

演化聚类在离散制造业质量管理中的应用

王鹏飞,舒红平,郑皎凌,文立玉

(成都信息工程学院软件工程学院,成都 610225)

摘要:针对离散制造业质量管理体系中维度高,且存在较多一致性数据的情形,设计了一种基于信息论中的信息熵,即互信息的改进聚类算法。通过实验分析,采用改进的聚类算法可有效提高聚类的正确率,并且通过演化聚类理论和方法的应用可对离散制造业质量管理提供有效的决策参考信息。

关键词:信息熵;互信息;演化聚类;质量管理

中图分类号:TP311

文献标志码:A

引言

制造业按其产品制造工艺过程特点总体上可概括为离散型制造业和流程制造业两种。典型的离散型制造行业包括电脑、汽车及工业用品制造等行业。本文即是以典型的离散型制造业——德阳东汽工模具分厂(以下简称工模具分厂)为研究对象。由于其主要生产复杂的机械产品,大多产品都可能包含若干组零部件,对于每个零部件又有其生产工艺和周期,而每道工序又涉及相应的加工资源,如机床、夹具、量具,甚至人力。所以离散制造业产品种类多,物料多样化,加工过程复杂的特点使得其质量管理与不合格原因分析更加困难^[1]。

然而在我国制造业低人力成本的竞争优势已逐年削弱的情形下,提升产品质量已经迫在眉睫。可喜的是,近年来我国大型制造企业的信息化水平有了长足进步,并且积累了大量的与质量相关的数据,为开展相关研究提供了丰富、真实可信的资料。由于传统的研究方

法难于驾驭多年积累的海量数据,数据挖掘相关理论与方法应运而生。

本文所述问题即来源于工模具分厂产品质量管理系统,下面通过对该厂的产品不合格原因的聚类研究为基础来说明本文将要解决的问题。

在工模具分厂中,很多因素影响产品的质量,导致决策者常常面对复杂的情况难于做出决策。而针对质量相关因素得出较好的聚类结果则可以使管理者抓住关键,对产品质量有较深入的了解,从而采取措施减少不合格品数量,提升产品整体质量。为了使数据集符合聚类相关算法的要求,对原始数据集进行抽取、转换、加载等操作后,质量相关的字段主要有:不合格时间、不合格原因 ID、不合格品级别、操作者、操作者所属部门、责任部门、本工序工时、前总工序工时、合计工时、交检数、不合格品数、回用数、报废数、返修数、图号、材质、是否进入经济责任反馈。具体见表 1。

由于数据的属性较多,且存在很多一致性数值,所以用传统的 K - Means 算法建立相应的聚类模型会使得

收稿日期:2013-03-16

基金项目:国家 973 计划子课题(2012CB518500);国家自然科学基金青年基金项目(61202250,61203172);四川省教育厅青年基金项目(11ZB088);四川省应用基础计划(2012JY0112);成都市科技计划项目(12DXYP100JH-002);成都信息工程学院中青年学术带头人科研基金(J201208,J201101);成都信息工程学院引进人才项目(KYTZ201110,KYTZ201111);四川省科技支撑计划项目(2011SZZ027)

作者简介:王鹏飞(1987-),男,河南安阳人,硕士生,主要从事数据库与知识工程、计算机在制造业中的应用方面的研究,(E-mail) 523774679@qq.com

表1 数据库仓库字段

| 字段 | 名称 |
|------------|------------|
| ADDTIME | 不合格时间 |
| REASONID | 不合格原因 ID |
| DISGRADE | 不合格品级别 |
| RUNNER | 操作者 |
| RUNNERDEPT | 操作者所属部门 |
| DUTYDEPID | 责任部门 |
| SEQUHOUR | 本工序工时 |
| BEFOREHOUR | 前总工序工时 |
| TOTALHOURS | 合计工时 |
| CHECKNUM | 交检数 |
| DISNUM | 不合格品数 |
| CIRCLENUM | 回用数 |
| REJECTNUM | 报废数 |
| FANXIUNUM | 返修数 |
| PICNUM | 图号 |
| MATERIAL | 材质 |
| ISECO | 是否进入经济责任反馈 |

