

基于滑动窗口的定性趋势分析方法及应用*

高东, 马昕[†], 许欣, 张贝克

(北京化工大学信息科学与技术学院自动化系, 北京 100029)

摘要: 传统定性趋势分析方法中, 存在划分片段宽度难以自适应、依赖人为设定阈值、算法复杂等问题, 针对这些问题, 提出了一种新的基于滑动窗口的定性趋势分析方法。方法采用一个滑动的窗口, 并对窗口内的数据进行线性拟合, 根据拟合的情况扩大或缩小窗口, 确定每个片段的最佳大小, 将每个片段识别为上升、下降或不变, 最终获取数据的定性趋势。在化工过程趋势分析应用的结果表明, 该方法能够更为有效地提取、识别出趋势, 具有较高的准确性和较低的复杂度, 为之后的数据压缩、故障诊断等打下坚实基础。

关键词: 定性趋势分析; 滑动窗口; 自适应

中图分类号: TP391.9

文献标志码: A

文章编号: 1001-3695(2014)05-1441-04

doi:10.3969/j.issn.1001-3695.2014.05.037

Method and application of qualitative trend analysis with sliding window

GAO Dong, MA Xin[†], XU Xin, ZHANG Bei-ke

(Dept. of Automation, College of Information Science & Technology, Beijing University of Chemical Technology, Beijing 100029, China)

Abstract: In the traditional qualitative trend analysis, the width of segments cannot be adaptive, the threshold is dependent on human experience, and the algorithm is complicate. Due to these shortcomings, this paper proposed a new qualitative trend analysis method with a sliding window. In the method, data in a sliding window was fitted linearly. According to the result of linear fitting, it expanded or reduced the sliding window to determine the best width of a segment. When all the segments were determined, they were identified as increasing, decreasing or steady. Finally it obtained the qualitative trend of data. Its application on chemical data trend analysis proves that the method can extract and identify qualitative trend effectively with higher accuracy and lower complexity. It can lay a solid foundation for data compression and fault diagnosis.

Key words: qualitative trend analysis; sliding window; adaptive

0 引言

化工过程的数据趋势信息往往包含了大量有价值的信息, 包括系统的运行状态、响应、人的操作等。这些信息对于分析、监测整个系统的运行、预警、故障诊断等都具有非常重要的意义。然而随着化工过程越来越大, 导致数据量往往很大, 形成信息爆炸, 将有用的信息淹没。

定性趋势分析就是一种基于数据驱动的分析方法, 将过程数据用定性的趋势来代表, 在化工过程数据分析、数据压缩、故障诊断等方面有着广泛应用。Cheung 等人^[1]提出了将测量数据转换为定性趋势的系统性方法。此方法使用多个三角形的片段及其组合来代表数据定性的趋势, 定性趋势分析的概念由此引出并发展。Janusz 等人^[2]提出一种新的定性趋势表示方法, 采用七个最基本的基元 (primitive) 的组合来表示定性趋势。该方法中, 多个基元组成片段, 多个片段组成趋势, 七种基元的方法成为之后应用较为广泛的方法之一。Bakshi 等人^[3,4]将用三角片段表达趋势的方法与小波分析结合起来, 从不同的尺度将数据转换为定性趋势, 并将其应用到过程工业故障诊断中。Rengaswamy 等人^[5,6]在前人工作的基础上, 对趋势分析的算法进行了改进和扩展, 加入了语法模式识别, 并将算

