Management Science and Engineering 管理科学与工程, 2016, 5(4), 186-192
Published Online December 2016 in Hans. http://www.hanspub.org/journal/mse
http://dx.doi.org/10.12677/mse.2016.54020

Hans汉斯

Evaluation and Analysis of the Regional Innovation Efficiencies Based on DEA and Decision Tree

Zhizong Chen

School of Economics and Management, Tongji University, Shanghai
Email: czz@tongji.edu.cn

Received: Dec. 3rd, 2016; accepted: Dec. 19th, 2016; published: Dec. 26th, 2016

Abstract

To scientifically and objectively evaluate and analyze the efficiencies of regional innovation systems is a good base to analyze and make policies of regional innovation. Data Envelopment Analysis (DEA) is one of the most used methods for evaluating the efficiency of regional innovation systems (which are referred to Decision Making Units). Decision tree model is a tree structure, which is a process of classification based on feature; it is an important, useful and visual tool of data mining. Combining the advantages of DEA and decision tree model, this paper presents a hybrid model to evaluate and analyze the efficiencies of regional innovation systems using DEA and decision tree method. Based on the evaluation of the efficiencies of 31 regional innovation systems in China, the classification decision tree is established and its role in strategic decision making (or policy making) of regional innovation systems is analyzed.

Keywords

DEA, Decision Tree, Innovation Efficiencies Evaluation, Regional Innovation System

基于DEA和决策树的区域创新效率评价与分析

陈志宗

同济大学经济与管理学院，上海
Email: czz@tongji.edu.cn

文章引用: 陈志宗. 基于 DEA 和决策树的区域创新效率评价与分析[J]. 管理科学与工程, 2016, 5(4): 186-192.
http://dx.doi.org/10.12677/mse.2016.54020

收稿日期：2016年12月3日；录用日期：2016年12月19日；发布日期：2016年12月26日

摘　要

科学、客观地评价与分析区域创新系统的效率是分析和制定创新政策的基础，数据包络分析(DEA)是评价区域创新系统(称为决策单元)相对有效性的常用评价方法。决策树模型是呈树形结构，表示基于特征对实例进行分类的过程，是一种应用广泛的可视化数据挖掘的重要工具。本文将结合DEA模型和决策树模型的各自优点，提出综合运用DEA与决策树方法的评价分析模式，在评价我国31个区域创新系统绩效的基础上，建立分类决策树并分析其在制定区域创新系统的战略决策(或政策制定)中的作用。

关键词

DEA，决策树方法，创新效率评价，区域创新系统

1. 引言

科技创新是经济增长和社会发展的重要源泉。区域创新系统是把创新人力资源和财力资源投入转化为创新产出的经济系统，是我国创新系统的重要组成部分、我国区域经济增长和科技发展的重要基础。客观、科学地评价与分析各区域创新系统的效率，对于明确区域自身优势、制定有效的创新政策，不断提高区域创新能力具有非常重要意义[1]。

由美国著名运筹学专家 Charnes 等人(1978)提出的数据包络分析(Data Envelopment Analysis, DEA)模型[2]是评价同类型投入、产出系统中各决策单元相对有效性的常用评价方法。不少学者已使用 DEA 模型对我国区域创新系统的效率进行了评价研究。刘顺忠、官建成(2002)运用 DEA CCR 模型评价了区域创新系统的规模与技术有效性，并进行创新系统的分类，提出针对性的对策建议[3]。池仁勇、唐根年(2004)运用 DEA 方法分别对我国 30 个行政区进行了技术创新效率的评价，结果显示我国的技术创新效率东部高、西部低[4]。白俊红、江可申、李婧(2009)对我国已有的区域创新系统的创新效率进行了测算，结果显示，我国区域创新效率普遍偏低，这是纯技术效率低下导致的，并呈现出规模报酬递减的态势[5]。

然而，DEA 模型是高维(多投入、多产出)抽象的数学模型，区域创新系统的决策者需要利用历史数据，对包括 DEA 在内的各种数量方法的分析结果进行数据挖掘，以便更好地因地制宜制定区域创新系统的政策与方针，其中，决策树方法是一种应用广泛的可视化数据挖掘的重要工具，主要有 Quinlan 提出的 ID3 和 C4.5 以及 Breiman 等人提出的 CART 算法[6]。决策树模型是呈树形结构，表示基于特征对实例进行分类的过程，其主要优点是模型具有可读性，可产生意义明确的决策规则，分类速度快和分类准确性高。

