基于分类挖掘的企业定单分析系统

Enterprise purchase request analysis system based on Data Classification Mining

崔玉华　（中科院研究生院 北京 100039）
李俊杰　（神华包头煤化工公司）
刘国华　王东方　（燕山大学）

摘要：企业订单分析作为 ERP 实施过程中的一个关键环节，为采购管理部门提供自动化的决策依据，对提高企业采购环节的经济效益具有关键作用，利用数据挖掘技术，开发企业采购订单分析系统，帮助企业管理人员及采购人员对采购计划做出适当的调整，在保证采购物料满足需求的同时获得更多的流动资金，降低库存成本，提高企业经济效益和整体竞争力。

关键词：数据挖掘　ERP　MRP　企业订单分析

1　引言

目前市面上流行的订单分析系统还主要是对订单数据的查询、统计和报表以及对供应商情况的分析，其处理方式主要是对指定的数据进行简单的数字处理，而不是对这些数据所包含的内在信息进行提取。随着采购决策所面对的数据量在不断增长，人工去整理和理解如此大的数据源已经存在效率、准确性等问题，采购管理部门希望能够提供更高层次的数据分析功能，自动和智能地将待处理的数据转化为有用的信息和知识，为此，我们利用数据挖掘技术，开发了企业采购订单分析系统，主要目标是利用数据挖掘技术对企业尚未发出的采购订单进行分类分析，将其分为必需提前的、可延迟的、可撤销的三类。从而帮助企业管理人员及采购人员对采购计划做出适当的调整，避免因采购物料缺乏而影响企业的正常运转，造成不必要的损失；同时也可以通过撤销不必要的采购订单或者推迟某些订单而降低库存量并获得更多的流动资金，降低库存成本，提高企业的资金使用效率，使企业在物资采购环节获得更高的经济效益。

2　分类数据挖掘方法

数据挖掘（Data Mining）是一个多学科交叉研究领域，它融合了数据库（（Database）技术、人工智能（Artificial Intelligence）、机器学习（Machine Learning）、统计学（Statistics）、知识工程（Knowledge Engineering）、面向对象方法（Object – Oriented Method）、信息检索 Information Retreval）、高性能计算（High – Performance Computing）以及数据可视化（Data Visualization）等最新技术的研究成果。经过十几年的研究，产生了许多新概念和新方法。

随着数据库容量的膨胀，特别是数据仓库（Data Warehouse）以及 Web 等新型数据源的日益普及，联机分析处理（On – Line Analytic Processing，OLAP）、决策支持（Decision Support）以及分类（Classification）、聚类（Clustering）等复杂应用成为必然。数据挖掘和知识发现使数据处理技术进入了一个更高级的阶段。它不仅能对过去的数据进行查询，而且能够找出过去数据之间的潜在联系，进行更高层次的分析，以便更好地做出理想的决策、预测未来的发展趋势等。通过数据挖掘，有价值的知识、规则或高层次的信息就能从数据库的相关数据集合中抽取出来，从而使大型数据库作为一个丰富、可靠的资源为知识的提取服务。

分类是数据挖掘中的一项非常重要的目标和任务，目前的研究在商业上应用最多。分类的目的是学会一个分类函数或分类模型（分类器），该模型能把数据库中的数据项映射到给定类别中的某一个类别。要构造分

类器，需要有一个训练样本数据集作为输入。由于数据挖掘是从源数据集中挖掘知识的过程，这种类知识也必须来自于源数据，应该是对源数据的过滤、抽取(抽样)、压缩以及概念提取等。从机器学习的观点，分类技术是一种有指导的学习(Supervised Learning)，即每个训练样本的数据对象已经有类标识，通过学习可以形成表达数据对象与类标识间对应的知识。从这个意义上说，数据挖掘的目标就是根据样本数据形成的类知识并对源数据进行分类、进而也可以预测未来数据的归类。用于分类的类知识可以用分类规则、概念树，也可能以一种学习后的分类网络等形式表示出来。

2.1 分类方法的基本概念和步骤

定义：给定一个数据库 $D=\{t_1,t_2,\cdots,t_n\}$ 和一组类 $C=\{C_1,C_2,\cdots,C_m\}$，分类问题是去确定一个映射 $f:D\rightarrow C$，每个元组 t_i 被分配到一个类中。一个类 C_j 包含映射该类中的所有元组，既 $C_j=\{t_i|f(t_i)=C_j,1\leqslant i\leqslant n$，而且 $ti\in D\}$。

