

文章编号:1671-251X(2014)04-0010-05

DOI:10.13272/j.issn.1671-251x.2014.04.003

吴乔,罗键,林金有.煤炭企业生产成本联机分析处理和数据挖掘研究[J].工矿自动化,2014,40(4):10-14.

煤炭企业生产成本联机分析处理和数据挖掘研究

吴乔¹, 罗键¹, 林金有²

(1.厦门大学自动化系,福建 厦门 361005; 2.福建煤电股份有限公司,福建 龙岩 364014)

摘要:针对煤炭企业生产成本分析与控制存在准确性低、决策支持不够等问题,提出了一种生产成本管理与决策支持设计方案。该方案采用数据集市采集整个生产范围内的生产成本,并对生产成本进行实时归集和核算,采用联机分析处理和数据挖掘技术获得隐藏在生产成本内部的规律和趋势,从而为决策者提供决策支持。实际应用表明,该方案实现了生产成本的实时分析与控制。

关键词:煤炭企业; 生产成本; 数据集市; 联机分析处理; 数据挖掘; 决策树

中图分类号:TD67 文献标志码:A 网络出版时间:

网络出版地址:

Research of online analytical processing and data mining for
production cost of coal enterprises

WU Qiao¹, LUO Jian¹, LIN Jinyou²

(1. Department of Automation, Xiamen University, Xiamen 361005, China;

2. Fujian Coal Industry and Electric Power Co., Ltd., Longyan 364014, China)

Abstract: In view of problems of low accuracy and inadequate decision support existed in analysis and control of production cost of coal enterprises, the paper proposed a design scheme of production cost management and decision support. The scheme uses data market to collect the production cost in whole production range and make real-time collection and assignment for the production cost, and it uses technologies of online analytical processing and data mining to get law and trend hided in internal of the production cost, so as to provide decision support for decision-maker. The actual application shows that the scheme realizes real-time analysis and control for the production cost.

Key words: coal enterprise; production cost; data market; online analytical processing; data mining; decision tree

0 引言

煤炭企业的生产成本管理是一个复杂的系统工程,传统的企业生产成本管理运用于煤炭企业会陷入一种分析粒度较粗、准确性和科学性不足的困境。

以数据仓库、联机分析处理(Online Analytical Processing, OLAP)和数据挖掘为核心的智能技术,已被应用到许多领域。参考文献[1]研究了联机分析处理在银行数据管理方面的应用;参考文献[2]

基于数据集市建立了电信经营分析系统模型,为经营提供决策依据;参考文献[3-4]利用联机分析处理进行高考招生和高考志愿数据分析;参考文献[5]建立了焊接车间数据集市,利用数据挖掘技术获得有效的决策支持;参考文献[6]构建了产品族供应链的数据仓库产品,并运用于汽车零部件企业。由于煤炭行业生产的特殊性,联机分析处理和数据挖掘技术在煤炭企业成本管理领域仍鲜有研究。

本文集合数据集市、联机分析处理和数据挖掘

收稿日期:2013-08-18;修回日期:2013-12-25;责任编辑:王晖。

基金项目:福建省产学研合作重大科技项目(2011H6027)。

作者简介:吴乔(1988—),男,湖北天门人,硕士研究生,主要研究方向为现代集成制造系统,E-mail:wuqiao@stu.xmu.edu.cn。

联机分析处理是一种归纳型的决策过程,主要作用是使企业的决策者能灵活地操纵企业的数据,

以多维的形式从多方面和多角度来观察企业的状态,掌握数据中的规律,完成相关的决策^[12]。

对于煤炭企业生产成本,生产部门的管理人员和集团领导关心的不仅仅是在某段时间内的生产成本投入和产出这样笼统的问题,可能更多是不同时间维的投入产出情况、不同部门维的绩效对比情况、不同成本维所占百分比情况和不同生产维的成本投入产出情况等。运用联机分析处理对数据进行多维分析使管理人员能够快速、准确地从多个角度来剖析反映企业生产成本的数据,发现隐藏在背后有价值的信息,辅助决策,并实现对生产管理的支持。

2.4 数据挖掘

数据挖掘是指从大量的数据中发现隐藏于其中的有着特定关系性的、对决策和预测起着重要作用信息的过程^[13]。利用已建立的生产成本数据集市,采用决策树对生产成本数据进行分析。

