户执行一个查询映射的时候 ,只需提交查询指令至数据映射中间件,中间件先是经过一系列的操作,去源数据源取出数据并将其转化为 XML 模式，然后访问目标数据源，读入 XML 文件,更新目标数据库，最后将映射结果提交返回给用户，系统实现的主要架构如下图 3-4 所示：

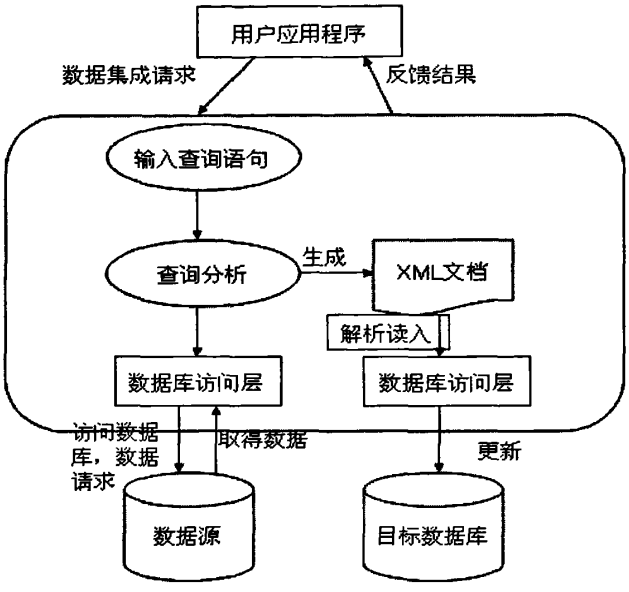

图 3-4 异构数据库集成系统架构图

Figure 3-4 heterogeneous database integration system architecture diagram

3.2.3 从 SQL Server 到 PostgreSQL 数据迁移的实现

SQL Server 和 PostgreSQL 都是目前比较流行的数据库管理系统。作为一种 Linux 下常用自由软件数据库系统，PostgreSQL 是一种功能强大复杂的对象—关系数据库管理系统，拥有一些商业数据库所没有特性，比如查询优化方面等。对其各方面性能的研究具有重要意义。

在 GB18240.7 中，业务、财务数据监控模块采用 SQL Server 作为后台数据库，其中存有大量的企业销售数据，包括月销售报表，月进货报表和库存月报表等。而税控服务器采用的 Linux 下 PostgreSQL 数据库，其中的存储的大量销售数据是通过嵌入在 pos 机内部的监控软件通过一定的网络协议上传上来的。因此将业务、财务服务器上的销售数据集成到服务器上来，对两者进行合理的比对和分析，对国家的税收监控和分析有着重大的意义。

3.2.3.1 从 SQL Server 中提取数据形成 XML 文档

XML 数据文档是异构数据库集成的中间件模式，其形成的过程如图 3-5：

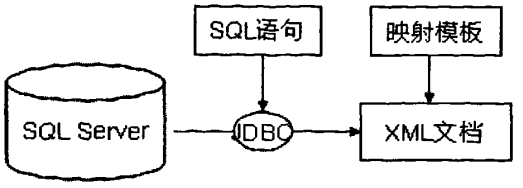

图 3-5 XML 文档形成过程
Figure 3-5 formation of XML documents

（1）用户可以通过界面输入源数据库地址和数据库类型的方式连接 SQL Server 数据库。本文采用 JDBC 来实现 SQL Server 数据库的访问：

```
Class.forName("com.microsoft.jdbc.sqlserver.SQLServerDriver");
String url ="jdbc:microsoft:sqlserver://192.168.1.5:1433;DatabaseName=Bus_Data";
Connection con = DriverManager.getConnection(url,"","");  //数据库用户名密码都为空
```

（2）以商品的日销售报表为例，XML 文档数据模型的定义如下：

```
<? xml version="1.0"? >
<mapping>
    <sourceData sql = "select * from DailySale">
    </sourceData>
    <root name = "dailySales" rowName = "dailySale" >
        <element name ="Date">     // 销售日期
            <attribute dateType = "dateType" >datetime</ attribute>
            <content>date</content >
        <element>
        <element name ="CCode">  // 商品编码
            <attribute dateType =" dateType">string</attribute>
            <content>cCode</content>
        <element>
        .... // 商品名称
        ....
    </root>
</mapping>
```

（3）创建 XML 数据文档

首先要成功解析以上映射文件，找出数据源即 sourceData。然后利用 sql 语句生成一个 Document 文档，根据已解析的文档映射关系检索根元素和行元素的信息。最后以 dailySale 为行元素的形式从头到尾循环地检索 Document 文档，并

同时向其中添加属性信息。生成的 XML 文档范式 dailySale.xml 如下：

```
<? xml version="1.0"? >
<dailySales>
    <dailySale date="25/3/2008" cCode="100000010001">
        <cName>水杯</cName> // 商品名称
        <bCode>200000010001</bCode> // 条形码
        <type>生活用品</type> // 类别
        <spec>800ml</spec> // 规格
        <unit>个</unit> // 单位
    <dailySale>
    <dailySale>
    ...
    ...
<dailySales>
```

3.2.3.1 XML 文档数据的导入

将 XML 文档数据导入到 PostgreSQL 数据库的流程如图 3-6：

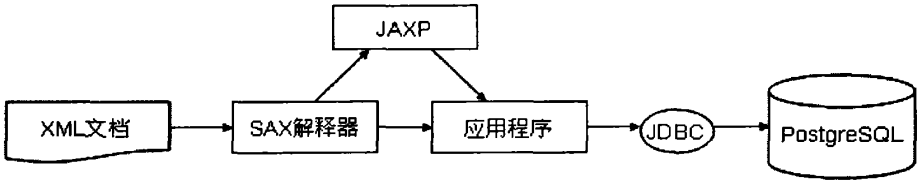

图 3-6 XML 解析导入流程图

Figure 3-6 flow chart of XML import and analysis

（1）通过界面输入目标数据库地址和数据库类型的方式使用 JDBC 连接 PostgreSQL 数据库。

```
Class.forName( " org.postgresql.Driver " ).newInstance();
String url = " jdbc:postgresql://localhost:5432/postgres " ;
Connection con = DriverManager.getConnection(url, " postgres " , "" ;
```

（2）采用 HandleBase 模式，应用程序通过 JAXP 以 SAX 的方式解析 XML，并将解析出的数据更新到 PostgreSQL。整个处理过程采用流的方式，将 XML 文档看作一个字符串集。SAX 事件处理器逐次扫描该字符串，并对遇到的标记和属性进行正确的解析读取，最后完成数据库的移植。程序实现的代码如下：

```
import javax.xml.parsers.*;
import org.xml.sax.*;
import org.xml.sax.helpers.*
import java.io.*;
public class XmlParse extends HandlerBase
{
    public void startDocument() throws SAXException{
```

```
        System.out.println("开始读取文档数据");
    }
    public void endDocument() throws SAXException{
        System.out.println("文档解析完毕");
    }
    Public void startElement(String names, AttributeList attributes) throws SAXException
    {
        For(int i = 0;i< attributes.getLength();i++)
        {
            /* 利用 attributes.getName(i)和 attributes.getValue(i)取得属性名称和属性值
        插入数据库*/
            …
        }
    }
    Public void characters(char[] ch, int start, int length) throws SAXException
    {
        // 如果有元素数据值则插入到目标数据库,读者可以自己实现
        …
    }
    …
    Public static void main()(String[] args)
    {
        Try{
                SAXParser parser = null;
                SAXParserFactory parserFactory = SAXParserFactory.newInstance();
                parser = parserFactory.newSAXParser();
                parser.parse("dailySale.xml", new XmlParse());  // 文档的读入
        }catch(Exception e) {e.printStackTrace();}
    }
}
```

在处理过程中，根据解析的字符串的不同会触发 DocumentHandler 的不同方法。例如，遇到开始标记<dailySale>，便会触发 startElement()方法，并将 dailySale 的名称和所有属性当作参数传给此方法。其中，可以对所需要的数据进行分析和处理，例如更新数据库等操作。另外，XMLReader 模式和 DefaultHandler 模式也都能很好的处理 XML，其处理方式与 HandlerBase 有所不同，由于篇幅所限这里就不在赘述。