我们把分类看作是从数据库到一组类别的映射。一般地，数据分类(Data Classification)分为两个步骤：建模和使用。

(1) 建立一个模型，描述预定的数据类集或概念集。通过分析由属性描述的数据库元组来构造模型。数据元组也称作样本、实例或对象。为建立模型而被分析的数据元组形成训练数据集。训练数据集中的单个元组称做训练样本，并随机地由样本群选取. 每个训练样本还有一个特定的类标签与之对应。由于提供了每个训练样本的类标号，该步也称做有指导的学习(即模型的学习在被告知每个训练样本属于哪个类的"指导"下进行)。它不同于无指导的学习(或聚类)，那里每个训练样本的类标号是未知的，要学习的类集合或数量也可能事先不知道。

通常，学习模型用分类规则、决策树或等式、不等式、规则式等形式提供。这些规则可以用来为以后的数据样本分类，也能对数据库的内容提供更好的理解。

(2) 使用模型进行分类。首先评估模型(分类法)的预测准确率。保持(Holdout)方法是一种使用类标号样本测试集的简单方法。这些样本随即选取，并独立于训练样本。模型在给定测试集上的准确率是正确被模型分类的测试样本的百分比。对于每个测试样本，将已知的类标号与该样本的学习模型类预测比较。如果模型的准确率根据训练数据集评估，评估可能是乐观的，因为学习模型倾向于过分拟合数据。因此，使用交叉验证法来评估模型比较合理。

如果认为模型的准确率可以接受，就可以用它对类标号未知的数据元组或对象进行分类。总之，我们可以把分类归结为模型建立和使用模型分类进行分类两个步骤，其实模型建立的过程也就是使用训练数据进行学习的过程，第二个步骤是对类标号未知的数据进行分类的过程。

2.2 分类方法的基本类型

分类在数据挖掘中是一项非常重要的任务，一般地，分类方法可归结为四种类型：基于距离的分类方法、决策树分类方法、贝叶斯分类方法和规则归纳方法。

和其他表示方法相比，分类器采用规则形式表达具有易理解性。常见的采用规则表示的分类器构造方法有：

· 利用规则归纳技术直接生成规则；· 利用决策树方法先生成决策树，然后再把决策树转换为规则；· 使用粗糙集方法生成规则；· 使用遗传算法中的分类器技术生成规则等。文章所述订单分析系统主要采用规则归纳算法，故这里重点讨论规则归纳方法。它可以直接学习规则集合，这一点与决策树方法、遗传算法有两点关键的不同。

它们可学习包含变量的一阶规则集合。这一点很重要，因为一阶子句的表达能力比命题规则要强得多。

规则归纳有四种策略：减法、加法、先加后减、先减后加策略。典型的规则归纳算法有 AQ、CN2 和 FOIL 等。这里讨论的算法使用序列覆盖算法。一次学习一个规则，以递增的方式形成最终的规则集合。

3 企业定单分析系统

系统适用对象为具有采购订单管理的各种企业和组织，主要目标是利用数据挖掘技术对企业尚未发出的采购订单进行分类分析，将其分为必需提前的、可延迟的、可撤销的三类。

通过对订单的分类分析，帮助企业管理人员及采购人员对采购计划做出适当的调整，避免因采购物料缺乏而影响企业的正常运转，造成不必要的损失；同时也可以通过撤销不必要的采购订单或者推迟某些订单而降低库存量并获得更多的流动资金，从而降低库存

成本，提高企业的资金使用效率，使企业在物资采购环节获得更高的经济效益。

本平台采用 VC + + 集成开发环境，SQL Server 2000 数据库管理系统，数据库的访问采用目前数据库系统开发中流行的 ADO 技术。可在 Windows 服务器上运行。在系统开发过程中还同时兼顾了系统的通用性、数据库访问的灵活性、系统的可扩展性和数据库连接的安全性以及方便快捷的操作界面等。

