文章编号:1671-6833(2006)03-0077-04

一种基于分类和预测技术的产品成本估算系统研究与应用

李向宁[1,2], 郝克刚[1]

(1.西北大学计算机科学系, 陕西 西安 710069; 2.西安电子科技大学机电工程学院, 陕西 西安 710071)

摘 要: 应用分类和预测技术对机械产品设计数据库进行数据挖掘,针对各种部件影响成本的关键因素建立了基于判定树的产品成本预测模型,并与专家系统相结合,通过数据挖掘来调整专家系统的规则库,从而使专家系统规则不仅可以来自专家经验的总结,而且可以来自对以往产品数据的挖掘,拓宽了知识获取的途径,提高了知识获取的效率.最后给出了一个基于分类和预测技术的产品成本评估系统的框架.

关键词: 数据挖掘; 专家系统; 成本估算; 分类; 预测

中图分类号: TP 391 **文献标识码**: A

0 引言

在机械产品设计和开发过程中,成本是决定产品竞争力和生命力的关键因素.影响产品成本的因素很多,对于机械产品的设计人员而言,产品设计方案对产品成本的影响非常明显.如果能够在方案设计阶段就能够大致给出给定设计要求下的产品预测成本,那么在设计方案选择、优化与评估,项目报价与利润估算,投标竞标等活动中将起到非常重要的指导作用.

产品设计方案与成本之间存在着某些联系.比如:如果轴的精度参数要求比较高,它的成本相应也会比较高,小模数齿轮的成本比大模数齿轮高等等.获取这些知识通常的做法是在领域专家的帮助下建立成本评估专家系统,以规则的形式把知识储存到规则库中.由于设计方案与成本之间关系的复杂性,知识的获取往往是很困难的.而企业在产品设计过程中积累了大量产品设计数据,其中包含产品的设计参数和成本.这些历史数据中蕴藏着很多产品设计参数与成本之间关联性的知识,如果能找出这些知识,就可以利用它们对新设计方案的成本进行预测(评估).

数据分类[1](Data Classification)是数据挖掘中一项非常重要的任务.分类是指通过分析训练数据集中由属性描述的数据元组,建立一个分类函数或分类模型(也常常称作分类器).预测是指利用分类得出的模型对未知变量的类别进行推测,也就是把要进行预测的数据项映射到给定类别中的某一个.因此可以利用分类技术对产品设计数据库中的数据元组进行分类,建立一个以成本为类别,以影响产品成本的关键设计参数为属性的分类器,利用该分类器对新设计方案进行成本预测.

应用数据分类和预测技术对产品设计方案(或设计要求)进行成本估算是一个两步过程.第一步,利用分类技术对产品设计数据库中筛选出的数据进行分析并建立一个分类模型.第二步,从模型中提取分类规则,利用分类规则对设计方案(或设计要求)进行成本预测.

笔者应用数据挖掘中的分类和预测技术,对产品设计数据库进行数据挖掘,建立了基于判定树算法的分类模型,并与专家系统结合,实现了一个基于数据挖掘技术的方案成本评估系统框架.

1 构造训练数据集

建立分类模型,首先要构造被分析的数据元组,也即训练数据集.构造训练数据集也就是准备被挖掘的数据的过程.由于训练数据集中的每个元组都被预先明确标注为一个成本类别,建立分类模型也称为有指导的学习(即模型的学习在被告知每个训练样本属于哪个类别的"指导"下进

收稿日期:2006-03-31;**修订日期**:2006-05-14

基金项目:国家"863"高技术研究发展计划资助项目(2004AA115090)

作者简介:李向宁(1976-),男,河北冀州人,西北大学在读博士研究生,主要从事工作流模型、数据挖掘、软件理论等方面的研究.

行).对训练数据集进行分析的目的是建立一个以成本为类别,以影响产品成本的关键设计参数为属性的分类器,下面以齿轮[2]的成本估算为例进行说明.