3.3 数据清洗与建模

我们知道，在对数据仓库进行数据分析前必须保证数据的完整性、准确性和可靠性。然而，从业务服务器集成过来的数据难免会出现噪声、空值或丢失值等

问题，所以在数据挖掘之前排除一切对分析不利的影响是非常有必要的。下面简要一下介绍处理数据异常的方法以及 ETL 处理之后数据集的情况。

3.3.1 数据清洗

1. 对离散的字段使用一个合适的常量填充丢失值。

虽然这种方法比较耗时，但是此方法能够产生出准确的数据模型，从而直接影响后期的挖掘效果。例如当商品类别或者商品的税种税目索引丢失时，可以根据商品名称进行估量。例如某钢笔的税目索引为空，可以根据其属于办公文具类将其归为税率为 10%的最低税率税目。

2. 对于一个连续的字段可以模糊计算出丢失值。

一种方法，可以使用周围临近值的均值计算出丢失值。以商品销售金额为例，当出现空值时，我们可以按照商品编号对整个数据集随机的取出一些样本实例，计算出均值或者中值予以填充。

另外一种方法就是利用数据挖掘技术计算出概率最大的数值取代不完整的数值。例如，可以使用回归、遗传算法，也可以使用最近临（用最近的已知数据点来估算该商品销售额）、基于统计的方差方法。

但是所有以上的方法都存在一个问题，因为不管采用何种方法，产生出来的数据都存在或多或少的误差。所以很有必要对填充的或者修改的值做一定的校验。例如，在商场销售时，周末的营业额会高出平时 50%或 100%，所以对计算出的不符合常规的数据要予以增补或消减。

3. 对 NULL 值的处理

对于数据库中存在的 NULL 值易采用关联规则算法和决策树算法进行处理，它们的训练时间短而且可以产生更加精确的模型。

3.3.2 数据建模

经过以上处理，数据的噪声、丢失问题已经解决，数据库的完整性已经确立，以业务数据库和税控服务器提供的商品日报表为例建模。由于两者提供的表结构除商品销售额外都是一致的，因此建模的目的就是将两者的销售数据整合到一张表上。可以采用内联接的方式对同种商品同一天的销售情况做对比，找出差额以及内在的联系。

首先要建立数据仓库事实表，以 4 月份的月销售为例建立 SQL 语句如下：

```
CREATE TABLE "DailySale-April"
(
  "商品编号" character varying(10) NOT NULL,
```

```
"商品名称" character varying,
"商品种类" character varying,
"税种税目索引" smallint,
"日期" date NOT NULL,
"业务销售额" double precision,
"税控销售额" double precision,
CONSTRAINT "dailySale-April_pkey" PRIMARY KEY ("商品编号", "日期
")
)
WITH (OIDS=FALSE);
ALTER TABLE "DailySale-April" OWNER TO postgres;
```

可以同时查看业务服务器和税控服务器的销售情况：

Select tableMis.*,tableTax.销售额 from tableMis inner join tableTax on tableMis.商品编码 = tableTax.商品编码 and tableMis.日期 = tableTax.日期;

其中 tableMis 为业务数据表，tableTax 为税控数据表，这样一来两者的销售数据可以在同一张表中显示。

最后建立事实表的数据模型，如下：

Insert into DailySale-April select tableMis.*,tableTax.销售额 from tableMis inner join tableTax on tableMis.商品编码 = tableTax.商品编码 and tableMis.日期 = tableTax.日期;

至此，待挖掘和分析的数据源已准备就绪（见下图 3-7），后面我将详细介绍数据挖掘和分析的方法和步骤。

Edit Data - PostgreSQL 8.3 (localhost:5432) - DWDB - monthsale

文件(F) 编辑(E) 视图(V) 帮助(H)

不限制

	商品编号 [PK] charact	商品单价 numeric	商品种类 character va	税种税目索引 smallint	日期 [PK] date	业务销售额 double preci	税控销售额 double preci
1	00100001	2	食品	1	2009-04-01	2000	2115
2	00100002	3	食品	1	2009-04-01	3210	4521
3	00100003	4.5	食品	1	2009-04-01	1586	1968
4	00100004	5	食品	1	2009-04-01	4560	4560
5	00100005	1.5	食品	1	2009-04-01	8000	9600
6	00100006	3.2	食品	1	2009-04-01	6582	7451
7	00100007	9.6	食品	1	2009-04-01	685	1250
8	00100008	7	食品	1	2009-04-01	658	890
9	00100009	7.1	食品	1	2009-04-01	5214	5962
10	00100010	8.6	食品	1	2009-04-01	10000.6	12035
11	00100011	8.5	食品	1	2009-04-01	3652	3895
12	00100012	2.5	食品	1	2009-04-01	60000	74152
13	00100013	2.8	食品	1	2009-04-01	2315	2365
14	00100014	3.6	食品	1	2009-04-01	12365	13568
15	00100015	6.5	食品	1	2009-04-01	5874	5896
16	00100016	8.1	食品	1	2009-04-01	4986	5026

图 3-7 挖掘数据源表结构

Figure 3-7 structure of mining table

3.4 本章小结

伴随着国家标准 18240.7 税控收款机商业自动化管理的出台，对数据迁移的要求也越来越高。针对此标准中两种异构数据库的不同，本章主要设计实现了 XML 文档的形成、解析等过程，完成了异构数据库的集成，为更好的理解和应用 XML 中间件技术提供了指导，并为下面章节的数据挖掘做了良好的铺垫。

第 4 章 数据挖掘平台的构建

从以上章节可以看出，数据挖掘和知识发现是本课题的研究重点，从众多的商场销售数据中找出模式是本文的关键。前几章节已经为此做了良好的铺垫，本章将从数据挖掘算法的角度讨论数据挖掘技术。针对税控、税收管理分析中的数据特点，选取合适的聚类算法，并选择良好的挖掘平台，将高效的算法整合于其中，构造出适合本课题研究的挖掘系统，进行优劣评估和决策分析，并努力做到具体问题具体分析。

4.1 基于划分的聚类算法的研究与改进

4.1.1 K-均值算法（KCM）

通过前面的讲解，相信读者对聚类算法已经有了一个大概的了解，本节将针对具体的一类算法作深入介绍。

K-均值算法（k-means algorithm）是一种基于划分的聚类算法，也称为 C 均值聚类算法，目前已经得到了广泛的应用。它的核心思想是把包含 n 个对象向量的数据集根据他们的属性分为 k 个分割 $G_i(i=1,2,\ldots,k)$，$k<n$。并求每个分割的聚类中心，使得非相似性（或距离）指标的价值函数（或目标函数）达到最小[21][28]。如下式：

$$V=\sum_{i=1}^{k}\sum_{x_j\in S_i}(x_j-\mu_i)^2 \qquad (4.1)$$

其中，Xj 为单个样本对象，Si 为 Xj 所属的聚簇集合，也可以理解为数据集合 N 的一个子集。μi 是群组 Si 内所有元素 xj 的重心，或叫中心点。它的目的在于找到集合中各个聚簇类的中心，目标是使各个群组内部的均方误差总和最小。

当选择欧几里德距离为组 j 中向量 xk 与相应聚类中心 ci 间的非相似性指标时，价值函数可定义为：

$$J=\sum_{i=1}^{c}Ji=\sum_{i=1}^{c}\Big(\sum_{k,x_k\in Gi}\|x_k-c_i\|^2\Big) \qquad (4.2)$$

这里 Ji 是组 i 内的价值函数。这样 Ji 的值依赖于 Gi 的几何特性和 ci 的位置。同样地，也可以用一个通用距离函数 $d(x_k,c_i)$代替组 i 中的向量 x_k，则相应的总价值函数可表示为：

$$J = \sum_{i=1}^{c} Ji = \sum_{i=1}^{c} (\sum_{k, x_k \in Gi} \mathrm{d}(x_k - c_i)) \qquad (4.3)$$

聚簇产生之后一般用一个 c×n 的二维隶属矩阵 U 来定义。如果第 j 个数据点 x_j 属于组 i，则 U 中的元素 u_{ij} 为 1；否则，该元素取 0。一旦确定聚类中心 ci，可导出如下使式（4.2）最小 u_{ij}：

$$u_{ij} = \begin{cases} 1 & \text{对每个} k \neq i, \text{如果} \left\| x_j - c_i \right\|^2 \leq \left\| x_j - c_k \right\|^2, \\ 0 & \text{其它} \end{cases} \qquad (4.4)$$