为了便于分类，在分类前要对数据进行预处理，即创建数据源，然后利用数据挖掘分类方法进行订单分类分析。因此本系统主要包括创建数据源和订单分析两个模块。

3.1 创建数据源模块

本模块的主要功能是接收用户（系统管理员或具有相应权限的操作人员）填写的关于所要连接和访问的数据库服务器的信息，为了保证系统的安全性，我们对用户输入的信息进行加密，然后存储在特定的文件中。当用户启动订单分析模块进行订单分析时，分析程序首先到加密文件中提取数据库服务器的信息，进行解密后，再进行数据库的连接。

因此，创建数据源模块并没有真正创建系统与数据源的连接，它只是将连接数据源的所需的相关信息经加密后存储在某个文件中，当订单分析程序真正创建数据源，进行数据库的连接时，程序到数据源连接信息文件中去读取。

3.2 订单分类分析数据挖掘模块

3.2.1 模块简介

本模块的主要功能是提取数据库中与订单分析相关的信息（vSQL = " select * from base_table" ;），利用分类数据挖掘算法进行订单的分类分析（ If 可用数量 > = 需求 and 后面近期还有需求 then delay ；If 可用数量 > = 需求 and 后面近期还有采购订单 then cancel；If 可用数量 < 需求 then ahead ），然后将分析结果显示给用户（ vSQLtd = " insert into SHOW values (" + wul + "'," + timm + "'," + danh + "'," + flagX + ")" ; ）。

要进行订单的分析工作，我们需要在数据库中提取以下几类相关数据。

（1）当前库存信息

当前库存信息包括企业生产运行所需的所有物料的当前库存信息，主要包括物料号、物料的采购周期（以天为单位）、物料的当前库存总量、以及物料的最小安全库存量。

（2）企业订单信息

企业订单信息包括企业中所作的所有物料的采购订单信息，主要包括订单的到货日期、订单中所定的物料号、物料的采购数量、采购订单的订单号。

（3）计划需求信息

计划需求信息包括企业为保障正常生产运行所需的所有物料的需求计划信息，主要包括物料的需求日期，所需物料的物料号、物料的需求数量。

注：此计划需求信息表是企业对未来某段时间的物料需求所做的计划，即某一天对某种物料的需求量。

3.2.2 订单分析分类模型

得到上述信息后，就可以建立分类模型对企业的物料采购订单进行分类分析了，分类算法将订单分为可撤销的、可延迟的和必需提前的三类。

在安全库存为 50，采购周期为 15 天的前提下，训练数据及分类规则如下：

（1）可延期的采购订单

在 2007 - 02 - 10 运行，结果如表 1 所示。

表 1　可延期的采购订单

日期	项目	数量	可用数量
2007 - 02 - 10	当前库存	100	50
2007 - 02 - 10	计划需求量	10	40
2007 - 02 - 12	采购订单量	30	70
2007 - 02 - 15	计划需求量	20	50
2007 - 02 - 20	计划需求量	30	20

考虑 2007 - 02 - 12 的采购订单量 30 个和 2007 - 02 - 15 的计划需求量 20 个，在 2007 - 02 - 10 的可用数量能够满足 2007 - 02 - 15 的计划需求，说明 2007 - 02 - 12 的采购订单可用延期收货，即可延期的采购订单。（ If 可用数量 > = 需求 and 后面近期还有需求 then delay ）

（2）可撤销的采购订单

在 2007 - 02 - 10 运行，结果如表 2 所示。

考虑 2007 - 02 - 12 的采购订单量 30 个、2007 - 02 - 23 的采购订单量 20 个和 2007 - 02 - 15 的计划需求量为 20 个、2007 - 02 - 20 的计划需求量 10 个，在 2007 - 02 - 10 的可以数量能够满足 2007 - 02 - 15 的

计划需求量和 2007 -02 -20 的计划需求量之和，说明 2007 -02 -12 的采购订单可以不收货，即此采购订单是可撤销的。(If 可用数量 > =需求 and 后面近期还有采购订单 then cancel)