表1 齿轮关键参数与成本评价

Tab.1 Key parameters of gear and its costs

轮型	齿型	精度	啮合关系	成本评价
圆柱	渐开线直齿	8	外啮合	f
圆锥	渐开线斜齿	7	外啮合	g
圆柱	人字齿	7	内啮合	j
……	……	……	……	……

齿轮的设计参数很多,包括齿轮的大小、齿型、轮型、材料、啮合关系、表面处理方式、精度、表面粗糙度等.这些参数对齿轮的成本都会产生影响,但影响的程度不一样.如果把这些参数都选为训练属性则会导致类别过多,而每一类的支持样本数过少,模型不具有普遍意义.因此在构造训练数据集时应该选择对成本影响最为关键的参数作为训练属性,舍弃次要参数.训练属性的选择对于分类模型的精确度和效率影响非常大,选择合适的训练属性通常需要领域专家的协助.作为示例,我们选择对齿轮成本起主要决定作用的轮型、齿型、精度、啮合关系等设计参数为训练属性,从设计数据库提取训练数据集如表1,表中最后一项属性为成本评价值,也即分类中所谓的"类别",由英文字母表示.由于各个产品设计数据库之间的异构性,通常这样的数据集不是天然存在的.这需要一个专门的程序负责从特定的数据库中提取相应属性以构造数据集,也就是我们后面将要介绍的"数据泵".以下为了方便问题的描述,假定上述数据集已获取.

2 建立基于判定树的分类模型

判定树[3,4]是一个类似流程图的树结构,其中每个内部节点表示在一个属性上的测试,每个分支代表一个属性输出,每个树叶节点代表一类或者类的分布,树的最顶层节点是根节点.判定树归纳的基本算法是贪心算法[1],它以自顶向下递归的构造判定树.判定树归纳算法的基本策略如下:

(1) 树以代表训练样本的单个节点开始.

(2) 如果样本都在同一个类,则该节点成为树叶,并用该类标记.

(3) 否则,算法使用称为信息增益的基于熵的度量作为启发信息,选择能够最好的将样本分类的属性.该属性成为节点的"测试"属性.

(4) 对测试属性的每一个已知的值,创建一个分支,并据此划分样本.

(5) 算法使用同样的过程,递归的形成每个划分上的判定树.一旦一个属性出现在一个节点上,则不再考虑出现该节点的任何后代上.

递归划分步骤仅当下列条件之一成立时停止:①给定节点的所有样本属于同一个类.②没有剩余属性可以用来进一步划分样本.在此情况下,使用多数表决,将给定的节点转换为树叶,并用样本中的多数所在的类标记它.③分支的当前测试属性没有样本.这种情况下,用样本中的多数所在的类创建一个树叶.在树的每个节点上使用信息增益(Information gain)度量选择度量属性.这种度量称为属性选择性度量或优良性度量.选择具有最高信息增益(或最大熵压缩)的属性作为当前节点的测试属性.这种信息理论方法使得对一个分类所需的期望测试数目达到最新,并确保找到一棵简单树.这个步骤如下:

设 S 是 s 个数据样本的集合,假定分类属性具有 m 个不同值,定义 m 个不同种类 $C_i(i=1,2,\cdots,m)$.设 s_i 是 C_i 中的样本数.对于一个给定的样本分类所需的期望信息由 $I(s_1,s_2,\cdots,s_m)=-\sum p_i\log_2(p_i)$ 给出,其中,p_i 是任意样本 C_i 的概率,用 s_i/s 估计.

设属性 A 具有 v 个不同值$(a_1,a_2,\cdots,a_v)$.可以用属性 A 将 S 划分为 v 个子集$(S_1,S_2,\cdots,S_v)$,其中 S_j 包含 S 中这样一些样本,他们在 A 上具有值 a_i.如果 A 选作测试属性(即最好的分裂属性),则这些子集对应于由包含集合 S 的节点生长出来的分支.设 s_{ij} 是子集 S_j 中类 C_i 的样本数.根据由 A 划分成子集的熵(Entropy)或期望信息由 $E(A)=\sum_{i=1}^{v}\frac{s_{1j}+\cdots+s_{mj}}{s}I(s_{1j},\cdots,s_{mj})$ 给出,其中项 $\frac{s_{1j}+\cdots+s_{mj}}{s}$ 充当第 j 个子集的权,并且等于子集(即 A 值为 a_i)中的样本个数除以 S 中的样本总数.熵值越小,子集划分的纯度越高.对于给定的子集 S_j,$I(s_{1j},s_{2j},\cdots,s_{mj})=-\sum_{i=1}^{m}p_{ij}\log_2(p_{ij})$,其中 p_{ij} 是 S_j 中的样本属于类的概率,则在 A 上分枝将获得的信息增益是:$\mathrm{Gain}(A)=I(s_1,s_2,\cdots,s_m)-E(A)$.