可以得出，如果 c_i 是 x_j 的最近的聚类中心，那么 x_j 属于组 i。由于一个给定数据只能属于一个组，所以隶属矩阵 U 具有如下性质：

$$\sum_{i=1}^{c} u_{ij} = 1, \qquad \forall j = 1, \ldots, n \qquad (4.5)$$

且

$$\sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij} = n \qquad (4.6)$$

另一方面，如果固定 u_{ij} 则使（6.2）式最小的聚类中心就是组 i 中所有元素向量的均值：

$$c_i = \frac{1}{|G_i|} \sum_{k, x_k \in G_i} x_k \ , \qquad (4.7)$$

这里$|G_i|$是 G_i 的规模或着说是所包含元素向量的个数。下面给出 K-均值算法的处理步骤（假定聚簇数为 k）：

（1） 初始化 k 个聚簇中心，从 n 个数据对象任意选择 k 个对象作为初始聚类中心；

（2） 循环执行步骤 3 和 4，直到每个聚簇内的元素不再发生变化为止。

（3） 依照当前的聚簇中心，依次计算出每个对象向量与各个聚簇中心的距离，并按照式（4.4）的约束，对每个对象进行重新划分，并更改矩阵 U。

（4） 重新计算聚簇中心。

K-均值算法聚类算法简单而且有效，一旦迭代过程的结果趋于稳定，训练集中的每个元素都被分配到离它最近的聚类中心，最终是将所有点到它们各自聚簇中心的距离平方和最小化。可以看出，这样处理的结果得到的是局部最小值，并

不是全局最优解，这也是所有聚类算法的弊端。事实上，通常找不到全局最优的聚类方法，这是当今聚类技术的现状。但是我们可以不断改进算法，使之无限的接近最优解。

4.1.2 基于概率的模糊聚类算法

上面一节介绍了 K-平均算法的基本思想，本小节将在其基础上提出一个新的聚类方法—基于概率的聚类。

4.1.2.1 模糊集基本概念

首先说明隶属度函数的概念。隶属度函数是表示一个对象 x 隶属于集合 A 的程度的函数，通常记做μA(x)，其自变量范围是所有可能属于集合 A 的对象（即集合 A 所在空间中的所有点），取值范围是[0,1]，即 0<=μA(x)<=1。μA(x)=1 表示 x 完全隶属于集合 A，相当于传统集合概念上的 x∈A。一个定义在空间 X={x} 上的隶属度函数就定义了一个模糊集合 A，或者叫定义在论域 X={x}上的模糊子集 $\underset{\sim}{A}$[5][6][7]。对于有限个对象 x1，x2，……，xn 模糊集合 $\underset{\sim}{A}$ 可以表示为：

$$\underset{\sim}{A} = \{(\mu_A(x_i), x_i) \mid x_i \in X\} \tag{4.8}$$

有了模糊集合的概念，一个元素隶属于模糊集合就不是硬性的了，在聚类的问题中，可以把聚类生成的簇看成模糊集合，因此，每个样本点隶属于簇的隶属度就是[0，1]区间里面的值。

4.1.2.2 概率模糊聚类算法

我们知道，聚类的目的是把具有相似性元素聚合到一起，没有相似关系的元素分离，从而达到聚簇内部相似度极高，而簇与簇之间相似度极低。但实际中并没有绝对结合与绝对离合的概念。因此采用概率来表示元素隶属于哪个类别将更加确切。与此同时，相对应的矩阵 U 允许有取值在 0，1 间的元素，并且照常满足式（4.5）。价值函数也变为以下形式：

$$J(U,c_1,\ldots,c_c) = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c}\sum_{j}^{n} u_{ij}^m d_{ij}^2 , \tag{4.9}$$

这里 u_{ij} 介于 0，1 间；c_i 为模糊组 I 的聚类中心，$d_{ij}=\|c_i-x_j\|$为第 I 个聚类中心与第 j 个数据点间的欧几里德距离；且 $m \in [1,\infty)$是一个加权指数。

构造如下新的目标函数，可求得使（4.10）式达到最小值的必要条件：

$$\begin{aligned}\overline{J}(U,c_1,...,c_c,\lambda_1,...,\lambda_n) &= J(U,c_1,...,c_c)+\sum_{j=1}^{n}\lambda_j(\sum_{i=1}^{c}u_{ij}-1)\\ &=\sum_{i=1}^{c}\sum_{j}^{n}u_{ij}^{m}d_{ij}^{2}+\sum_{j=1}^{n}\lambda_j(\sum_{i=1}^{c}u_{ij}-1)\end{aligned} \tag{4.10}$$

这里λ_j，j=1 到 n，是（6.9）式的 n 个约束式的拉格朗日乘子。对所有输入参量求导，使式（6.10）达到最小的必要条件为：

$$c_i=\frac{\sum_{j=1}^{n}u_{ij}^{m}x_j}{\sum_{j=1}^{n}u_{ij}^{m}} \tag{4.11}$$

和

$$u_{ij}=\frac{1}{\sum_{k=1}^{c}\left(\frac{d_{ij}}{d_{kj}}\right)^{2/(m-1)}} \tag{4.12}$$

由上述两个必要条件，模糊 C 均值聚类算法是一个简单的迭代过程。具体迭代方式类似于 K-均值算法，如下：

（1）用值在 0，1 间的随机数初始化隶属矩阵 U，使其满足式（4.5）中的约束条件。

（2）用式（4.11）计算 k 个聚类中心 c_i。

（3）计算价值函数，如果它小于某个确定的阀值，或它相对上次价值函数值的改变量小于某个阀值，则算法停止，否则执行下一步。

（4）依照式（4.12），重新计算出每个对象向量隶属于各个聚簇的概率，更改矩阵 U，返回第二步。

从上可以看出，K-均值算法是对数值的划分是一种硬性的，在一些场合并不合适，而概率模糊聚类算法是一种软分类方法，使我们能够根据实际情况灵活掌握元素的隶属关系从而更能够体现聚类价值。

4.1.3 基于划分的聚类算法的改进

4.1.3.1 基于划分聚类算法的缺点和局限性

通过以上的学习我们可以知道，不管是 K-平均算法还是模糊聚类算法，都存在以下的弊端和不足，需要改进：

首先，算法中的聚类数 k 必须是事先给定的确定值,即初始值。然而实际中

很难精确确定,因而无法解决该核算法的实际问题。一种解决方案是对不同的可能个数进行实验,即找出那个能使所有点离开它们聚类中心的距离平方和总和达到最小的 k 值。可以从一个给定的最小数开始,例如 k=1,一直实验到一个满足客户需要的最大值。我们知道如果用距离平方总和来衡量 k,那么将选择和数据元素一样多的聚类,由此我们要抑制聚类数目过多造成聚类无意义的结果,仔细斟酌 k 的取值范围[15]。

其次,聚类初始中心点的确定对整个聚类分析起着决定性作用,中心点选的好坏直接影响聚类的结果以及价值函数的大小。如果随机的选取聚类的中心,盲目性过大,将导致偏离实际的结果。例如,对平行四边形而言,很明显短边上的两个点分别形成两个聚簇。但是倘若选取长边的两个中点作为聚簇初始点,很有可能的结果是两个聚类中的每一个都将拥有位于长边两端的两个实例。因此,选择合适的初始中心也是均值算法改进的方向。

再次,在聚类的每次迭代过程中,全部元素必须迭代完一次后更新聚类中心,这样容易使后计算元素对各个聚类中心距离计算的误差加大。由于前者所属类别已经改变即相应的聚类中心有所变化,如果后者再按照变化前的中心计算所属类,将必会产生一定的偏差。

最后,基于划分的聚类算法需要多次迭代,每次都要计算每一个实例元素相对于 k 个聚簇中心的距离或是概率分布,显然算法复杂度是很高的。为此我们可以采用树的数据结构来处理。通过对每一个结点计算出它的隶属关系,从而间接确定他的孩子结点的从分配动向,这样就大大降低了程序的复杂度,提高了时间效率。

4.1.3.2 实际应用中的改进策略

针对聚类算法中的弊端,并结合本课题的研究对象—税控数据仓库,围绕研究内容—商场业务服务器和税控服务器的数据差异,将概率模糊聚类算法做一定的改进:

(1) 在实际商场税收中,税控税种税目有六种,也就是说商场是按照六种税率来纳税的,因此很有必要将商品销售情况分为六大类,针对不同税率的类别将从属商品分到相应的类中,从而形成六大聚类的初始状态,然后决定是否值得再将它们分裂。通常在每个聚类变化最大的方向、距离聚类中心一个标准差的地方建立一个新的聚类中心,最后要注意验证新增加的聚类是否合理。

