

文章编号:1007-6735(2012)04-0394-05

基于 DEA 和决策树的我国工业劳动生产率 影响因素规则提取

陶 洪^{1,2}, 徐福缘²

(1. 绍兴文理学院 经济与管理学院, 绍兴 312000; 2. 上海理工大学 管理学院, 上海 200093)

摘要: 利用数据包络分析将影响劳动生产率增长的因素分解为技术效率、纯技术进步、人均资本的规模效率、资本深化 4 个方面,并结合分类与规则提取的决策树方法,揭示了各因素影响劳动生产率变动的规律;然后结合我国工业统计数据,分析了我国省域工业劳动生产率的变动特征. 研究显示对我国省域劳动生产率差异影响最大的指标是资本深化,其次是技术效率和人均资本的规模效率,而技术进步基本不对省域劳动生产率差异产生影响.

关键词: 劳动生产率; 数据包络分析; 决策树算法; 人均资本

中图分类号: F 404 **文献标志码:** A

The Rule Extraction for China Industrial Labor Productivity Based on DEA and Decision Tree

TAO Hong^{1,2}, XU Fu-yuan²

(1. School of Economy and Management, Shaoxing University, Shaoxing 312000, China ;
2. Business School, University of Shanghai for Science and Technology, Shanghai 200093, China)

Abstract: Based on data envelope analysis (DEA), the labor productivity was decomposed into 4 aspects: technical efficiency, pure technical progress, changes of scale efficiency of capital per labor and capital deepening. The rule extraction based on decision tree classification algorithms was used to get the effects of labor productivity's affecting factors and to study how to affect provincial industrial labor productivity. The research finds that the most important factor affecting provincial industrial labor productivity is the capital deepening. The second factor is the technical efficiency, and the next factor is the scale efficiency of capital per labor. Pure technical progress almost doesn't affect the difference between provincial industrial labor productivities.

Key words: labor productivity; data envelope analysis; decision tree algorithm; capital per labor

收稿日期: 2011-04-25
基金项目: 浙江省哲学社会科学规划重点资助项目(08CGJJ001Z); 国家自然科学基金资助项目(71171135);
上海市重点学科建设资助项目(S30504)
作者简介: 陶 洪(1971-),男,副教授.研究方向:工业经济学. E-mail: taohongem@163.com

劳动生产率是指劳动者在一定时期内创造的劳动成果与其相适应的劳动消耗量的比值. 劳动生产率因计算简便、含义直观, 以其固有的优势一直在经济管理的理论和实践领域保持着重要的地位. 劳动生产率作为衡量劳动生产效率的方法之一, 不仅可以反映技术进步所造成的效率提高, 也可以衡量其它替代生产因素对生产效率的影响. 劳动生产率既是衡量组织竞争力的一项重要经济统计指标, 又可帮助决策者分析经济状况、预估所需劳动力以及了解促进组织成长的方法. 但劳动生产率指标又是一项非常综合的指标, 包含了丰富的经济信息, 通过对影响劳动生产率变动的因素进行分类和规则提取, 有助于更好地把握劳动生产率的增长规律.

1 理论基础

1.1 劳动生产率影响因素的 DEA 分析

设 $D_1(\mathbf{X}_0, \mathbf{Y}_0)$ 为规模报酬不变的相对效率, 则规模报酬不变的 C^2R 模型为^[1]

$$D_1(\mathbf{X}_0, \mathbf{Y}_0) = \min \theta$$

$$\text{s. t. } \begin{cases} \sum_{j=1}^n \lambda_j x_j \leq \theta \mathbf{X}_0 \\ \sum_{j=1}^n \lambda_j y_j \geq \mathbf{Y}_0 \\ \lambda_j \geq 0 \end{cases} \quad (1)$$

式中, θ 是相对效率; $\mathbf{X}_0$ 是投入向量; $\mathbf{Y}_0$ 是产出向量.

设 $D_2(\mathbf{X}_0, \mathbf{Y}_0)$ 为规模报酬可变的相对效率, 则规模报酬可变的 BC^2 模型为^[2]

$$D_2(\mathbf{X}_0, \mathbf{Y}_0) = \min \theta$$

$$\text{s. t. } \begin{cases} \sum_{j=1}^n \lambda_j x_j \leq \theta \mathbf{X}_0 \\ \sum_{j=1}^n \lambda_j y_j \geq \mathbf{Y}_0 \\ \sum_{j=1}^n \lambda_j = 1 \\ \lambda_j \geq 0 \end{cases} \quad (2)$$

Kumar 等^[3]利用规模报酬不变模型, 将劳动生产率变动分解为技术进步、技术效率和资本深化 3 方面的影响因素.