法应用到流化催化裂化装置 (FCCU) 的故障诊断中。Dash 等人^[7]提出基于模糊逻辑的多变量推理系统。系统具备自动推理功能, 对两个剧情中的对应变量进行分析比较, 如果剧情中所有的变量都是相似的, 那么这两个剧情就被确定为相似。最后通过模糊逻辑计算出对应变量趋势之间的相似度。在前人工作基础上, Dash^[8]提出一种基于半分法 (interval halving) 的数据趋势提取、识别算法框架, 并将其应用到故障诊断中, 该方法也是目前应用较多的方法之一。Charbonnier 等人^[9]提出了一种定性趋势分析方法。与之前的采用其中基元的趋势提取算法不同, 此方法采用了三个基元 {上升, 下降, 不变} 的组合来描述趋势片段, 但趋势提取识别中需要人为设定几个阈值, 获取趋势的结果与设定的阈值有很大相关。Scali 等人^[10]提出一种定性趋势分析方法, 用于对控制回路中阀门的粘滞问题进行识别。在半分法基础上 Maurya 等人^[11]提出了一种在线的定性趋势提取识别方法以满足在线故障诊断的要求。

国内也已经展开了相关研究, 张海洋等人^[12,13]在 Charbonnier 的基础上改进了趋势识别方法, 并将其应用到化工过程故障诊断中。宋政辉等人^[14]将模糊逻辑与定性趋势分析结合, 大大缩减了趋势匹配时所需的数据库的规模。张静哲等人^[15]将趋势分析与符号有向图方法相结合进行故障诊断。陈

收稿日期: 2013-06-24; 修回日期: 2013-08-08 基金项目: 中央高校基本科研业务费资助项目 (ZY1208)

作者简介: 高东 (1982-), 男, 河北保定人, 讲师, 博士, 主要研究方向为化工仿真、故障诊断 (gaodong@mail.buct.edu.cn); 马昕 (1975-), 女 (通信作者), 讲师, 博士, 主要研究方向为故障诊断、数据挖掘; 许欣 (1981-), 男, 博士, 主要研究方向为化工仿真建模; 张贝克 (1976-), 男, 副教授, 博士, 主要研究方向为化工仿真、软件工程、安全工程。

俊平等^[16]在半分法的基础上加入了阶跃信号的识别,解决了阶跃信号对趋势提取、识别干扰的问题。邹莹等人^[17]将定性趋势分析应用到泡沫浮选故障诊断中。

综上所述,定性趋势分析方法已经广泛应用于化工过程以及其他行业的数据监测、故障诊断等领域。然而传统的定性趋势分析方法依然存在不少问题,如提取的片段宽度难以自适应、算法复杂、依赖人为设置阈值等。针对上述问题,本文提出一种新的基于滑动窗口的定性趋势分析方法。

1 定性趋势分析方法面临的问题

定性趋势分析是从含有噪声的过程数据中将变量的定性趋势提取并识别出来。描述趋势的基元包括一种基元法、三种基元法、七种基元法等。

目前定性趋势分析面临的主要问题包括:

a) 片段宽度的选择。这里的宽度指的是针对要分析的数据,需要确定采用多少个片段来描述它的趋势,即需要确定每个片段的宽度。片段宽度的大小会对最后趋势提取的结果产生较大的影响,甚至是不同的宽度得出完全不同的结果。

当片段的宽度较小时,虽然能够捕捉到数据细微的变化情况,但是过小的宽度会导致陷入到数据的细节中去,难以获得整体的趋势变化情况。而较大的宽度一方面可能会漏掉一些有用的细节,另一方面导致拟合的误差较大,难以反映真实的数据变化情况。

b) 趋势提取策略的选择。趋势提取中策略的选择也非常重要,这包括数据拟合时拟合阶次以及拟合方法的选择等。理论上讲,拟合的时候采用的多项式的阶次越高,拟合出的结果越精确,误差越小,但阶次越高,计算的复杂度越高,不利于实时计算。很多时候,对于化工过程数据,只要能够提取出趋势的基本特点就足以满足要求了。

c) 噪声的影响。化工过程中,信号中往往包含噪声,噪声可能会对趋势分析产生影响,以致不能获取真正的数据趋势。为了避免噪声对趋势分析的影响,通常在趋势分析之前加入去噪环节。

d) 计算复杂度。对于实时性要求很高的系统来讲,尤其是故障诊断系统,要求实时提取趋势,因此计算的复杂度要低。

每种定性趋势分析的方法都会面临上述问题,目前还没有一种方法能够将这些问题全部解决。

2 基于滑动窗口的定性趋势分析方法

针对目前定性趋势分析面临的一些问题,提出一种新的基于滑动窗口的定性趋势分析方法。此方法采用一个滑动的窗口对数据进行线性拟合,根据拟合的情况自适应地增大或减小窗口宽度,以实现趋势片段的最优分割,然后采用上升、下降、不变三个基元来描述所有的片段,最终完成定性趋势的识别。