算法计算每个属性的信息增益.具有最高信息增益的属性选作给定集合 S 的测试属性.然后创建一个节点,并以该属性标记,对属性的每个值

创建分枝，并据此划分样本.

在判定树构造时，由于数据中的噪声和孤立点，许多分枝反映的是训练数据中的异常，也就是说某项数据项所表现的特征并不具有普遍规律，这时候需要用剪枝方法[4]处理这种数据问题.这种方法使用统计度量，剪去最不可靠的分枝，加快分类的速度，提高树独立于测试数据的正确分类能力.通常有两种常用的剪枝方法，先剪枝(prepruning)和后剪枝(postpruning).先剪枝由于可以在树的构造过程中进行而较多的被采用.在先剪枝方法中，通过提前停止树的构造(通过决定在给定的节点熵不再分裂或划分训练样本的子集)而对树"剪枝".一旦停止，节点变成树叶.该树叶中将持有子集样本中最频繁的类.

根据先剪枝策略，在构造树时采用信息增益评估分裂的优良性.如果在一个节点划分样本将导致低于信息增益预定义阈值的分裂，则给定子集的进一步划分停止.选取一个适当的阈值是很关键的，较高的阈值可能导致树的划分过于简单，而阈值太低可能使得树没怎么被修剪.

图 1 是由表 1 中数据生成的判定树.其中每个内部节点(用矩形表示)代表在一个属性上的测试，每个分支代表一个测试输出，而每个树叶节点(用圆表示)代表一个估算类别.树的最顶层节点是根节点.由这个判定树我们也可以大致了解到，精度是对齿轮成本影响最为明显的因素(首选的分裂属性).

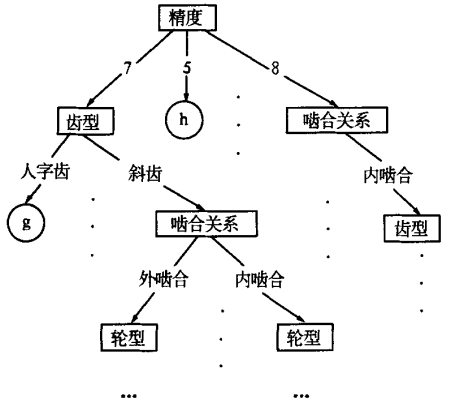

图 1　由表 1 数据生成的判定树

Fig.1　Decision tree built from data in Tab.1

3　基于数据挖掘技术的方案成本评估原型系统设计与实现

基于分类和预测技术的产品成本评估原型系统的结构如图 2 所示.该系统设计为数据挖掘工具与专家系统相结合，其中带灰色边框的矩形为软件模块，无边框矩形为数据，另外还有外部数据源，表示各种异构的产品设计数据库.首先由数据泵(Data Pump)根据训练属性从产品设计数据库(Product Design Data)中提取数据构造训练数据集(Training Data Set).实现了判定树分类算法的分类引擎(Classification Engine)对训练数据集进行分析，建立分类模型并输出分类规则(Classification Rules).分类规则提取方法如下：

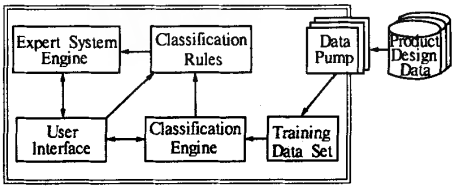

图 2　基于分类和预测的产品成本评估系统

Fig.2　Product cost evaluating system based on data classification and predication

对判定树从根到树叶的每条路径创建一个规则，沿给定路径上的每个属性－值对形成规则前件("IF"部分)的一个合取项.叶节点包含类型预测，形成规则后件("THEN"部分).

以下是由图 1 所表示的判定树提取的两条规则：

IF 精度 = "5"THEN 成本评价 = "h"

IF 精度 = "7" AND 齿型 = "人字齿"THEN 成本评价 = "g".