设由 C^2R 模型确定的前沿面为纯前沿生产面, 由 BC^2 模型确定的为前沿生产面. t 时期的前沿生产面和纯前沿生产面分别为 f_t, g_t ; $t+1$ 时期的前

沿生产面和纯前沿生产面分别为 f_{t+1}, g_{t+1} (如图 1). 从 g_t 到 g_{t+1} 表明从 t 时期到 $t+1$ 时期所有决策单元不考虑规模报酬变化情况下, 生产技术水平提高对生产效率的影响.

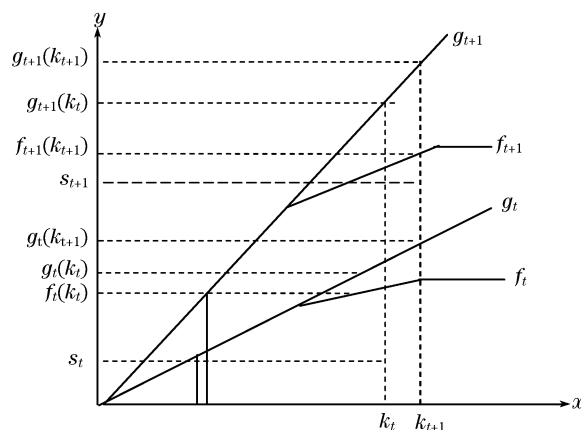


图1 劳动生产率四重分解

Fig.1 Quadruplicity decomposed to labor productivity

设 k 表示人均资本, t 期人均资本为 k_t , $t+1$ 时期人均资本为 k_{t+1} , t 期和 $t+1$ 期对应的实际劳动生产率、前沿劳动生产率和纯前沿的劳动生产率分别为 $s_t, f_t(k_t), g_t(k_t)$ 和 $s_{t+1}, f_{t+1}(k_{t+1}), g_{t+1}(k_{t+1})$. $g_{t+1}(k_t)$ 表示 $t+1$ 时期人均资本为 k_t 时, 其纯前沿的劳动生产率; $g_t(k_{t+1})$ 为 t 期人均占有资本为 k_{t+1} 时, 其纯前沿的劳动生产率. 如果以 s_t 为参照, s_t 和 $f_t(k_t)$ 的差距为技术效率, 这指标反映了该决策单元和人均资本占有量相同的决策单元对现有技术利用能力的差异. 技术效率主要表现为管理能力、对新技术的运用能力和资源的配置能力. 某一决策单元从 t 期到 $t+1$ 期劳动生产率的变化率为 $\frac{s_{t+1}}{s_t}$, 记为 LP . 技术效率的变化对劳动

生产率的影响为 $\frac{s_{t+1}}{f_{t+1}(k_{t+1})} \times \frac{f_t(k_t)}{s_t}$, 记为 TP . 该指标大于 1, 表明可通过挖掘生产潜力提高生产效率. 规模效率对生产效率的影响是 $\frac{f_{t+1}(k_{t+1})}{g_{t+1}(k_{t+1})} \times \frac{g_t(k_t)}{f_t(k_t)}$, 记为 SP . 该指标度量在 t 期和 $t+1$ 期该决策单元人均资本占有量达到最佳人均资本占有量时对生产效率的影响. 处于 k_t 和 k_{t+1} 的纯技术进步变化率分别是 $\frac{g_{t+1}(k_{t+1})}{g_t(k_{t+1})}$ 和 $\frac{g_{t+1}(k_t)}{g_t(k_t)}$, 其几何平均是 $\left[\frac{g_{t+1}(k_{t+1})}{g_t(k_{t+1})} \times \frac{g_{t+1}(k_t)}{g_t(k_t)} \right]^{\frac{1}{2}}$, 记为 TC . 该项指

标度量从 t 期到 $t+1$ 期技术变化对生产效率的影响,大于1,说明整体的技术水平得到提升并促进了生产效率的提高.对纯前沿生产面 g_t ,人均资本由 k_t 变为 k_{t+1} ,劳动生产率的变化为 $\frac{g_t(k_{t+1})}{g_t(k_t)}$.对纯前沿生产面 g_{t+1} ,人均资本由 k_t 变为 k_{t+1} ,劳动生产率的变化为 $\frac{g_{t+1}(k_{t+1})}{g_{t+1}(k_t)}$.资本深化对劳动生产率影响的几何平均为 $\left[\frac{g_t(k_{t+1})}{g_t(k_t)} \times \frac{g_{t+1}(k_{t+1})}{g_{t+1}(k_t)} \right]^{\frac{1}{2}}$,记为 KP .该项指标度量了在 t 期和 $t+1$ 期人均资本的变化对生产效率的影响.由此得到劳动生产率的四重分解式