(2) 按照 1 所讲的初始聚类规则,初始值的确定应以每个聚簇内部各元素的销售差额为基准。首先应该计算出聚类子集中各商品的差额平均值,然后求得内部所有商品差额的中心点。注意,由于商品的类别及其价格高低不一,所以不可笼统的求平均值,而应该采用加权系数的方式得到初始聚簇中心。

(3) 在所属关系 u_{ij} 的计算过程中,应结合商品销售的时间维度和所属类别

（税种），增加一个距离修正函数来修正训练集的概率隶属度，尽可能的将商品归到所属税种的一类。例如，计算欧几里德距离时，通过增加修正系数来提高商品与所属税率之间的隶属度。

（4）在聚类的每次迭代过程中，摒弃传统的全部元素迭代完一次后更新聚类中心的方法而是采用时时更新聚类中心的算法，即对每个元素的归属判定后更新聚类中心，从而降低了后计算元素对各个聚类中心距离计算的误差，减少了系统迭代次数，并使聚类的效果得到了改善，提高了系统效率。

4.2 可视化数据挖掘

数据可视化与数据挖掘技术虽说是两种不同的技术，但如果想创建和应用成功的商业智能企业解决方案必须将两者紧密的结合起来，通过创建二维或三维立体数据分析图，发现潜在的、未知的趋势和模式，使用户易于决策，提高预测和洞察力。

4.2.1 数据可视化

近年来，随着数据仓库技术、网络技术、电子商务技术等的发展，可视化技术涵盖了更广泛的内容，并进一步提出了数据可视化的概念。所谓数据可视化（Data Visualization）是对大型数据库或数据仓库中的数据的可视化，它是可视化技术在非空间数据领域的应用，使人们不再局限于通过关系数据表来观察和分析数据信息，还能以更直观的方式看到数据及其结构关系。数据可视化技术的基本思想是将数据库中每一个数据项作为单个图元元素表示，大量的数据集构成数据图像，同时将数据的各个属性值以多维数据的形式表示，可以从不同的维度观察数据，从而对数据进行更深入的观察和分析[34]。

数据可视化技术包含以下几个基本概念：

（1）数据空间：是由 n 维属性和 m 个元素组成的数据集所构成的多维信息空间；

（2）数据开发：是指利用一定的算法和工具对数据进行定量的推演和计算；

（3）数据分析：指对多维数据进行切片、块、旋转等动作剖析数据，从而能多角度多侧面观察数据；

（4）数据可视化：是指将大型数据集中的数据以图形图像形式表示，并利用数据分析和开发工具发现其中未知信息的处理过程。

目前数据可视化已经提出了许多方法，这些方法根据其可视化的原理不同可以划分为基于几何的技术、面向像素技术、基于图标的技术、基于层次的技术、

基于图像的技术和分布式技术等等。

4.2.2 企业可视化数据集

大多数企业的业务数据都是以二维表的形式组织其内部结构的，本课题也不例外。通过之前建立的数据仓库，训练集已准备就绪，它反映的是业务数据与税控数据的对比关系，从中找出两者之间的差距，以图形图表的形式呈现给客户，最终为税务部门纳税评估提供有力的科学依据[32]。以四月份的月销售为例，六种税率商品汇总的柱形图可以模拟创建如下图 4-1：

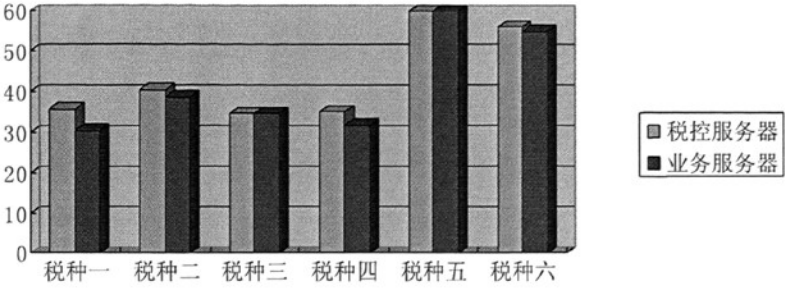

图 4-1 商品销售数据汇总柱形图

Figure 4-1 column chart of summary sales data

针对不同的业务需求，我们可以采用不同的可视化工具。多维的可视化工具能够让客户非常直观的在坐标系中比较单个数据维和其他数据维之间的关系。轮廓图、直方图、条形图、点图等都是比较直观的数据分析模式，实际应用中需示具体情况来采用不同的视图。针对本课题的聚类算法研究，拟采用散点图的形式进行处理与分析，见图 4-2：

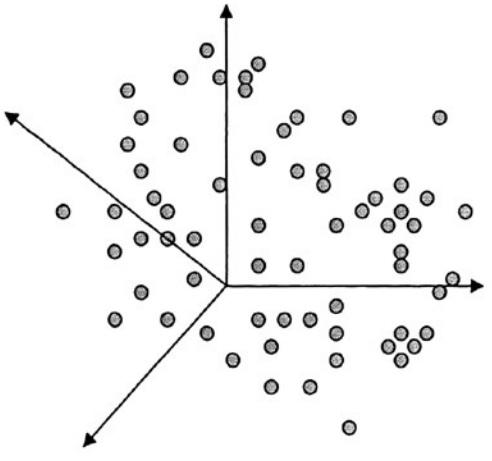

图 4-2 N 维散点图

Fogure 4-2 Scatter chart

4.2.3 可视化数据挖掘工具

利用可视化的数据挖掘工具能够创建多维图表，辅助客户制定决策。伴随越来越多的软件供应商加入数据挖掘这一行列，使得现有的挖掘工具的性能得到进一步的增强，使用更加便捷，也使得其价格门槛迅速降低，为应用的普及带来了可能。当今比较著名挖掘工具有 IBM Intelligent Miner、SAS Enterprise Miner、SPSS Clementine 等。

（1）Intelligent Miner

由美国 IBM 公司开发的数据挖掘软件 Intelligent Miner 是一种分别面向数据库和文本信息进行数据挖掘的软件系列，采用了多种统计方法和挖掘算法，主要有：单变量曲线，双变量统计，线性回归，因子分析，主变量分析，分类，相似序列，序列模式，预测等。在商业智能中得到了良好的应用和推广，缺陷是与 DB2 连接的限制问题。

（2）Enterprise Miner

这是 SAS 公司开发出的一个全功能、易于使用、可靠和易于管理的系统。在我国的企业中得到采用的数据挖掘工具。作为一种通用的数据挖掘工具，按照“抽样--探索--转换--建模--评估”的方法进行数据挖掘，适合于企业在数据挖掘方面的应用以及 CBM 的全部决策支持应用。

（3）SPSS Clementine

SPSS Clementine 是 Spss 公司收购 ISL 获得的数据挖掘工具。它提供了多种图形化技术，有助理解数据间的关键性联系，指导用户以最便捷的途径找到问题的最终解决办法。

以上介绍的都是一些商业化的挖掘软件，在实际应用中非常广泛。而 Weka 作为一种开源的数据挖掘软件，蕴含了大量的机器学习相关算法，对我们研究和学习算法有重要的知道和帮助。

4.3 分析平台的建立

要进行数据挖掘和分析必须有一个良好的平台或者工具。而 Weka 正是名气最大的开源机器学习和数据挖掘软件。用户可以通过 Java 编程和命令行来调用其分析组件。同时 Weka 也为普通用户提供了图形化界面以及用于修改和改进其平台算法的接口[35]。

4.3.1 Weka Explorer 挖掘

在 Weka 主界面的 Applications 中（下图 4-3），列出了其开发的主要应用程序，Weka Explorer（Weka 探索数据的环境），Weka Experimenter（运行算法试验、统计检验的环境）和 Weka Knowledge Flow（支持增量学习的改进 Explorer 环境），本节重点介绍探索开发数据的环境[40]。