2.1 趋势基元

采用上升、下降、不变三个基元来描述定性趋势,如图 1 所示。

上升代表数据处于上升的趋势,一阶导数为正;下降代表数据处于下降的趋势,一阶导数为负;不变代表数据没有变化,一阶导数为零。以上三个基元以及它们的组合,能够基本描述所有的化工过程趋势,如线性的上升、下降;非线性的变化可以

通过组合实现。采用三个基元的主要优点在于:a) 通过基元以及基元的组合,能够基本描述所有的化工过程趋势,满足定性趋势分析的要求;b) 基元个数少,且为线性,提取、识别时计算量小、复杂度低。

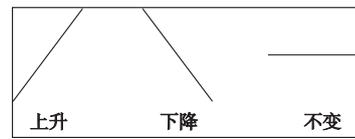


图1 三个基元

2.2 算法框架

算法主要分为提升格式滤波、片段提取、片段识别三部分。

1) 滤波去噪部分 实际工业中对数据进行实时采样是很重要的环节,但由于数据在激励、传输和检测过程中,可能不同程度地受到随机噪声和粗大误差的污染,这些包含噪声的数据进入到定性趋势分析中,可能会对最终的结果产生影响。因此,采用提升格式去除数据中可能包含的噪声,包括白噪声、粗大误差等^[18]。

2) 趋势提取 采用一个滑动窗口对数据进行线性回归,获得组成整个趋势的各个片段。由于算法采用了最小二乘来对数据进行线性回归,但事先并不能确定变量与时间之间具有线性关系,需要对拟合的结果进行显著性检验,验证拟合的效果。

算法主要步骤如下:

a) 假设经过去噪后的原始数据长度为 L 的数据为 y_1, y_2, \dots, y_L , 时间变量为 t_1, t_2, \dots, t_L 。其中下标代表序号,序号越大,代表数据越新。滑动窗口初始宽度为 M , 将 M 个原始数据 y_1, y_2, \dots, y_M 放入滑动窗口内;同时原始数据删除相应的数据。

b) 采用最小二乘法对滑动窗口内的数据用式(1)进行一元线性拟合

$$y(x) = f(x) = ax + b \quad (1)$$

采用 F 检验判断拟合效果,如果拟合的结果满足要求,即窗口的数据为线性,即可以继续扩大窗口,从原始数据中移入新的数据,继续拟合;F 检验判断是否满足要求,直到不能满足要求。这样一个线性的片段确定出来,转到步骤 d) 对滑动窗口内片段进行进一步划分;如果不能满足要求则转到步骤 c)。

c) 对于拟合效果不能满足要求的分两种情况。一种情况是初始滑动窗口设置大了,窗内数据是非线性,导致拟合效果不能满足要求,这种情况就需要减小窗口宽度,重新拟合;另一种情况是窗口内数据是不变的,需要对这两种情况进行区分:

(a) 针对滑动窗口内的数据,计算 y 的方差,并与阈值比较,如果小于阈值,说明数据没有变化,继续从原始数据移入新的数据,计算方差,判断是否没有变化,直到方差大于阈值,这样一个不变的片段被提取出来。转到步骤 e)。

(b) 如果 y 的方差大于阈值,说明数据有变化,滑动窗口设置大了,就减小窗口,将新移入的数据放回原始数据中,重新拟合,F 检验,直到满足要求或者窗口内数据方差小于阈值。这样一个新的片段被提取出来。转到步骤 e)。

d) 滑动窗口内的数据已经过 F 检验,满足线性要求。但是由于采用 F 检验判断窗口内数据是否线性的标准较为宽松,需要对大于等于滑动窗口初始宽度的数据作进一步划分。假设窗口宽度为 W , 数据为 y_1, y_2, \dots, y_w 。对窗口数据分成两部分分别进行线性拟合,即对 y_1, y_2, \dots, y_i 以及 y_i, y_{i+1}, \dots, y_w