$$L_i(x^{t+1}, y^{t+1}; x^t, y^t) = \frac{s_{t+1}}{s_t} = \left[\frac{s_{t+1}}{f_{t+1}(k_{t+1})} \times \frac{f_t(k_t)}{s_t} \right] \times \left[\frac{f_{t+1}(k_{t+1})}{g_{t+1}(k_{t+1})} \times \frac{g_t(k_t)}{f_t(k_t)} \right] \times \left[\frac{g_{t+1}(k_{t+1})}{g_t(k_{t+1})} \times \frac{g_{t+1}(k_t)}{g_t(k_t)} \right]^{\frac{1}{2}} \times \left[\frac{g_t(k_{t+1})}{g_t(k_t)} \times \frac{g_{t+1}(k_{t+1})}{g_{t+1}(k_t)} \right]^{\frac{1}{2}} \quad (3)$$

1.2 分类与规则提取的决策树算法

决策树算法是一种归纳分类算法,它通过对训练集的学习,挖掘出有用的规则. C4.5 是目前广泛使用的决策树算法,该算法使用信息增益比例的概念,以信息增益比例作为选择属性的标准,在每个节点上选择具有最高信息增益比例的属性对数据集进行划分.这种理论方法使得对一个对象分类所需的期望测试数目最小,以确保找到一棵简单的树^[4].

设 S 是 s 个数据样本的集合.假定类标号属性具有 m 个不同值,定义 m 个不同类 $C_i (i=1, \dots, m)$, 设 s_i 是类 C_i 中的样本数. 对一个给定的样本分类所需的期望信息为

$$I(s_1, s_2, \dots, s_m) = - \sum_{i=1}^m p_i \log_2(p_i) \quad (4)$$

其中, p_i 为任意样本属于 C_i 的概率,并用 $\frac{s_i}{s}$ 估计.

设属性 A 具有 v 个不同值 $\{a_1, a_2, \dots, a_v\}$, 可以用属性 A 将 S 划分为 v 个子集 $\{s_1, s_2, \dots, s_v\}$. 其中, s_j 包含 S 中这样一些样本, 它们在 A 具有值 a_j , 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)或期望信息为

$$E(A) = \sum_{j=1}^v \frac{s_{1j} + \dots + s_{mj}}{s} I(s_1, s_2, \dots, s_m) \quad (5)$$

$\frac{s_{1j} + \dots + s_{mj}}{s}$ 是第 j 个子集的权, 对于给定的子集 s_j ,

$$I(s_1, s_2, \dots, s_m) = - \sum_{i=1}^m p_{ij} \log_2(p_i) \quad (6)$$

其中, $p_{ij} = \frac{s_{ij}}{s_j}$ 是 s_j 中样本属于类 C_i 的概率.

在 A 上分枝将获得的编码信息, 即信息增益为

$$G(A) = I(s_1, s_2, \dots, s_m) - E(A) \quad (7)$$

$G(A)$ 是由于知道属性 A 的值而导致的熵的期望压缩.

一个属性的增益比例为

$$GR(A) = \frac{G(A)}{SI(A)} \quad (8)$$

其中, $SI(A) = - \sum_{j=1}^v P_j \log_2(P_j)$.

2 实证结果分析

2.1 数据的选择与处理

研究的产出数据用 1988~1992 年工业净产值作为产出指标, 1992 年之后选择了工业增加值. 由于是相邻两年指标的比, 《中国统计年鉴 1993》提供了 1992 年工业净产值和工业增加值的数据, 故 1991~1992 年的变化率按照这两年的工业净产值计算, 而 1992~1993 年的变化率按照这两年的工业增加值计算. 分析表明前期使用工业净产值和后期使用工业增加值在整个分析年度具有很强的一致性. 为消除价格影响因素, 对 1992~2006 年的工业增加值数据、1988~1992 年的工业净产值数据, 按统计年鉴提供的“工业品出厂价格指数”以 1988 年为基期进行平减.

选择各省的工业从业人员平均人数和固定资本存量作为投入指标, 固定资本存量是以 1988 年为起点, 利用永续盘存法计算而得, 公式为

$$K_{i,t} = (1 - \delta) K_{i,t-1} + P_t I_{i,t} \quad (9)$$

式中, $K_{i,t}$ 和 $K_{i,t-1}$ 表示第 i 个省在 t 年和 $t-1$ 年的固定资本存量; δ 表示固定资产的折旧率, 采用代表几何效率递减的余额折旧法得出其值为 9.6%^[5]; $I_{i,t}$ 表示第 i 个省在第 t 年的固定资产投资; P_t 为固定资产投资在第 t 年的价格指数 (1988 年为 1). 所用数据来自于《中国工业经济统计年鉴》和《中国统计年鉴》.