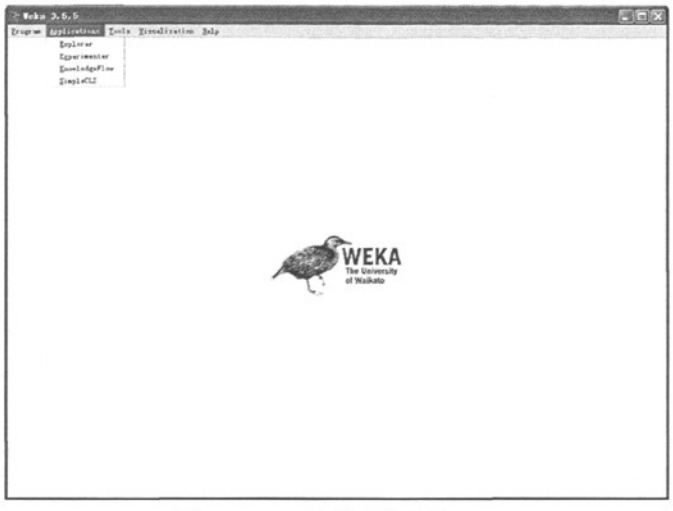

图 4-3 Weka 集成开发环境

Figure 4-3 Weka Integrated Development Environment

当 Explorer 首次启动时，只有第一个标签页是活动的；其他均是灰色的（下图）。这是因为在探索数据之前，必须先打开一个数据集(可能还要对它进行预处理)。由于本课题的训练集之前已经完成了数据的转移、抽取、清洗等预处理过程，因此将不在详细讲解其处理步骤，感兴趣的读者可以自行阅读学习。

我们通过点击 Open file，打开挖掘训练集进行聚类分析。注意 Weka 所能分析的文件数据的扩展名都是以 arff 的命名的，见下图 4-4：

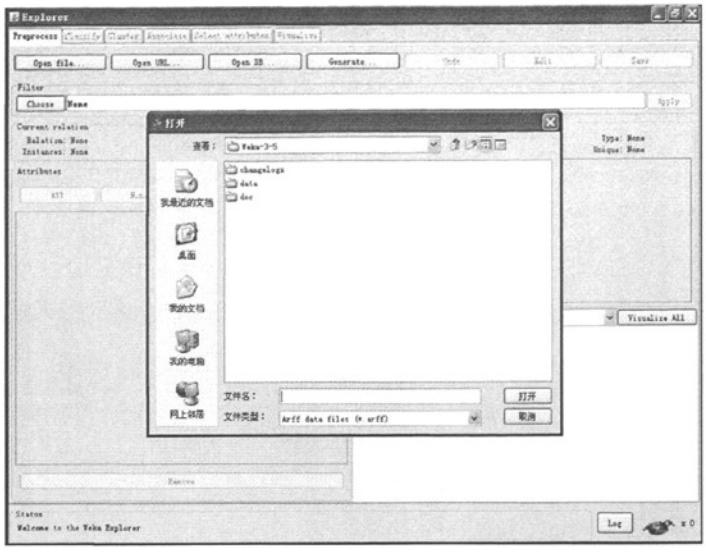

图 4-4 数据加载

Figure 4-4 load data

接着我们以天气数据集为例，进行典型的 K-均值聚类分析，并查看聚类分析的结果：

（1）点击 choose 按钮，选择 SimpleKMeans 算法，见下图 4-5：

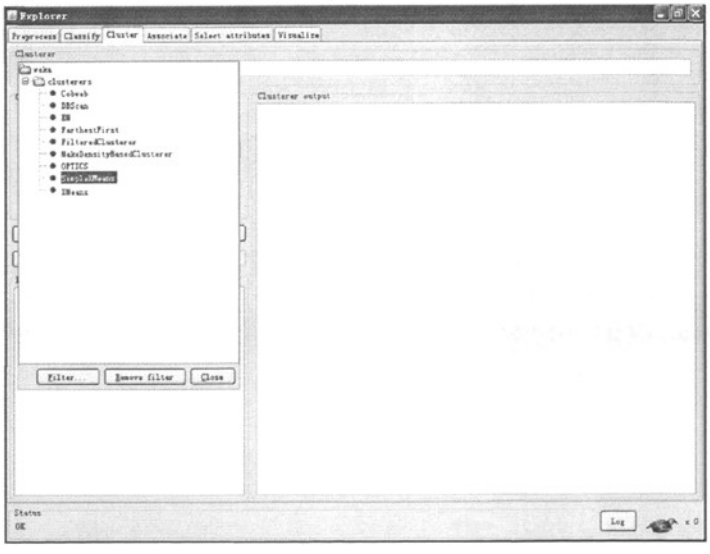

图 4-5 聚类算法的选择

Figure 4-5 selection of clustering algorithm

（2）点击 Start,聚类的结果便在后侧 output 窗口中显示出来，见下图 4-6：

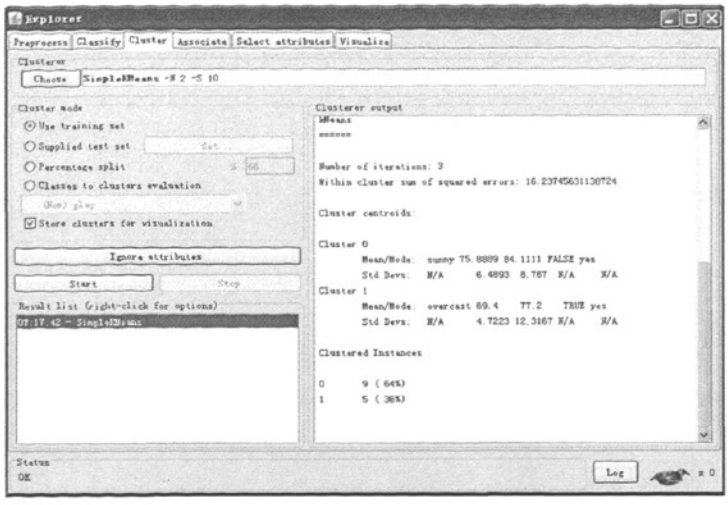

图 4-6 挖掘结果输出

Figure 4-6 output of mining result

4.3.2 在 Weka 中加入自己的算法

Weka 开源的好处就在于我们可以向其加入自己的机器学习算法，改善运行效率。现将我们改进的基于划分的聚类算法加入其中。

首先，编译源文件（AdvancedKMeans.java），生成相应的.class 文件。

接着，为新算法建立合理的工作目录。由于 weka 的目录和它的代码的包结构是一致的，因此将算法加入到聚类（Cluster）目录中理所当然。

然后，修改 weka 启动时加载的系统配置文件，即 weka.gui.GenericPropertiesCreator.props 文档，由于新添加的算法属于聚类所以此处不用修改，因为包 weka.clusterers 已经存在。若加入新的包时则必须修改这里，添加新的包的完整路径。

最后，再修改 weka.gui.GenericObjectEditor.props 配置文件，因为 Weka 中可选择的具体算法都需要在 weka.gui.GenericPropertiesCreator.props 文件中配置，新添加算法的也不例外。在#Lists the Clusterers I want to choose from 的 weka.clusterers.Clusterer=\下加入：weka.clusterers. AdvancedKMeans。

4.4 本章小结

本章首先从算法研究入手，详细讨论了均值聚类算法，提出了其存在的一些不足之处并进行了一定的改进，使其符合税控挖掘的要求。接着介绍了一种开源的可视化数据挖掘平台—Weka。最后，并将自己改进的聚类算法加入其中，从而构造了一个良好可行的数据挖掘平台。

第 5 章 挖掘系统实现和决策分析

5.1 系统设计

本课题的研究和工作平台都是 Weka。Weka 是基于 Java 语言开发的开源软件，包含能处理所有的标准数据挖掘问题的方法：分类、聚类、关联规则等。它的主要的使用方式就是将一种学习方法应用于特定的数据集上，分析并输出结果，供客户参考与决策。围绕这个分析平台，本课题的模型设计可以分为以下几个主要模块[11] [13]：

（1）Weka 启动并加载数据库模块

由于 Weka 中没有特定的连接数据库接口，这需要我们用户去专门定制。本课题中，用到了 PostgreSQL 数据仓库，因此需要建立相关的连接。数据加载完毕后还要进行数据的筛选、维度的建立等工作。

（2）基于划分的聚类算法改进模块

本课题紧密围绕 GB18240.7 税控收款机的实施要求，以对比业务数据库和税控服务器采集到的销售数据之间的差距为需求，从中找出里面所含有的一些模式，从而达到实施监控企业销售情况的作用。

采用基于划分的聚类算法，改进了 K-均值算法中聚类数和初始聚簇中心值随机确定的不良因素，并借鉴模糊 C-均值算法，添加了商品单价、税率权重的概念与处理方式，使算法的运行更加高效、合理。