分类规则为专家系统引擎(Expert System Engine)所用.用户界面实现各模块之间的交互，包括与专家系统外壳的交互以及与分类引擎的交互，此外领域专家还可以通过用户界面对分类规则进行修改和筛选，调整误差过大的规则，剔除重复或错误的规则.

数据泵是一种用来进行数据提取和迁移的软件组件，它按照应用程序的要求从数据库中抽提取所需数据.数据泵对应用程序屏蔽了不同数据源库结构的异构性，使得应用程序可以面对一个统一的数据接口进行操作.图 2 中数据泵跨越了系统的边界并显示为集合，这是由于不同企业和单位的产品数据库结构多种多样，数据泵通常需要定制以适应不同结构的数据源.

与齿轮的判定树类似，对于轴、箱体、飞轮、蜗杆等部件建立相应的预测模型并提取分类规则.用户界面部分接收用户对于某种产品(如减速器)的设计参数，对其中的构件分别应用相应的分类

规则,最后系统对这些结果进行加权,输出整个产品的成本估值区间.

此外对于成本评估类别与货币值之间的对应,另外建立一张评估类别与货币值之间的映射表,如表2,用来表示各级评估类别与当前等价货币值之间的对应关系.这是张表是可调的,根据当前实际物价水平或运行过程中反馈的评价误差调整评估类别对应的货币值.这样就实现了抽象评价与确切数值之间的解耦.在预测结束得到评估类别后,通过查表,就能得到该类别对应的货币值区间.

表2 成本评估类别与货币值对应关系

Tab.2 Correspondence of estimated cost sorts and their monetary values

成本评价	值区间/元
a	0~50
b	51~100
……	……
y	1250~

4 结语

介绍了一种应用分类和预测技术对机械产品数据库进行数据挖掘以实现产品设计方案阶段的成本预测的方法,针对各种部件影响成本的关键因素建立了基于判定树的预测模型,并与专家系统相结合,实现了通过对产品数据库进行数据挖掘来调整专家系统的规则库,从而使专家系统规则不仅可以来自专家经验的总结,而且可以来自对大量历史数据的挖掘.尽管数据挖掘技术不能完全代替专家知识的总结,并且所获取的知识需要专家进行调整和筛选,但是它拓宽了知识获取的途径,大大提高了知识获取的效率.

在以上研究的基础上,我们用Java语言开发了基于判定树算法的分类引擎和一个数据泵,并采用Jess(Java expert system shell)[5]作为规则引擎(专家系统引擎).Jess提供了丰富的外部用户接口,据此开发了与Jess以及规则引擎交互的用户界面.利用该原型对我校机加工中心机械设计信息数据库进行数据挖掘,验证了上述方法的有效性,取得了良好的效果.

参考文献:

[1] HAN J W, KAMBER M. Data mining: concept and techniques[M]. Hong Kong: Morgan Kaufmann Publishers, 2000.149~183.

[2] 杨可帧,程光蕴.机械设计基础(第三版)[M].北京:高等教育出版社. 1997.54~73.

[3] BRODLEY C E, UTGOFF P E. Multivariate decision trees[J]. Machine Learning, 1995,19:45~77.

[4] QUINLAN J R. Simplifying decision trees[J]. International Journal of Man-Machine Studies, 1987,27:221~248.

[5] HILL E F. Jess[CP]. http://herzberg.ca.sandia.gov/jess/.

[6] 雷文平,黄式涛,石金彦.粗糙集与决策树结合诊断故障的数据挖掘方法[J].郑州大学学报(工学版),2003,24:109~112.

Research and Application of A Product Cost Estimating System Based on Data Classification and Predication

LI Xiang-ning[1,2], HAO Ke-gang[1]

(1.Department of Computer Science, Northwest University, Xi'an 710069, China; 2.School of Mechanical and Electronic Engineering, Xi'an University of Electronic Technology, Xi'an 710071, China)

Abstract: This paper applies data classification and predication techniques to mechanical products design database, and presents a predication model based on decision trees according to the key factors influencing the cost of each kind of mechanical components. Combined with expert system, this approach realized adjusting rules through data mining. Thereby the rules in expert system come not only from the experience of experts but also from mining the product data. This widens the source of knowledge and enforces the system with better flexibility. Finally a framework of product cost evaluating system based on data classification and predication techniques is given.

Key words: data mining; expert system; cost estimate; classification; predication