2.2 省域劳动生产率影响因素的分析

运用式(3)基于 DEA 的劳动生产率四重分解模

型和中国工业统计的省份数据,对中国 30 个省份 1988~2006 年的劳动生产率的变化率进行了解,在计算过程中采用了 DEAP2.1 软件^[6],各省域的平均结果如表 1 所示.

表 1 省域工业劳动生产率的变化率四重分解结果
Tab.1 Annual average of change for Chinese industrial labor productivity

省份	技术效率	规模效率	纯技术进步	资本深化	劳动生产率变化率
北京	1.019	0.991	1.006	1.157	1.152
天津	1.029	0.983	1.006	1.192	1.193
河北	1.014	0.996	1.006	1.144	1.153
山西	1.008	0.996	1.006	1.148	1.151
内蒙古	1.055	0.968	1.006	1.214	1.227
辽宁	1.001	0.986	1.006	1.171	1.153
吉林	1.012	0.983	1.006	1.185	1.180
黑龙江	1.037	0.988	1.006	1.180	1.199
上海	0.998	0.982	1.006	1.158	1.128
江苏	0.990	0.997	1.006	1.162	1.142
浙江	0.975	1.000	1.006	1.151	1.106
安徽	0.980	1.006	1.006	1.173	1.148
福建	1.007	0.999	1.006	1.117	1.113
江西	0.977	1.025	1.006	1.173	1.150
山东	1.020	1.005	1.006	1.109	1.137
河南	1.031	1.010	1.006	1.130	1.152
湖北	0.978	0.977	1.006	1.200	1.139
湖南	1.003	1.013	1.006	1.173	1.179
广东	1.011	1.008	1.006	1.075	1.080
广西	0.974	0.994	1.006	1.155	1.114
海南	0.981	0.969	1.006	1.195	1.124
四川	1.009	0.981	1.006	1.272	1.235
贵州	0.996	0.982	1.023	1.183	1.171
云南	0.992	0.980	1.006	1.165	1.127
西藏	1.009	0.987	1.006	1.205	1.145
陕西	1.006	0.984	1.006	1.188	1.174
甘肃	0.998	0.989	1.006	1.146	1.130
青海	1.003	1.009	1.006	1.183	1.156
宁夏	0.995	0.985	1.016	1.157	1.142
新疆	1.068	0.983	1.007	1.170	1.217

注:由于重庆市 1996 年前的统计数据无法获取,本文将重庆并入四川省分析.

2.3 影响省域劳动生产率变动的决策树建立和提取规则步骤

本研究中将劳动生产率作为分析目标,总结技术进步、技术效率、规模效率、资本深化 4 个因素对它的影响,所以将劳动生产率作为类别属性,其它 4 个属性称为非类别属性,提取劳动生产率变动影响规则.在决策树分析过程中,首先分别对 30 个省劳动生产率、纯技术效率、技术进步、人均资本的规模效率、资本深化这 5 个指标共 18 个跨年度变化率的数据进行加权平均,然后利用决策树提取影响省域劳动生产率的规则.

2.3.1 劳动生产率指标的离散化分类

由于劳动生产率指标是连续数据,首先需要将 30 个省份的工业劳动生产率变动率的平均数据进行离散化处理,本文使用 SPSS 统计软件的层次聚类分析将劳动生产率增长率的数据作离散化处理,结

果见表 2.

表 2 省域工业劳动生产率聚类分析结果
Tab.2 Cluster analysis result of provincial industrial labor productivity

4 类	省份
1	北京、河北、山西、辽宁、上海、江苏、浙江、安徽、福建、江西、山东、河南、湖北、广西、海南、贵州、云南、甘肃、西藏、青海、宁夏
2	天津、湖南、吉林、黑龙江、陕西
3	内蒙古、四川、新疆
4	广东

在用决策树分析软件 Clementine 对劳动生产率影响因素规则提取过程中,分别对 2,3,4,5 类聚类结果给予了分析,发现纯技术进步是对劳动生产率影响最小的因素,几乎不对劳动生产率指标产生影响;而且由 4 个聚类类别增加到 5 个聚类类别,分

类的准确率由 90% 下降到 86.7%, 所以本次分类选取 4 个分类作为规则提取的分类标准.