（3）结果分析模块

利用改进的算法，基于模拟的商场销售数据值，以商品 4 月份销售情况为例，展开挖掘分析，并将处理结果反馈给用户。

系统处理的流程如下图 5-1：

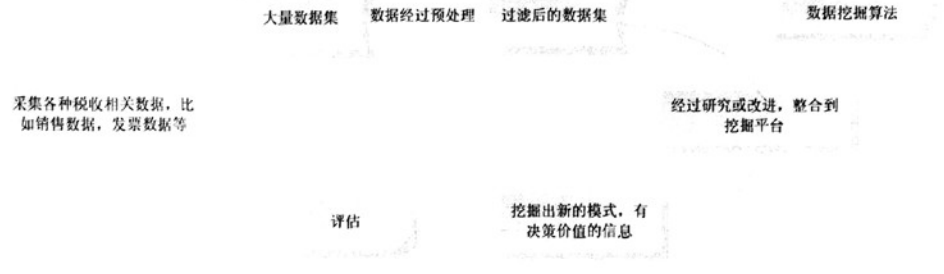

图 5-1 挖掘系统流程图

Figure 5-1 Mining System flow chart

5.2 系统实施

5.2.1 数据加载和维度确定

首先将 Java 连接 postgresql 的 JDBC 驱动 postgresql-8.3-604.jdbc4.jar 添加到项目的 ClassPath 中，然后在连接数据库 URL 中输入 jdbc:postgresql://localhost:5432/DWDB，Username 用 postgres，密码为空，来进行数据库的连接（图 5-2）：

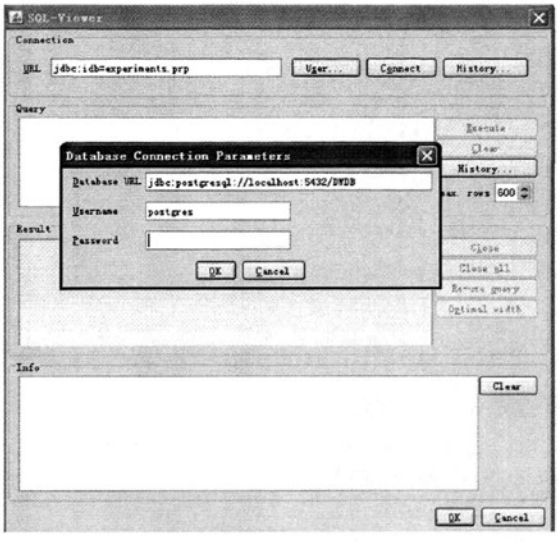

图 5-2 数据库连接

Figure 5-2 connect to database

选择我们要分析的数据仓库表 monthsale 进行数据加载，见下图 5-3：

图 5-3 数据加载

Figure 5-3 load data

在商品销售表中不是所以的字段都对评价税控数据和业务数据的差额有意义。例如，商品的编号与两者的差额就不会有任何联系，另外由于本例是对 4 月的整个销售情况做分析，所以日期也应被舍去。其次，销售差额是由税控服务器的销售总额减去业务服务器上传的销售总额得到的，所以业务服务器销售额也不被纳入到分析维度集合中[8]，具体操作见下图 5-4：

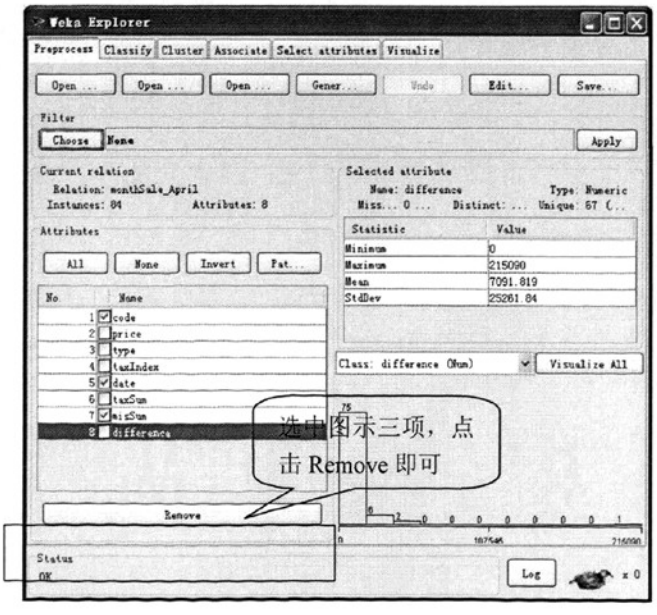

图 5-4 维度确定

Figure 5-4 establishment of dimensions

5.2.2 算法的改进

针对 K-均值算法的部分局限性，并结合本课题税收分析的实际情况，对其进行了如下改进，使其更加符合税收管理的需求。

算法的伪码描述如下：

Input：the number of cluster and dataset

Output：the detail of clusters and make value errors smallest

BEGIN

F：Calculate the initial value for these clusters

S：repeat:

// 根据每个聚簇的中心值，将每个数据元素赋予最类似的簇

a: for i=1 to n(number of dataset), assign each x_i the cluster which has the closest mean and then update the changed center

// 更新簇的平均值

b: for j=1 to k(number of clusters), calculate c_i, the average value for each clusters, $c_i = \frac{1}{|G_i|}\sum_{k, x_k \in G_i} x_k$

// 检查每个簇有无变化

c: check the changes of each cluster;

until there is no changes of these clusters;

L：compute the value errors:

$$J = \sum_{i=1}^{c} Ji = \sum_{i=1}^{c} (\sum_{k, x_k \in Gi} \| x_k - c_i \|^2)$$

END

（1）初始值 k 的选定

算法中 k 必须要进行初始化，否则算法无法进行。当然如果取值偏小或者无限大都会使聚类没有意义。我们可以借鉴税率分类的原则，将数据分为六大类（每一类对应一种税目的商品，当然这也一定是绝对的）。因为毕竟相同税率的商品产生的销售数据具有相似点，将它们归为一类作为初始值，可能会提高聚类的运算效率。K 定义如下：

```
// number of clusters to generate
private int m_NumClusters = 6;
```

（2）初始聚簇中心的设定

初始聚簇中心的选择对聚类起着决定性作用，因为中心选择的好坏直接影响价值

函数的大小，从而影响评估的效果。由于我们已经采用了以税率分类做为初始值的策略，所以应从初始每个类的所以元素中找到它们的中心点。核心代码如下：

```
// initialize clusterer
public void initialize(Instances instances)
{
    Instances[] initClusters = new Instances[m_NumClusters];
    for(int i=0; i<m_NumClusters; i++)
    {
        initClusters[i] = new Instances(instances, 0);
    }
    // classify the instances
    for(int i=0; i<instances.numInstances(); i++)
     {
          initClusters[(int)instances.instance(i).valueSparse(2)].add
          (instances.instance(i));
    }
    double[][] vals = new
            double[m_NumClusters][instances.numAttributes()];
    //  find and record the cluster center
    for(int i=0; i<m_NumClusters; i++)
    {
        for(int j=0; j<instances.numAttributes(); j++)
        {
            vals[i][j] = initClusters[i].meanOrMode(j);
        }
        if(m_ClusterCentroids != null)
        {
            m_ClusterCentroids.add(new Instance(1.0, vals[i]));
        }
    }
}
```

（3）距离权值的计算方法

我们知道聚类的目的是使同一聚簇内部的元素相似度增高，而不同聚簇的元素之间的相似度降低。如何计算找出他们之间的相似度，即距离，是本文的关键。由于商品销售元素集既存在数值型的也有非数值型的属性，因此采用纯数值距离

来考虑两者之间的距离是不可取的，必须将两种属性的联系紧密结合起来综合考虑差异度。为此我们可以采用基于概率的计算方法：

对于非数值型属性，即商品类别和税目，我们采用不同税率的商品差别为0.5，相同税率的商品如果类别不同则差别为 0.25，相同的为 0 的计算方法。部分代码如下：

```
switch (m_ClusterCenters.attribute(index).type()) {
    case Attribute.NOMINAL :
        // If attribute is nominal
        if (Instance.isMissingValue(val1)
                || Instance.isMissingValue(val2)
                || ((int) val1 != (int) val2)) {
            if(m_ClusterCentroids.attribute(index).name() ==
"type"){
                return 0.25;
            }else{
                return 0.5;
            }
        } else {
            return 0;
        }
```