($1 < i < w$) 进行线性拟合,并计算两部分的误差总和,取误差总和最小的 i 为最佳分割点。一个新的片段 y_1, y_2, \dots, y_i 产生,剩余的数据放回原始数据中。转到步骤 e)。

e) 新的片段产生,为了保证所有片段的连续性,将新片段最后一个数据放入原始数据中,清空滑动窗口数据,继续从原始数据装入 M 个数据,转到步骤 b)。如果原始数据长度为零,表明所有片段都被提取出来,转到步骤 f)。

f) 对所有提取出来的片段进行识别。如果提取时,片段是没有变化的,即为不变;否则,如果 $a > 0$, 片段为上升, $a < 0$, 片段为下降。

具体的框架流程如图2所示。

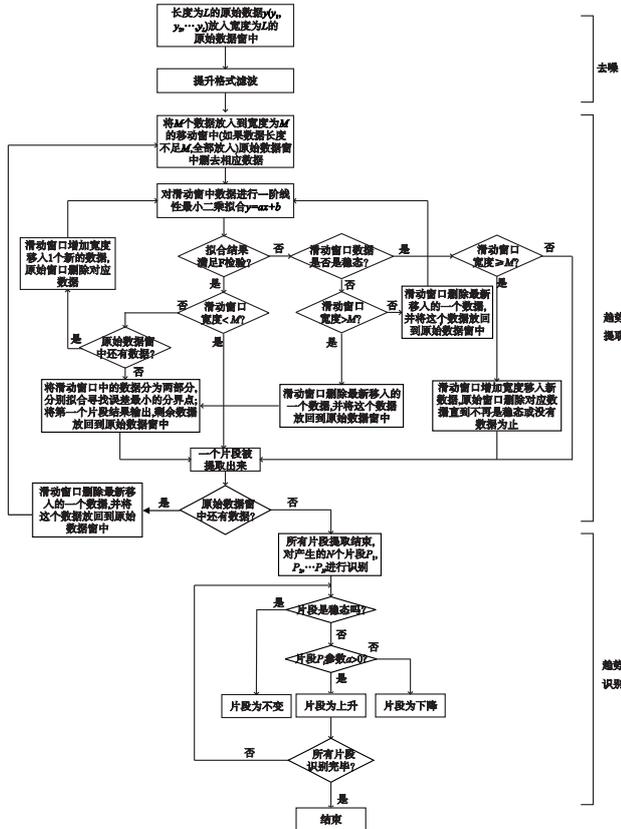


图2 基于滑动窗口的定性趋势分析算法框架

与传统方法相比,基于滑动窗口的定性趋势分析方法主要的优点如下:

- a) 趋势片段划分更为精确。采用一个自适应的滑动窗口对数据进行拟合,根据拟合的情况扩大或者减小窗口的宽度,最终获取更为精确的片段划分。
- b) 采用上升、下降、不变三个基元,通过基元以及基元的组合能够基本描述所有的化工过程趋势,满足定性趋势分析的要求,并且基元个数少,且为线性,提取、识别时计算量小、复杂度低。
- c) 趋势提取、识别过程不依赖人为设定的阈值,避免了不同阈值导致不同识别结果的情况。
- d) 采用基于提升格式的方法对数据进行去噪处理,避免随机误差和粗大误差的干扰。
- e) 整个算法复杂度低,计算量小。

3 案例分析

针对某化工过程发生故障后蒸汽流量信号,原始数据长度

为715,如图3所示。分别采用固定窗口宽度线性拟合方法、基于滑动窗口的线性拟合方法对数据进行拟合识别,结果如图4~6所示。最终结果对比如表1所示。

如表1和图4~6所示,固定宽度线性拟合的方法最为机械,误差最大;基于滑动窗口的方法在不同初始窗宽度的情况下,对趋势的拟合都明显比固定宽度的方法好,尤其是当实际数据变化波动较大时,固定宽度的方法误差会更大。