决策树算法是一种常用的数据挖掘算法,是对现实情况的抽象化,其本质是一种贪心算法^[14]。最有影响的决策树算法是 Quinlan 提出的 ID3 算法和 C4.5 算法。ID3 算法选取信息增益最大的属性来划分训练样本,使分枝时熵值最小。该算法偏向于取值较多的属性,而实际中取值较多的属性往往并不是最优的,这类属性可能不会提供太多有价值的信息^[15]。而 C4.5 算法作为对 ID3 算法的一种改进算法,它采用信息增益率作为选择分枝属性的标准,弥补了 ID3 算法的不足之处。C4.5 算法的基本实现过程如下:

(1) 给定一组训练样本,设 S 为 s 个样本数据集,其具有 m 种不同的类标号属性 $C_i (i=1, 2, \dots, m)$, s_i 为属于 C_i 的样本数。

(2) 对数据源进行预处理,将连续型的属性变量进行离散化形成决策树的训练集(如果没有连续取值的属性,则可以忽略)。

(3) 对样本进行分类,初始的熵值为

$$I(S) = - \sum_{i=1}^m p_i \log_2 p_i$$

式中: p_i 为样本属于 C_i 的概率, $p_i = \frac{s_i}{s}$ 。

(4) 假定属性 A 具有 n 个不同的值 $\{a_1, a_2, \dots, a_n\}$, 可以将样本集划分为 n 个不同的子集 $\{s_1, s_2, \dots, s_n\}$ 。如果选择 A 作为分类属性,根据 A 划分样本集的熵为

$$E(S | A) = \sum_{j=1}^n \frac{s_j}{s} I(s_j)$$

式中: $I(s_j) = - \sum_{i=1}^m p_{ij} \log_2 p_{ij}$, $p_{ij} = \frac{s_{ij}}{s_j}$, s_{ij} 为 s_j 中属于 C_i 的样本数。

属性 A 的信息熵为

$$E(A) = \sum_{j=1}^n \frac{s_j}{s} \log_2 \frac{s_j}{s}$$

属性 A 的信息增益为

$$\text{Gain}(A) = I(S) - E(S | A)$$

属性 A 的信息增益率为

$$\text{Gain_Ratio}(A) = \text{Gain}(A) / E(A)$$

通过计算不同属性下的信息增益率,选取具有最高增益率的属性作为测试属性,创建一个节点,并据此属性划分样本,依次递归,构造决策树。

3 应用实例

笔者对福建煤电股份有限公司生产成本进行管理决策支持,系统采用 B/S 架构,在集成企业现有的生产管理系统、薪资管理系统、物料管理系统和设备管理系统的基础上,对不同子项成本和生产数据进行 ETL 处理后装入到成本数据集市。

客户端作为程序主体部分,负责确定分析和挖掘主体,即定义数据分析和成本挖掘目标及维度属性。根据规则,对样本数据进行分析与挖掘,将分析结果显示到用户界面。

3.1 联机分析处理

结合煤炭企业生产特点,以部门级别、时间跨度、成本类别和生产类别为主要维度,其中对部门的粒度细化到各生产队,对时间粒度细化到各生产月份。联机分析主要包含实际成本与计划成本对比分析、成本各子项构成对比分析及单位产出成本投入对比分析等。

实际成本与计划成本对比分析就是对部门维度上的同一级别部门在不同的时间维和不同环境维下,各个类别成本的实际生产成本与计划生产成本的对比分析,并以图表形式呈现给用户,让管理人员清楚了解企业生产成本实际投入与计划投入的对比和差异情况。具体的煤炭企业实际成本与计划成本对比分析见表 1。

3.2 决策树的应用

运用 C4.5 算法将生产队别和生产环境作为分类属性,将每月生产成本差异作为挖掘目标,对福建煤电股份有限公司下属某个子矿产煤生产成本差异数据进行挖掘。取该子矿 5 个采煤队伍过去 8 年的生产成本差异数据,见表 2。

表 1 煤炭企业实际成本与计划成本对比分析							元
环境维	部门维	3 月份实际生产成本			本季度累计生产成本		
		物料成本		...	物料成本		...
		实际成本	计划成本	...	实际成本	计划成本	...
正常块段	采煤一队	31 240	33 000	...	81 240	82 500	...
	采煤二队	32 560	33 000	...	80 560	82 500	...
	采煤三队	34 250	33 000	...	84 560	82 500	...
	合计	98 050	99 000	...	246 050	247 500	...
回采块段	采煤四队	39 440	38 000	...	98 440	95 000	...
	采煤五队	37 320	38 000	...	97 320	95 000	...
	合计	76 760	76 000	...	195 760	190 000	...
极薄煤层	采煤六队	35 430	33 000	...	85 430	82 500	...
	合计	35 430	33 000	...	85 430	82 500	...