多次调整参数得出的聚类结果极其相似,且随着每一次同方向的调整,结果无限接近于最优,聚类效果较不理想,难于理解,不利于实践中的理解和应用。

针对上述问题,改进传统 K - Means 算法的距离计算指标,采用信息论中的“互信息”,即 J. R. Quinlan 提出的 ID3 算法中的“信息增益”来计算每两条记录之间的相异度,替代传统的欧几里得距离,将改进后的算法嵌入开源数据挖掘工具 weka 中反复进行建模,并利用演化聚类理论来获得(解释)较好的聚类结果,并在实践中加以运用。

1 理论基础

传统的基于划分的聚类研究主要是基于欧几里得距离来对数据进行标准化处理和计算对象之间的相异度的。计算公式如下:

$$d(i, j) = \sqrt{|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \dots + |x_{ip} - x_{jp}|^2} \quad (1)$$

这里的 $i = (x_{i1}, x_{i2}, \dots, x_{ip})$ 和 $j = (x_{j1}, x_{j2}, \dots, x_{jp})$ 是数据集中的两个 p 维的数据对象,并使用平方误差准则作为收敛函数的。计算公式如下:

$$E = \sum_{i=1}^k \sum_{p \in C_i} |p - m_i|^2 \quad (2)$$

这里的 E 是数据集中的所有对象的平方误差的总和, p 是空间中的点,表示给定的数据对象, m_i 是簇 c_i 的平均值(p 和 m_i 都是平均值)。然而这样的 K - Means 算法仅适用于对象数值型数据较多,数据维度不高的情形,当面对数据维度很高,存在分类型数据,并且数据属

性值存在较多一致性数值时,便会导致聚类模型过于相似,聚类结果难以解释和运用^[2]。

目前国内外很多学者致力于聚类算法的研究和改进。孙吉贵等即挑选出了 8 种常用聚类算法,采用 UCI 机器学习数据集储存库中的知名数据集分别进行了随机性实验分析。其中为判断 k 最近邻一致强制和保留算法是否明显优于 K 均值(K - Means)算法^[3], kNN 一致性与聚类质量之间有何关系,其采用 Imagine、Iris、Wine、Glass、Ionosphere 等数值型数据集进行了 20 次随机聚类实验。从聚类结果的正确率对比来看,可以得出聚类结果在很大程度上依赖于所用的相似性度量方式。

然而对于数据集一致性数据过多、数据维度较高的情形,采用可有效体现对象之间的相异度的度量方式如信息熵才可取得较好的聚类结果。本文即利用信息论中的“互信息”来计算对象之间的相异度,改进传统的 K - Means 算法,提高了聚类的正确率和可理解性。

2 基于传统的 K - Means 聚类算法研究及其实现

2.1 传统 K - Means 聚类算法简介

K - Means 聚类算法是聚类分析中使用最为广泛的算法之一。该算法对大数据集的处理效率较高,特别是对于发现凸面形状的簇很是适合,可以达到较优的聚类结果。该算法是以确定的类数 k 且对象之间的距离可以度量的条件下,并以选定的初始聚类中心为前提下进行的,当选用的准则函数收敛时结束。通常采用欧几里得距离计算对象之间的相异度,因此更适用于数值型数据,计算方法如式(1)所示。而准则函数则选用平方误差准则,计算方法如式(2)所示。

算法的基本思想是:给定一个有 n 个数据对象的数据集,以及要生成的簇的数目 k , 首先选出 k 个对象作为初始聚类中心,按最小距离原则将各对象分配到 k 类中的某一类,之后不断计算类心和调整各对象的类别,最终使各对象到其判属类别中心的距离平方之和最小。算法步骤如下^[4]:

(1) 针对 n 个对象,任选 k 个对象作为初始聚类中心(z_1, z_2, \dots, z_k)。

(2) 对每个对象 x_i , 找到离它最近的聚类中心 z_r , 并将其分配到 z_r 所标明的类 u_r 。

(3) 采取平均的方法计算重新分类后的各类心。

(4) 计算式(2)的最小值。

(5) 如果式(2)收敛,则 $\text{return}(z_1, z_2, \dots, z_k, U)$ 并终止本算法,否则转 2)。

2.2 基于传统的 K - Means 聚类算法的实验及分析

本次实验的数据集来自工模具分厂自 2006 年至 2010 年积累的生产数据,真实可靠。数据对象的属性见表 1,部分数据见表 2。

表 2 不合格表(部分数据)

| ADDTIME | DUTYDEPID | PICNUM | ISECO |
|------------|-----------|--------------|-------|
| 2006090005 | 1006 | 9940 - 012 - | N |
| 2006100033 | 1007 | 4685 - 879 | N |
| 2007010024 | 0907 | 9930 - 558 | Y |

2.2.1 数据预处理

本次实验采用的原始数据,2006 年共 1069 条,2007 年 1289 条,2008 年 1157 条,2009 年 1286 条,2010 年 1583 条。由于数据量总体较大,很容易出现噪声数据、空缺数据,脏数据等不利于聚类实验的数据。且原始数据属性较多,共 40 余项,经过筛选影响产品质量的属性如表 1 所示。并且实验所用的聚类算法要求数据是离散数据、数值型数据,而且为了属性之间的可比性,数据还需经过标准化。本次实验采用著名开源数据挖掘工具 weka 来进行,所以不必自行进行标准化的操作。总之,我们利用数据清理、数据集成、数据变换、数据规约、数据替换等技术对原始数据集进行预处理,且转换成 weka 的标准数据文件后,文件注释如下所示:

```
@ attribute ADDTIME numeric
@ attribute DUTYDEPID { 1006,1007, ... }
@ attribute PICNUM { 9940 - 012, ... }
@ attribute ISECO { N, Y }
```

2.2.2 聚类过程与分析

我们采用 K - Means 聚类算法对预处理后的数据集分年份进行聚类,其中当设置 $k = 4$, $seed = 50$ 时,2007 年的数据集聚类结果如图 1 所示。图 1 是由 weka 生成的聚类结果可视化展示图,其中每一种颜色代表一个簇。由图 1 可以看出,簇之间分不开,聚类结果较不合理。

同时变换参数值进行实验时,聚类结果如图 2 所示。图 2 中,横坐标表示 k 的取值,纵坐标表示对应的 k 和 $seed$ 的取值(选取的种子数)时簇内方差的取值。取值越小则簇内越紧凑,聚类结果越合理。由图 2 知随着 k 值的不断增大,聚类评价指标值越来越小,聚类结果越来越理想。且数据每年分为 10 个近似簇时,记录的簇归属有较明显变化。后来经过深入分析发现由

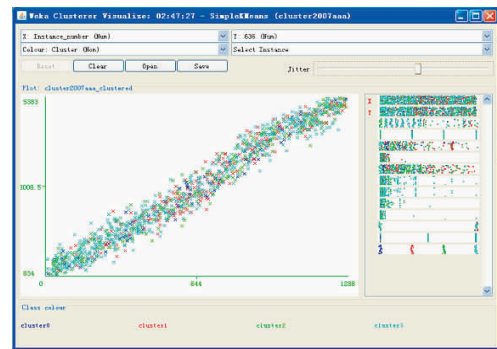


图 1 2007 年聚类结果散点图(传统算法)

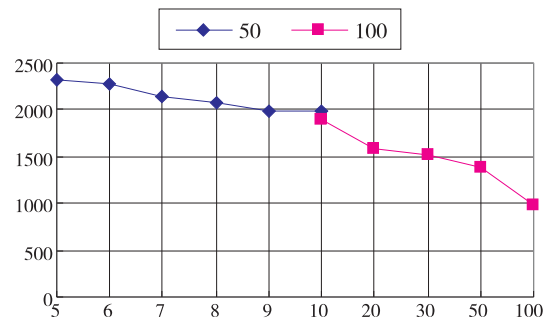


图 2 2007 年数据集聚类结果

于数据集某些属性经过手工转换和 weka 的自动标准化后,数值过于接近,而采用了传统的欧几里得距离难于有效度量对象之间的相异度,所以阻碍了实验的良好效果。

3 基于信息熵的 K - Means 聚类算法设计及其实现

3.1 信息熵的定义及计算

信息论是 C. E. Shannon 为了解决信息传递(通信)过程问题而建立的理论。其把通信过程看成是在随机干扰的环境中传递信息的过程,由发送端(信源)和接收端(信宿)以及连接两者的通道(信道)三者组成。