对于数值型属性，首先我们找出该属性上数值的最大值与最小值，并计算他们的差做为绝对距离，将最大值距离最小值的距离（差异度）记为 1。然后分别计算两比较元素在该属性上的取值相对于最小值的差异度，并将两者差异度的差值做为两者之间的差异度或相似度。可以看出，不论两者的属性值如何变化，他们之间的差异度始终在 0 到 1 之间的闭区间上变化。

// val1, val2 分别为两者实际属性值，attributeIndex 为属性索引

```
difference = transferData(val1, attributeIndex) - transferData (val2,
attributeIndex);
```

transferData 函数原型为：

```
private double transferData (double x, int index) {
    return (x - m_Min[index]) / (m_Max[i] - m_Min[index]);
}
```

最后我们可以将各属性差异度之和做为两元素之间的距离来进行聚类分析：

```
for (int i=0; i<instance.numAttributes; i++){
    .....
    .....
```

```
        distance += difference * difference;
    }
```

（4）时时更新聚类中心的算法

在每次迭代过程中，对每个元素的归属判定后更新聚类中心，从而降低了后计算元素对各个聚类中心距离计算的误差，减少了系统迭代次数，并使聚类的效果得到了改善：

```
    int emptyClusterCount=0;
    Instances[] tempClusters = new Instances[m_NumClusters];
    m_ClusterCentroids = new Instances(instances, m_NumClusters);
    for (int i = 0; i < m_NumClusters; i++) {
        tempClusters [i] = new Instances(instances, 0);
    }
    for (int i = 0; i < instances.numInstances(); i++) {
        tempClusters[ClusterAssignments[i]].add
                                        (instances.instance(i)); }
    for (int i = 0; i < m_NumClusters; i++) {
        double[] vals = new double[instances.numAttributes()];
        if (tempClusters [i].numInstances() == 0) {
            emptyClusterCount++;
        } else {
            for (int j = 0; j < instances.numAttributes(); j++) {
                vals[j] = tempClusters[i].meanOrMode(j);
                m_ClusterNominalCounts[i][j] = tempI[i]
                        .attributeStats(j).nominalCounts;
            }
            m_ClusterCentroids.add(new Instance(1.0, vals));
        }
    }
```

5.2.3 结果评估与决策分析

针对数据集成后建立的 886 条训练实例集，并利用上节建立的聚类算法我们可以迅速对模拟训练集建立聚类分析模型。如图，单击 Choose，并选择 AdvancedKMeans 方法。点击 Start，即可得出聚类的结果模型。下图 5-5 是改进的 K-均值算法的运行结果：

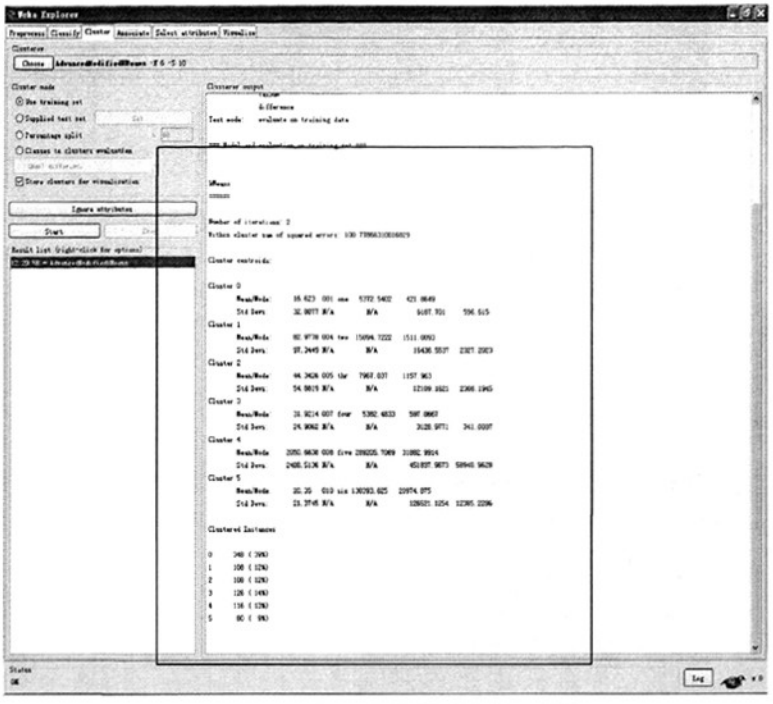

图 5-5 改进的 K-均值算法的输出结果

Figure 5-5 output of improved k-means algorithm

而采用一般 K-均值算法的输出结果为下图 5-6：

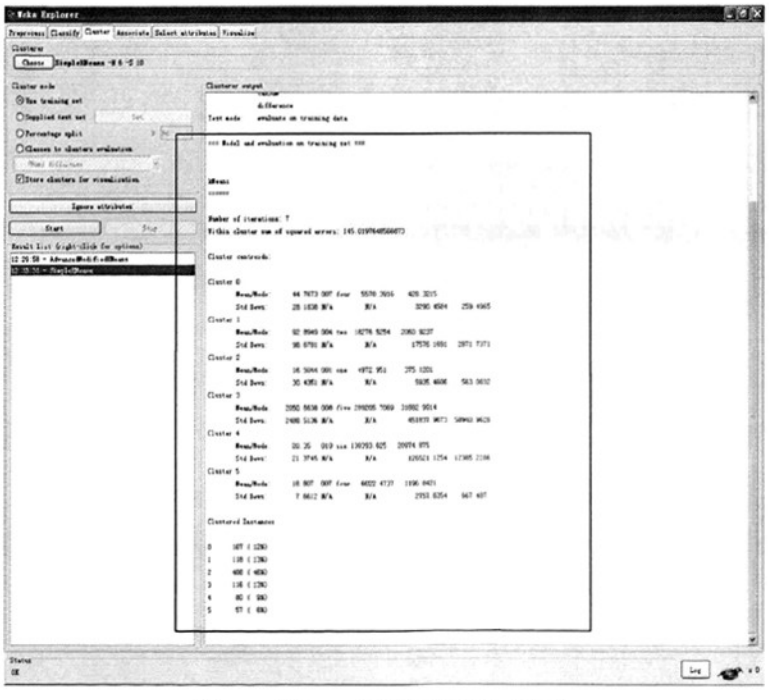

图 5-6 传统 K-均值算法的输出结果

Figure 5-6 output of Traditional k-means algorithm

从图中可以看出，应用初始聚类中心初始化和隶属度概率算法的循环迭代次数和相似性价值函数分别为 2 和 100.7787，所有的记录被分配到以税率划分的类中；而随机选择初始值聚类算法的迭代次数和相似度分别为 7 和 145.0198。并且两者对训练集的聚类结果（元素和聚簇的隶属关系）也有一定的差别。可以看出前者在一定程度上优于后者。用户可以根据每个聚类中心的各项指标了解两个服务器之间的差异度，并可以对各属性之间的关联关系，以及对差异度的影响建立一种概念模型，来进行决策制定与分析[9] [14] [36]。

5.3 本章小结

本章具体设计实现了数据挖掘的简单模型。通过将税控数据仓库数据的一系列处理，找出数据的相似度并进行聚类拟合，分析了数据维度对数据聚类的影响及相互关系，帮助税务部门进行合理的纳税评估和决策分析。本章主要从具体的算法入手，通过对原有聚类算法的部分改进，使之更加高效且更贴近当前的税控挖掘目标。

结论

随着数据库和数据挖掘技术的发展，与之相对应的应用愈演愈烈。本文以数数据挖掘技术为根本出发点，研究讨论了其在当今商业中的广泛应用，并与商业智能紧密结合起来，从聚类算法的角度入手，详细讨论了数据挖掘技术中聚类分析的一些方法，并结合国标 18240.7 商场自动化收款机系统，基于多种数据源建立了良好的分析模型，为税务稽查人员纳税评估提供了积极有力的决策依据。

纵观全文，主要研究解决了以下几个问题：

（1）数据挖掘技术的发展以及当今的应用。数据挖掘对我们来讲已经不是一个陌生的概念，已经走过了足足二十个春秋，逐渐演变为当今的数据分析处理、商业智能等技术。

（2）数据仓库技术的研究与应用。针对数据源庞大、分散且异构的问题，我们采取了异构数据库集成的方案，利用 XML 中间件技术，将分散在多台服务器上的销售数据汇总到统一的数据库中，并构建了相应的数据仓库，为后来的挖掘工作做好了准备。