表 2 采煤队伍生产成本差异数据							
差异类别	环境维	部门维					
		采煤一队	采煤二队	采煤三队	采煤四队	采煤五队	合计
有利差异	正常块段	56	30	40	46	20	192
	回采块段	7	7	16	7	20	57
	极薄煤层	6	11	3	4	17	41
	合计	69	48	59	57	57	290
不利差异	正常块段	13	21	6	21	12	73
	回采块段	4	13	24	12	15	68
	极薄煤层	10	14	7	6	12	49
	合计	27	48	37	39	39	190

初始时刻根据每月生产成本差异是否符合要求分为 2 类:一类是有利差异(实际成本小于计划成本);另一类是不利差异(实际成本大于或等于计划成本)。

初始时刻的熵值为

$$I(S) = -\frac{290}{480} \log_2 \frac{290}{480} - \frac{190}{480} \log_2 \frac{190}{480} = 0.968\ 5$$

如果选取生产队别作为测试属性,则条件熵为

$$E(S \mid \text{Team}) = \frac{96}{480} (-\frac{69}{96} \log_2 \frac{69}{96} - \frac{27}{96} \log_2 \frac{27}{96}) + \frac{96}{480} (-\frac{48}{96} \log_2 \frac{48}{96} - \frac{48}{96} \log_2 \frac{48}{96}) + \frac{96}{480} (-\frac{59}{96} \log_2 \frac{59}{96} - \frac{37}{96} \log_2 \frac{37}{96}) + \frac{96}{480} (-\frac{57}{96} \log_2 \frac{57}{96} - \frac{39}{96} \log_2 \frac{39}{96}) + \frac{96}{480} (-\frac{57}{96} \log_2 \frac{57}{96} - \frac{39}{96} \log_2 \frac{39}{96}) = 0.953\ 6$$

生产队别信息熵为

$$E(\text{Team}) = -\frac{96}{480} \log_2 \frac{96}{480} - \frac{96}{480} \log_2 \frac{96}{480} -$$
$$\frac{96}{480} \log_2 \frac{96}{480} - \frac{96}{480} \log_2 \frac{96}{480} = 1.857\ 5$$

如果选取生产环境作为测试属性,则条件熵为

$$E(S \mid \text{Environment}) = \frac{265}{480} (-\frac{192}{265} \log_2 \frac{192}{265} - \frac{73}{265} \log_2 \frac{73}{265}) + \frac{125}{480} (-\frac{57}{125} \log_2 \frac{57}{125} - \frac{68}{125} \log_2 \frac{68}{125}) + \frac{90}{480} (-\frac{41}{90} \log_2 \frac{41}{90} - \frac{49}{90} \log_2 \frac{49}{90}) = 0.914\ 2$$

生产环境信息熵为

$$E(\text{Environment}) = -\frac{265}{480} \log_2 \frac{265}{480} - \frac{125}{480} \log_2 \frac{125}{480} - \frac{90}{480} \log_2 \frac{96}{480} = 1.431\ 5$$

生产队别的信息增益为

$$\text{Gain}(\text{Team}) = 0.968\ 5 - 0.953\ 6 = 0.014\ 9$$

生产环境的信息增益为

$$\text{Gain}(\text{Environment}) = 0.968\ 5 - 0.914\ 2 = 0.054\ 3$$

生产队别的信息增益率为

$$\text{Gain_Ration}(\text{Team}) = \frac{0.014\ 9}{1.857\ 5} = 0.008\ 0$$

生产环境的信息增益率为

$$\text{Gain_Ration}(\text{Environment}) = \frac{0.054\ 3}{1.431\ 5} = 0.037\ 9$$

因为生产环境的信息增益率较大,所以选择生产环境作为决策的根节点。对每一个分支重复上述步骤,生成的决策树如图 3 所示,T₁,T₂,T₃,T₄,T₅ 分别代表不同的采煤队别,Y 表示产煤实际成本小于计划成本,N 表示产煤实际成本超出计划成本。从图 3 可看出,在采煤过程中,煤层状况为影响生产成本的第 1 因素,而生产队别为第 2 因素。其中,生

产环境因素层面上:在正常块段进行生产时,实际成本低于计划成本的比率高;在极薄煤层进行生产时,实际成本低于计划成本的比率适中;在回采块段进行生产时,实际成本低于计划成本比率较低。生产队别因素层面上: T_1 在不同的生产环境下实际成本低于计划成本比率均保持较高水平; T_2 在不同的生产环境下实际成本低于计划成本比率相对较低; T_3 受生产环境因素的影响较大,在正常块段实际成本低于计划成本比率高,在其他生产块段实际成本低于计划成本比率较低; T_4 比较稳定,不易受生产环境因素的影响,但实际成本低于计划成本比率水平不是很高; T_5 在回采块段和极薄煤层体现出较高的实际成本低于计划成本比率水平。