信源用值域 $U\{u_1, u_2, \dots, u_r\}$ 表示,信宿用值域 $V\{v_1, v_2, \dots, v_q\}$ 表示。相关定义如下:

定义 1 消息(符号)

$u_i (i = 1, 2, \dots, r)$ 的发生概率 $P(u_i)$ 组成信源数学模型(样本空间和概率空间)

$$[U, P] = \begin{bmatrix} u_1 & u_2 & \dots & u_r \\ P(u_1) & P(u_2) & \dots & P(u_r) \end{bmatrix} \quad (3)$$

定义 2 自信息

消息 u_i 发生后所含有的信息量反映了消息 u_i 发生前的不确定性(随机性),定义为:

$$I(u_i) = \log \frac{1}{P(u_i)} = -\log P(u_i) \quad (4)$$

log以2为底,所得的信息量单位为b。

定义3 信息熵

自信息的数学期望,即信源发出消息后,信源消息所提供的信息量,也反映出了信源发出消息前的平均不确定性^[5]。定义为:

$$H(u) = - \sum_i P(u_i) \log(P(u_i)) \quad (5)$$

3.2 基于信息熵的K-Means聚类算法的实验及分析

式(5)是一维变量时信息熵计算公式,如果是n维, U_1 表示 u_1 的值域,则信息熵表示为:

$$H(u_i, u_j) = - \sum_{u_1 \in U_1, u_2 \in U_2} \dots \sum_{u_n \in U_n} P(u_1 u_2 \dots u_n) \log P(u_1 u_2 \dots u_n) \quad (6)$$

本文即采用式(6)来计算二维对象间的距离,由于实验数据集中存在较多分类型、标称变量型对象,所以利用信息熵作为衡量对象间的相似性的度量可以有效避免欧几里得距离对标称变量噪声比较敏感的缺点,获得较好的聚类结果^[6]。

为了检验改进聚类算法的效果,以信息熵作为对象之间的相似性即相异度的度量,同时以K-Means聚类算法的思想作为基础对给定的数据集进行聚类。结果如图3所示。

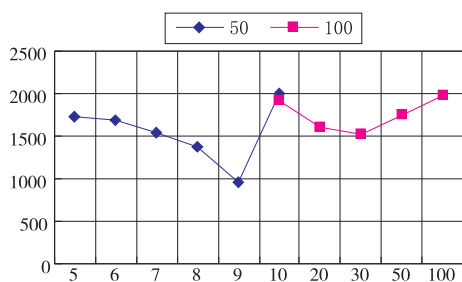


图3 2007年数据集聚类结果(改进算法)

由图3可见,衡量对象间的相异度的方法与聚类结果有着直接的关系,同样的参数设置下,本次聚类结果各簇分得较开,聚类结果较合理,且获得了最佳聚类数^[7]。在面对数据集维度较高,一致性数据过多的情形,即工模具分厂所属的情况时,采用信息论中的信息熵来计算对象间相异度可以获得较好的聚类结果。

3.3 聚类结果的意义及演化聚类的应用

上述我们已经通过改变距离计算方式得出比较合理的聚类结果,下面我们分析一下聚类的实际意义。由表3可知,在2006年9月,且责任部门ID为1006,在不

计入经济责任时,根据图号9940-012-生产容易产出不合格产品。在2009年9月,且责任部门ID为0902,在不计入经济责任时,根据图号2282-146-生产同样容易产出不合格产品。这样我们就可以很直观地从海量数据中找出规律,为领导决策提供依据。