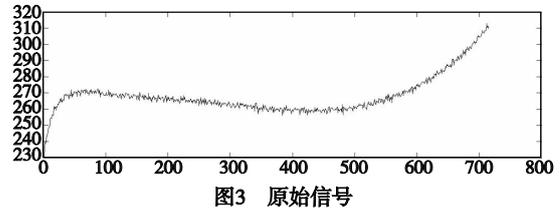


图3 原始信号

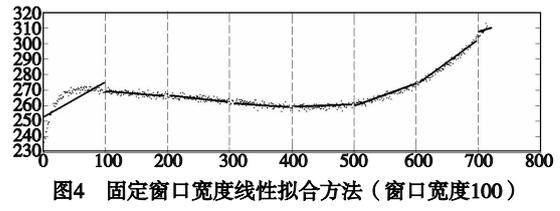


图4 固定窗口宽度线性拟合方法(窗口宽度100)

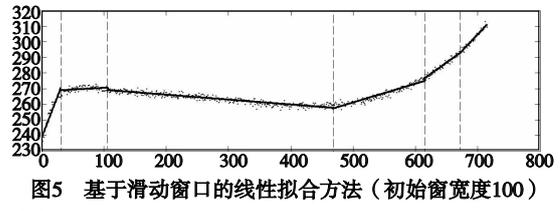


图5 基于滑动窗口的线性拟合方法(初始窗宽度100)

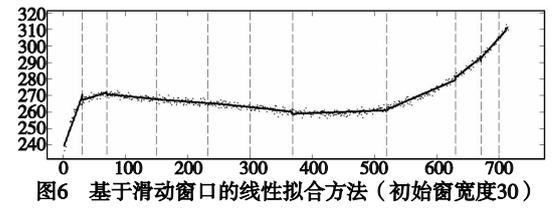


图6 基于滑动窗口的线性拟合方法(初始窗宽度30)

表1 不同定性趋势分析方法比较

方法	片段	误差
固定宽度线性拟合(100)	8	1 330.67
基于滑动窗口的方法(100)	6	809.31
基于滑动窗口的方法(30)	12	754.86

对于基于滑动窗口的方法,初始窗口为100的时候得到的效果最好,基本反映了数据的整体趋势。一般来讲,初始窗口越小,得到的片段越多,误差越小,但过多的片段可能会影响整体趋势的判断,需要将片段进行合并;初始窗口较大时,得到的片段会少些,误差会比初始窗口较小时大些,但容易获得整体趋势。实际应用中,可以根据不同情况进行选取,也可以采用多个窗口进行拟合识别,最终根据片段多少和误差情况综合比较选取。

综上所述,基于滑动窗口的定性趋势提取、分析方法采用的滑动窗口保证了片段划分的自适应,拟合时不再依赖人为设定阈值,且采用线性拟合在满足趋势分析需要的同时,确保算法较低的复杂度,便于实际应用。

4 结束语

针对传统定性趋势分析方法中存在划分片段宽度难以自适应、依赖人为设定的阈值、算法较为复杂等问题,并结合实际化工过程数据特点,提出了一种新的基于滑动窗口的定性趋势

分析方法,采用上升、下降、不变三个基元及其组合来描述定性趋势。通过一个滑动的窗口,对窗口的数据进行最小二乘线性拟合,并根据 F 检验判断拟合效果,根据拟合效果增大或缩小窗口,实现趋势片段更为精确的划分,将片段识别为上升、下降、不变三个基元之一,从而最终获取数据的定性趋势。化工过程实际数据趋势分析的结果表明,该方法能够更为有效地提取、识别出趋势,具有较高的准确性和较低的复杂度,可以为后续的数据分析、故障诊断等提供有效的支持。

参考文献:

- [1] CHEUNG J T, STEPHANOPOULOS G. Representation of process trends, part I: a formal representation framework [J]. *Computer & Chemical Engineering*, 1990, 14(4/5): 495-510.
- [2] JANUSZ M, VENKATASUBRAMANIAN V. Automatic generation of qualitative description of process trends for fault detection and diagnosis [J]. *Engineering Applications of Artificial Intelligence*, 1991, 4(5): 329-339.
- [3] BAKSHI B R, STEPHANOPOULOS G. Representation of process trends, part III: multiscale extraction of trends from process data [J]. *Computer & Chemical Engineering*, 1994, 18(4): 267-302.
- [4] BAKSHI B R, STEPHANOPOULOS G. Representation of process trends, part IV: induction of real-time patterns from operating data for diagnosis and supervisory control [J]. *Computer & Chemical Engineering*, 1994, 18(4): 303-332.
- [5] RENGASWAMY R, VENKATASUBRAMANIAN V. A syntactic pattern recognition approach for process monitoring and fault diagnosis [J]. *Engineering Applications of Artificial Intelligence*, 1995, 8(1): 35-51.
- [6] RENGASWAMY R, HAGGLUND T, VENKATASUBRAMANIAN V. A qualitative shape analysis formalism for monitoring control loop performance [J]. *Engineering Applications of Artificial Intelligence*, 2001, 14(1): 23-33.
- [7] DASH S, RENGASWAMY R, VENKATASUBRAMANIAN V. Fuzzy-logic based trend classification for fault diagnosis of chemical processes [J]. *Computers and Chemical Engineering*, 2003, 27(3): 347-362.
- [8] DASH S. Data-driven qualitative and model-based quantitative approaches to fault diagnosis [D]. West Lafayette: Purdue University, 2001.
- [9] CHARBONNIER S, GARCIA-BELTAN C, CADET C, et al. Trends extraction and analysis for complex system monitoring and decision support [J]. *Engineering Applications of Artificial Intelligence*, 2005, 18(1): 21-36.
- [10] SCALI C, GHELARDONI C. An improved qualitative shape analysis technique for automatic detection of valve stiction in flow control loops [J]. *Control Engineering Practice*, 2008, 16(12): 1501-1508.
- [11] MAURYA R M, PARITOSH P K, RENGASWAMY R. A frame work for on-line trend extraction and fault diagnosis [J]. *Engineering Application of Artificial Intelligence*, 2010, 23(6): 950-960.
- [12] 张海洋, 张贝克, 吴重光. 用于监测复杂系统的计算机辅助定性趋势分析 [J]. *计算机工程与应用*, 2007, 43(17): 41-44.
- [13] 张贝克, 张海洋, 马昕. 定性趋势分析在过程故障诊断中的应用研究 [J]. *系统仿真学报*, 2008, 20(10): 2750-2753.
- [14] 宋政辉, 税爱社, 李明. 基于模糊定性趋势分析的故障诊断方法研究 [J]. *计算机测量与控制*, 2009, 17(11): 2164-2169.
- [15] 张静哲, 杜文莉, 钱锋. 基于时序过程片段分析的符号有向图实时故障诊断方法 [J]. *计算机与应用化学*, 2010, 27(10): 1375-1379.
- [16] 陈骏平, 税爱社, 宋政辉, 等. 一种改进的区间半分定性趋势提取方法 [J]. *仪表技术*, 2011(2): 57-60.
- [17] 邹莹, 梁旭, 黄易. 基于定性趋势分析的泡沫浮选故障诊断 [J]. *中国科技博览*, 2012(33): 595.
- [18] 高东, 张贝克, 吴重光. 基于提升格式的过程数据在线去噪方法及其应用 [J]. *计算机应用研究*, 2008, 25(10): 3198-3220.
- [19] 量 [J]. *管理世界*, 2010, 26(1): 133-141, 161.
- [11] 杨清香, 俞麟, 陈娜. 董事会特征与财务舞弊——来自中国上市公司的经验证据 [J]. *会计研究*, 2009, 30(7): 64-70.
- [12] 王斌, 梁欣欣. 公司治理、财务状况与信息披露质量——来自深交所的经验证据 [J]. *会计研究*, 2008, 29(3): 31-38.
- [13] 于晓强, 刘善存. 治理结构与信息披露违规行为——来自我国 A 股上市公司的经验证据 [J]. *系统工程*, 2012, 30(6): 43-52.
- [14] 樊纲, 王小鲁, 朱恒鹏. 中国市场化指数——各地区市场化相对进程 2011 年报告 [M]. 北京: 经济科学出版社, 2011: 1-100.
- [15] 郭泳植. 克劳修斯讲的熵的故事 [M]. 昆明: 云南教育出版社, 2012.
- [16] 张继国, 辛格. 信息熵——理论与应用 [M]. 北京: 中国水利水电出版社, 2012.
- [17] 李特, 冯琦, 张翌. 基于熵权灰色关联和 D-S 证据理论的威胁评估 [J]. *计算机应用研究*, 2013, 30(2): 380-382.
- [18] 汪妮, 方正, 解建仓. 改进的熵权法在再生资源价值评价中的应用 [J]. *西安理工大学学报*, 2012, 28(4): 416-420.
- [19] 王清源, 潘旭海. 熵权法在重大危险源应急救援评估中的应用 [J]. *南京工业大学学报: 自然科学版*, 2011, 33(3): 87-91.
- [20] HSU L C. Investment decision making using a combined factor analysis and entropy-based topsis model [J]. *Journal of Business Economics and Management*, 2013, 14(3): 448-466.
- [21] 任政亮, 徐飞, 徐红年. 基于熵权信息披露质量的测度: 算法与实证 [J]. *管理现代化*, 2013, 35(3): 10-12.
- [22] 张近乐, 任杰. 熵理论中熵及熵权计算式的不足与修正 [J]. *统计与信息论坛*, 2011, 26(1): 3-5.
- [23] 杨红, 杨淑娥, 张栋. 基于熵理论的上市公司信息披露质量测度 [J]. *系统工程*, 2007, 25(9): 16-21.