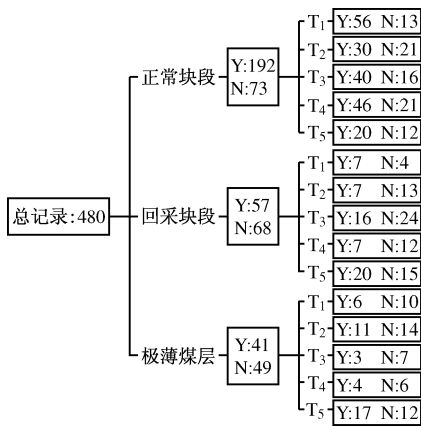


图3 煤炭企业生产成本差异决策树

决策树的分类规则如下:

规则 1:在不同的生产环境下,成本差异率较大。

规则 2:不同的生产队别其生产绩效体现出较大的差异,并且有些生产队别在不同的生产环境下均能保持较高的生产绩效,而有些生产队别受生产环境的变化,绩效变化较大。

对于生产环境因素产生的影响,在生产调度时要根据生产队别表现出来的具体特性进行安排。例如,在回采块段与极薄煤层可优先安排 T_4 进行生产,在正常块段可以优先安排 T_1 和 T_3 进行生产。而对于生产队别因素产生的影响,需要在管理上加强力度。例如,可以总结 T_1 的生产经验,并据此加强对其他队别的管理。同样情况下,可以对生产维中的其他生产方式所投入的生产成本数据进行挖掘,也可以对成本维中各子项生产成本数据进行挖掘得到指导性的意见,以支持决策。

4 结语

从煤炭企业生产成本管理的需要出发,将基于

数据集市的联机分析处理和数据挖掘引入到生产成本管理领域,在不同的维度对生产成本进行了分析与挖掘。通过在煤炭企业的应用实例,细化企业生产成本管理的粒度到各生产队别,验证了联机分析处理和数据挖掘的有效性,为成本管理和生产决策提供了支持,为成本管理提供了有效的工具。

参考文献:

[1] 江键,陈福生. OLAP 在银行数据仓库中的设计和实现[J]. 计算机工程与设计,2006,27(20):3884-3886.

[2] 李艳,杨永健,李树秋. 基于数据集市的电信经营分析系统模型[J]. 山东大学学报:理学版,2007,42(11):62-65.

[3] 何小明,张自力,肖灿,等. 基于联机分析处理与数据挖掘的高考招生数据分析[J]. 计算机科学,2012,39(6):175-178.

[4] 殷员分,张自力,蔡海敏,等. 数据仓库与 OLAP 技术在高考志愿数据分析中的应用[J]. 计算机科学,2010,37(5):162-164.

[5] 李迪,张春华,赖乙宗. 焊接车间数据集市和数据挖掘的研究[J]. 机械工程学报,2003,39(4):79-82.

[6] 王志宏,顾新建. 基于数据仓库的产品族供应链管理系统[J]. 浙江大学学报:工学版,2009,43(7):1197-1202.

[7] 刘威,初延刚,王冠,等. 钢铁企业生产成本控制的体系结构[J]. 东北大学学报:自然科学版,2004,25(5):467-470.

[8] 王珊. 数据仓库与联机分析处理[M]. 北京:科学出版社,1998.

[9] 朱德利. SQL Server 2005 数据挖掘与商业智能完全解决方案[M]. 北京:电子工业出版社,2007.

[10] CHAN K C C, LEE V, LEUNG H. Generating fuzzy rules for target tracking using a steady-state genetic algorithm[J]. IEEE transactions on Evolutionary Computation,1997,1(3):189-200.

[11] 刘晓冰,张浩,蒙秋男,等. 钢铁企业生产成本数据集市及数据挖掘研究[J]. 计算机集成制造系统,2006,12(10):1708-1712.

[12] 宋旭东,胡墨谦,刘晓冰. 基于模型驱动体系架构的联机分析处理开发方法[J]. 计算机集成制造系统,2010,16(2):423-430.

[13] 马丽娜,刘弘,张希林. 数据挖掘、联机分析处理在决策支持系统中的应用[J]. 计算机应用研究,2001,18(11):10-12.

[14] KANTARDZIC M. 数据挖掘——概念、模型、方法和算法[M]. 闪四清,陈茵,程雁,等,译. 北京:清华大学出版社,2003.

[15] 曲开社,成文丽,王俊红. ID3 算法的一种改进算法[J]. 计算机工程与应用,2003,39(25):104-107.