（3）聚类算法的改进与演变。算法的研究是本文的关键。传统的 K-均值算法不能满足税控数据分析的要求。结合数据仓库的特点和分析目标，选取合适的初始点作为聚簇中心，并将基于距离的聚类算法和基于概率的算法结合起来形成了一种新的基于距离权值的聚类方法，一定程度上提高了系统的效率，并使聚类的效果相比以前有了较大提高

（4）挖掘平台的选择搭建问题。本文以研究讨论建立模型为目的，由此选择一种开源的挖掘平台尤为重要。Weka 作为一种集成了大量机器学习算法的开源平台是我们的不错选择。将改进的聚类算法依据它我们提供良好的应用接口嵌入其中，方便的进行数据分析。

数据挖掘的应用和算法分析涉及方方面面，本文由于时间所限没有详细到每一点，只是讨论了其中的技术，尚存在以下需要继续研究改进的地方：

（1）当数据量增大时，算法的复杂度偏高，系统运行缓慢。我们可以设计一种树的结构予以解决。树的每一点对应数据集合的每一个元素。从树根到树叶自顶向下形成行成聚类。在分析树中某一点的隶属关系时，也就间接的处理了其子树的从属类别。

（2）统计、分析的处理流程、界面有待丰富。鉴于本文的研究点，并没有在界面和报表处理上耗费过多精力。在后续的研究工作中可以添加表格汇总、模式分类等业务综合，使商业智能的价值充分体现。

（3）随着当今 Web service 技术的发展，智能客户端的应用已经开始逐渐推

广。针对税收数据分析，我们可以利用其建立一个智能分析平台，将面向服务的XML Web Service技术与智能客户端结合起来构建优秀的客户定制化解决方案。

参考文献

1 Armando Fox, Steven D. Gribble, Yatin Chawathe, Eric A. Brewer, Paul Gauthier.Cluster-Based Scalable Network Services. Symposium on Operating Systems Principles.1997

2 Data Mining: An Introduction. http://databases.about.com/od/datamining/a/datamining.htm

3 Neena Buck. Eureka! Knowledge Discovery. Software Magazine. December 2000/January 2001 cover story.

4 Knowledge Discovery In Databases: Tools and Techniques. http://www.acm.org/crossroads/xrds5-2/kdd.html

5 R. Vilalta, T. Stepinski, M. Achari .An efficient approach to external cluster assessment with an application to martian topography . Data Mining and Knowledge Discovery. 2007.1

6 Ian Davidson, S. S. Ravi . The complexity of non-hierarchical clustering with instance and cluster level constraints . Data Mining and Knowledge Discovery. 2007.1

7 J. Pei, J. Han, and R. Mao. CLOSET: An efficient algorithm for mining frequent closed itemsets. In ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, pages 21--30, 2000

8 刘明吉，王秀峰，黄亚楼.数据挖掘中的数据预处理.计算机科学，2000，27(4):54-57

9 杨晓超，朱兴友.纳税评估的特点和作用.税收理论．2002.10

10 杨光，张雷，艾波.OLAP 技术及其发展.计算机应用与研究.1999(7):7-10

11 Ed Wilson . The Knowledge Discovery Process , A Problem SoldngMethodology[M].Computer Associates International.Inc. 1998.

12 Ian H witten,EibeFrank.数据挖掘:实用及其学习技术及 Java 实现[M].北京:机械工业出版社.2003.9

13 数据挖掘管理系统规范说明. http://www.dmgroup.org.cn/ppt19.ppt

14 杨萌藜.基于数据挖掘的纳税评估系统的研究与实现.南京航空航天大学硕士论文.2007.1

15 John Shafer, Rakesh Agrawal, Manish Mehta. SPRINT: A Scalable Parallel Classifier for Data Mining. Proc. 22nd Int. Conf. Very Large Databases, VLDB.1996

16 毛国军.数据挖掘的概念、系统结构和方法.计算机工程与设计.2002，23(8):13-17

17 Ian H. Witten, Eibe Frank. Data Mining Practical Machine Learning Tools and Techniques. 机械工业出版社.2006.2

18 王珏，周志华，周傲英.机器学习及其应用.清华大学出版社.2006.3

19 周志华.机器学习与数据挖掘.南京大学计算机软件新技术国家重点实验室，南京 210093

20 Rakesh Agrawal, Manish Mehta, John Shafer, Ramakrishnan Srikant, Andreas Arning, Toni Bollinger. The Quest Data Mining System. Proc. 2nd Int. Conf. Knowledge Discovery and Data Mining, KDD. 1996

21 Ramakrishnan Srikant, Rakesh Agrawal. Mining Generalized Association Rules. Future Generation Computer Systems.1995

22 Adriaans, P., and D. Zantige. 1996. Data mining. Harlow, England: Addison-Wesley

23 Bergadano, F., and D. Gunetti. 1996. Inductive logic programming: From machine learning to software engineering. Cambridge, MA: MIT Press

24 Berry, M. J. A., and G. Linoff. 1997. Data mining techniques for marketing, sales, and customer support. New York: John Wiley

25 谭旭,王丽珍,卓明.利用决策树发掘分类规则的算法研究.云南大学学报.2000,22(6):415-419

26 毛国君,段立娟,王实,石云.数据挖掘原理与应用.清华大学出版社. 2005.7

27 税控收款机 第7部分:商业自动化管理

28 Moshkovieh, H.M., Mechitov, A.I., &Olson. Rule induction in data mining:effect of ordinal scales. Expert System with Application. 2002

29 税收分析是税收管理的眼睛.
http://www.chinatax.gov.cn/n480462/n480498/n480707/n483158/n572330/n572567/5047643.html

30 人性化打造税收分析系统.
http://www.cweek.com.cn/2006/0424/234712.shtml

31 Ramakrishnan Srikant, Quoc Vu, Rakesh Agrawal. Mining Association Rules with Item Constraints.Proc. 3rd Int. Conf. Knowledge Discovery and Data Mining, KDD

32 Ming-Syan Chen, Jiawei Han, Philip S.Yu Data Mining: An Overview from a Database Perspective. Ieee Trans. On Knowledge And Data Engineering.1996

33 Bing Liu. Integrating Classification and Association Rule Mining. Knowledge Discovery and Data Mining.1998

34 Tom Soukup Ian Davidson. Visual Data Mining. PUBLISHING HOUSE OF ELECTRONICS INDUSTRY. 2004.1

35 Ian H.Witten Eibe Frank. Data Mining, Practical Machine Learning Tools and Techniques. China Machine Press. 2006.2

36 王韬,许评.基于数据挖掘的税收决策支持系统设计.管理科学.2003,16(1)

37 马宏鹏,赵新,李明.数据仓库原型系统设计[J] .计算机工程与应用.2000,(11):109-111.

38 数据仓库.
http://baike.baidu.com/view/19711.htm

39 绍库普,戴维森,朱建秋,蔡伟杰.可视化数据挖掘.电子工业出版社.2004

40 WEKA Software..
http://www.cs.waikato.ac.nz/ml/weka/

41 侯要红,栗松涛. Java XML 应用程序设计.北京:机械工业出版社.2007.9

42 Alexander R. Dunegan. A metadata approach to managing XML in relational database. Emory University.2002

攻读硕士学位期间发表的学术论文

1 王志亮,于书举. 国标 18240.7 异构数据库集成技术的研究. 计算机应用研究, 2010 年

致谢

研究生三年短暂的时光马上就要过去，在此向我的导师于书举老师和许向众老师表示最诚挚的感谢！感谢你们对我的教导，为我创造的良好的科研环境和氛围，让我得到了锻炼，各方面素质得到了提高，感谢你们在学业上孜孜不倦的教导以及在生活上无微不至的关怀，这些都让我获益终身。在我完成本论文的过程中，两位老师提出了宝贵的意见和指导，再次向你们表示感谢！

除此之外，我要感谢我的舍友们。尽管他们在研究生期间与我不在一起工作，但在生活上却给了我极大的帮助，帮助我战胜一切困难并不时地给我思想教导和精神动力，使我在失败和情绪低落时重新找回了必胜的信心。真诚的感谢你们！

再次，我还要感谢三年来一起学习工作的同学们，特别是是实验室的各位同学，感谢你们对我的帮助与支持，你们身上的优秀品质将激励我在今后的学习和工作中更加的努力。

最后，我要感谢我的父母，亲人和所有朋友们，正是你们无时无刻地默默地给我鼓劲，鼓舞我奋勇向前，才使我取得了今天的成果，谢谢你们！