表3 聚类后簇的中心点(部分数据)

| 簇 | ADDTIME | DUTYDEPID | PICNUM | ISECO |
|---|------------|-----------|-----------|-------|
| 1 | 2006090005 | 1006 | 9940-012- | N |
| 2 | 2009090176 | 0902 | 2281-146 | N |

由上述信息熵的定义可知,条件熵反映了整体数据集的不确定性,由于其是由每个取值的正例与反例计数组成计算因子,ID3算法即是根据此来挑选决策树的根节点,信息熵越小,则互信息越大,通过此信息来反映数据集的整体特征^[8-11]。同时熵越小则记录集合的无序性越小,即数据集的属性越有顺序越有规律。由表4可知,2008年由于地震灾害的影响数据较不稳定,其余年份则稳定在0.03左右。演化聚类的理论即是结合数据集的稳定性来进行聚类,当数据集越稳定则越适合聚类,进行质量管理的分析更切合实际,反映真实生产情况。

表4 数据集分年份信息熵值

| 年份 | 2006 | 2007 | 2008 | 2009 | 2010 |
|----|--------|--------|--------|--------|--------|
| 熵 | 0.0339 | 0.0386 | 0.0479 | 0.0323 | 0.0328 |

4 结束语

传统的K-Means聚类算法中采用欧几里得距离来计算对象间的相异度,但面对数据集维度较高,且存在较多一致性数据、数值相差不多情形时无法获得较优的聚类结果。本文结合信息论中的信息熵,可有效衡量对象之间的相似度。在此基础上提出了一种新的改进的K-Means聚类算法,理论研究和实验结果证明以上算法方案的合理性和有效性,同时结合演化聚类理论更好地为决策者提供辅助参考信息。

参考文献:

- [1] 舒红平,游志胜,蒋建民.基于信息熵的决策属性分类挖掘算法及应用[J].计算机工程与应用,2004(1):186-189.
- [2] 舒红平,郭远远,龚双龙,等.决策树分类在离散制造业中的研究与应用[C].//2009国际信息技与应用论坛论文集.IEEE Computer Society,2009:117-119.
- [3] 孙吉贵,刘杰,赵连宇.聚类算法研究[J].软件学报,

- 2008,19(1):48-61.
- [4] Kamrani A, Wang R, Ricardo G. A genetic algorithm methodology for data mining and intelligent knowledge acquisition[J].Computers & Industrial Engineering 2001, 40:361-377.
- [5] 高坚.基于 C2 均值和免疫遗传算法的聚类分析[J].计算机工程,2003,29(12):65-66.
- [6] 徐志伟,冯百明,李伟.网格计算技术[M].北京:电子工业出版社,2004.
- [7] 庞彦军,刘立民,刘开第.未知均值聚类[J].河北工程大学学报:自然科学版,2010,27(4):98-100.
- [8] 姜园,张朝阳,仇佩亮,等.用于数据挖掘的聚类算法[J].电子与信息学报,2005,27(4):655-661.
- [9] 韩凌波.K-均值算法中聚类个数优化问题研究[J].四川理工学院学报:自然科学版,2012,25(2):77-80.
- [10] 崔勇,吴建平,徐恪.基于模拟退火的服务质量路由算法[J].软件学报,2003,14(5):877-884.
- [11] Avila J L, Gibaja E L, Zafra A, et al. A gene expression programming algorithm for Multi-Label classification[J].Journal of multiple-valued logic and soft computing,2011,17(2-3):183-206.

Application of Evolution of Clustering in Discrete Manufacturing Quality Management

WANG Peng-fei, SHU Hong-ping, ZHENG Jiao-ling, WEN Li-yu

(Software Engineering college of Chengdu University of Information Technology, Chengdu 610225, China)

Abstract: For high dimension and much consistent data in discrete manufacturing quality management system, an improved clustering algorithm is designed based on information entropy in information theory which is also called mutual information. Experimental analysis shows that using the improved clustering algorithm can effectively improve the correct rate of clustering, and the application of the evolution of the cluster theory and method can provide effective decision-making reference information for discrete manufacturing quality management.

Key words: information entropy; mutual information; evolution of clustering; quality management