(上接第 1440 页)体系的合理选择和更多科学方法的引入和运用方面,因此,如何通过进一步改进修正熵权算法或引入更科学、合理的方法,并将其与信息披露质量评价指标的选择进行紧密结合,从而使得评价结果更具说服力,是未来需要深入研究的方向。

参考文献:

- [1] 张宗新. 上市公司信息披露质量与投资者保护 [M]. 北京: 中国金融出版社, 2009: 1-23.
- [2] 黄方亮, 武锐. 证券市场信息披露质量: 基于内容分析法的研究框架 [J]. *江苏社会科学*, 2011, 32(6): 32-37.
- [3] 李忠. 中国上市公司信息披露质量研究: 理论与实证 [M]. 北京: 经济科学出版社, 2012: 1-146.
- [4] 高明华. 中国上市公司信息披露指数报告 (2012) [M]. 北京: 经济科学出版社, 2012: 24-26.
- [5] 陈君兰, 谢赤. 上市公司信息披露质量测度与评价 [J]. *证券市场导报*, 2013, 23(3): 25-30.
- [6] 屈文洲, 蔡志岳. 我国上市公司实证研究的动因实证研究 [J]. *中国工业经济*, 2007, 24(4): 96-103.
- [7] 任政亮, 徐飞, 徐红年. 投资者视角下信息披露质量的测度: 基础与框架 [J]. *现代管理科学*, 2013, 34(5): 3-5.
- [8] 任政亮, 徐飞, 徐红年. 改进熵权模型在信息披露质量测度中的应用——来自深交所的实证分析与检验 [J]. *软科学*, 2014, 28(1): 135-139.
- [9] 吴国萍, 黄政. 信息披露违规——公司治理与投资者保护 [M]. 北京: 科学出版社, 2012: 59-131.
- [10] 伊志宏, 姜付秀, 秦义虎. 产品市场竞争、公司治理与信息披露质