2.3.2 基于决策树的影响省域劳动生产率变动规则提取结果

根据以上步骤, 在计算过程中运用 SPSS Clementine11 软件, 得到劳动生产率影响因素决策表的一个规则集, 共包括 4 条规则:

a. 中国工业劳动生产率增长率处于高位的省份有 21 个, 占总数的 70%; 处于中上水平的省份有 5 个, 占总数的 16.7%; 处于中下水平的省份有 3 个, 占总数的 10%; 处于低位的省份有 1 个, 占总数的 3.3%.

b. 资本深化是影响省域工业劳动生产率差异最主要的因素. 资本深化以 1.165 作为临界点, 处于临界点以上(包括临界点)和临界点以下的省份各占 50%, 均为 15 个. 而工业劳动生产率增长率处于高位的 21 个省份中, 资本深化大于等于 1.165 的省份有 14 个, 占资本深化大于 1.165 的省份的 93.3%. 另外, 工业劳动生产率变化率处于低位的 1 个省份资本深化也大于等于 1.165. 资本深化小于 1.165 的 15 个省份中, 处于高位的省份有 7 个, 占资本深化小于 1.165 的省份的 46.7%; 处于中上水平的省份有 5 个, 占资本深化小于 1.165 的省份的 33.3%; 处于中下水平的省份有 3 个, 占资本深化小于 1.165 的省份的 20%.

c. 对省域工业劳动生产率差异的影响处于第二位的指标是技术效率. 从决策树分析可以知道, 技术效率只对资本深化小于 1.165 的省份的劳动生产率变化率具有显著影响. 对资本深化小于 1.165 的 15 个省份, 可以按照技术效率是否大于 1.003 进行划分. 资本深化小于 1.165 而技术效率大于等于 1.003 的省份总共有 6 个, 并且全处于劳动生产率增长率的高位. 资本深化小于 1.165, 技术效率也小于 1.003 的省份总共有 9 个, 其中 1 个处于劳动生产率增长率的高位, 5 个处于劳动生产率增长率的中上水平, 3 个处于劳动生产率增长率的中下水平.

d. 对工业劳动生产率影响的最后一个指标是人均资本的规模效率, 并且这一指标只对资本深化小于 1.165 而且技术效率也小于 1.003 的省份具有显著影响. 满足资本深化小于 1.165 而且技术效率也小于 1.003 的省份共有 9 个, 可以按照人均资本

的规模效率是否大于等于 1.19 分为两类. 人均资本的规模效率大于等于 1.19 的省份有 6 个, 其中有 5 个省份的劳动生产率增长率处于中上水平, 有 1 个省份的劳动生产率增长率处于中下水平. 人均资本的规模效率小于 1.19 的省份有 3 个, 其中有 1 个省份的劳动生产率增长率处于高位, 有 2 个省份的劳动生产率增长率处于中下水平.

3 结 论

利用决策树分类和规则提取方法, 提取了劳动生产率各影响因素对省域工业劳动生产率增长差异的影响、影响方式以及影响规则. 以上研究显示, 中国省域间工业劳动生产率差异的决定因素是投资, 即 1988~2006 年期间, 中国工业增长仍然是依靠外延的扩展, 而非依靠技术、管理等方面提升的内涵式增长. 当然, 由于使用的数据是多年的平均数据, 无法反映时间趋势的变化, 结论的精确性将不可避免受到影响, 但本研究是将决策树算法用于劳动生产率分析的一次有益尝试, 其结论也符合中国省域工业劳动生产率增长现状, 并对揭示中国省域工业劳动生产率增长的规律也具有一定的指导意义.

参考文献:

- [1] Charnes A, Cooper W W, Rhodes E. Measuring the efficiency decision making units[J]. European Journal of Operational Research, 1978, 2(6): 429 - 444.
- [2] Banker R, Charnes A, Cooper W W. Some models for estimating technical and scale inefficiencies in data envelopment analysis[J]. Management Science, 1984, 30(9): 1078 - 1092.
- [3] Kumar S, Russell R R. Technological change, technological catch-up, and capital deepening: relative contributions to growth and convergences[J]. American Economic Review, 2002, 92(3): 527 - 548.
- [4] Han J, Kambr M. 数据挖掘概念与技术[M]. 范明, 孟小峰, 译. 北京: 机械工业出版社, 2003.
- [5] 张军, 吴桂英, 张吉鹏. 中国省际物质资本存量估算: 1952—2000[J]. 经济研究, 2004(10): 35 - 44.
- [6] Coelli T. A guide to DEAP version 2.1: a data envelopment analysis program [R]. Armadillo: University of New England, 1996.