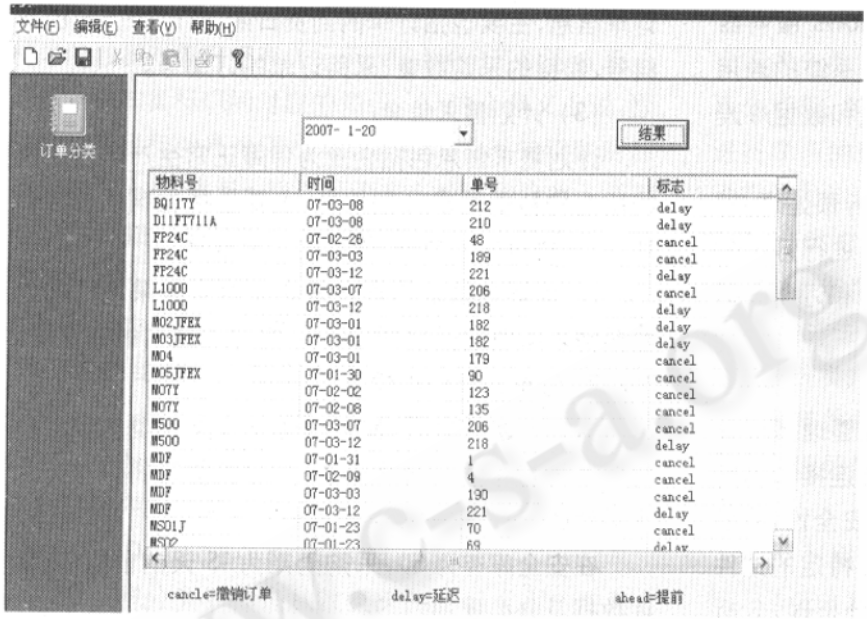

图 1　订单分析结果

(3) 必须提前的采购订单

在 2007 -02 -10 运行，结果如表 3 所示。

表 2　可撤销的采购订单

日期	项目	数量	可用数量
2007 -02 -10	当前库存	100	50
2007 -02 -10	计划需求量	10	40
2007 -02 -12	采购订单量	30	70
2007 -02 -15	计划需求量	20	50
2007 -02 -20	计划需求量	10	40
2007 -02 -23	采购订单量	20	60

表 3　必须提前的采购订单

日期	项目	数量	可用数量	单号/上级物料
2007 -02 -05	当前库存	100	50	
2007 -02 -10	计划需求量	10	40	产品编码
2007 -02 -15	计划订单量	20	20	采购订单号
2007 -02 -20	计划需求量	40	-20	产品编码
2007 -02 -28	采购订单量	30	10	产品编码

考虑 2007 -02 -28 的采购订单量 20 个和 2007 -02 -10 的计划需求量 10 个、2007 -02 -15 的计划需求量 20 个、2007 -02 -20 的计划需求量 40 个，在 2007 -02 -20 的可用数量为短缺 20 个，该物料的采购周期为 15 天，当前日期为 2007 -02 -05，只要在 2007 -02 -20 能够收到货就能满足需求，说明 2007 -02 -28 的采购订单是必须提前的。(If 可用数量 < 需求 then ahead)

3.3　模块应用

当用户想进行订单分析时，运行订单分类分析数据挖掘程序。为了确保分析开始的时间为当前日期，我们在界面上设置了一个日历，其默认显示的是系统中的当前日期。如果系统日期不准确的话，允许用户重新设置日期，程序将从用户设置的时间开始进行订单分析。

确定日历显示的日期为当前日期后，点击“结果”按钮，开始进行分类分析。数据挖掘程序将分类分析结果显示在界面中间的表格中，如图 1 所示。

图中表格显示的数据就是订单的分析结果，每一行代表一条订单信息，包括物料编号、订单到货日期、订单数量和订单分类标志。

“标志”列为“cancle”表示该条订单为可撤销的；

“标志”列为“delay”表示该条订单为可延迟的；

“标志”列为“ahead”表示该条订单为必须提前的。

4　结束语

目前，本订单分析系统已通过用户的实际使用测试，分类分析的准确性及各方面的性能已达到用户要求，使用效果良好，希望能为用户在采购管理方面获得更好的经济效益。

参考文献

1　毛国君等编著，数据挖掘原理与算法，北京：清华大学出版社，2006 年 1 月.

2　刘国华等编著，数据库新理论方法及技术导论，北京：电子工业出版社，2006 年 12 月.

3　M. Berthold, D. J. Hand, Intelligent Data Analysis. An Introduction. Springer - Verlag, Berlin, Heidelberg, New York (1999).