结合 DEA 模型和决策树方法一并分析具有吸引力，已有学者在 IT、企业规模等领域做了相关研究[7][8]，但在区域创新绩效评价领域，据本文作者所知，尚未有相应的研究文献。本文将结合 DEA 模型和决策树模型的各自优点，提出综合运用 DEA 模型与决策树方法的评价分析模式，对我国 31 个区域创新系统的效率做客观、科学的评价与分析。本研究的主要目标包括：在评价我国 31 个区域创新系统效率的基础上，分析分类决策树方法在制定区域创新系统的战略决策(或政策制定)中的作用。

2. 综合 DEA 与决策树方法的评价分析模式

区域创新系统效率的综合评价分析模式包括两个过程：首先，应用 DEA 模型，把区域创新系统分为

有效和无效两个类别；其次，使用决策树方法构建分类决策树，提取重要特征变量并形成可知识化的决策规则。下面具体讨论这两个过程中所使用的模型和方法。

2.1. DEA 评价模型

假设 n 个区域创新系统，称为决策单元(decision making unit, DMU)，每个决策单元有 m 种投入和 s 种产出，第 j 个决策单元 DMU_j 的投入和产出分别为 $\left(x_{1j}, x_{2j}, \cdots, x_{mj}\right)$ 和 $\left(y_{1j}, y_{2j}, \cdots, y_{sj}\right)$，$j=1,\cdots,n$。决策单元 DMU_0 的相对效率可以通过下列分式规划得到：

$$\begin{aligned} &\max \quad \theta_0 = \frac{\sum_{r=1}^{s} u_r y_{r0}}{\sum_{i=1}^{m} v_i x_{i0}} \\ &s.t. \quad \theta_j = \frac{\sum_{r=1}^{s} u_r y_{rj}}{\sum_{i=1}^{m} v_i x_{ij}} \le 1, \quad j=1,\cdots,n \\ &u_r, v_i \ge \varepsilon, \quad r=1,\cdots,s; i=1,\cdots,m. \end{aligned} \tag{1}$$

其中 DMU_0 为被评价的 DMU，$v_i\left(1,\cdots,m\right)$ 和 $u_r\left(1,\cdots,s\right)$ 是决策变量，ε 是非 Archimedes 无穷小。利用 Charnes-Cooper 变换，可将分式规划模型(1)化为如下等价的线性规划模型：

$$\begin{aligned} &\max \theta_0 = \frac{\sum_{r=1}^{s} u_r y_{r0}}{\sum_{i=1}^{m} v_i x_{i0}} \\ &s.t. \theta_j = \frac{\sum_{r=1}^{s} u_r y_{rj}}{\sum_{i=1}^{m} v_i x_{ij}} \le 1, \; j=1,\cdots,n \\ &u_r, v_i \ge \varepsilon, r=1,\cdots,s; i=1,\cdots,m. \end{aligned} \tag{2}$$

如果线性规划问题(2)存在最优解 $u_r^*\left(1,\cdots,s\right)$ 和 $v_i^*\left(1,\cdots,m\right)$，使得 $\theta_0^*=1$，则 DMU_0 称为 DEA 有效；否则，其称为 DEA 无效(inefficient)。所有 DEA 有效的 DMUs 构成了有效性前沿。

在使用 DEA 模型识别出具体的有效与无效决策单元之后，可进入决策树模型分析阶段。

2.2. 决策树方法

在分类问题中，设 N 个样本(即决策单元)的数据集 S 具有 I 个特征 A (即 DEA 模型中的投入与产出变量)，以及取 K 个值(例如：DEA 有效、无效等)的分类变量 C，分类决策树方法的目的是找出一颗决策树，选择重要的特征，由其值来预测分类变量的值。

分类决策树算法假设决策树是二叉树结构。决策树由结点和有向边(左侧分支和右侧分支)组成。结点有两种类型：内部结点(internal node)和叶结点(leaf node)。内部结点表示一个特征，其取值为“是”和“否”，左侧分支是取值为“是”的分支，右侧分支是取值为“否”的分支，叶结点表示一个类。这样，决策树等价于递归地二分每个特征，将输入空间(即特征空间)划分为有限个区域，并在这些区域上确定预计的概率分布。

分类决策树用最小化不纯(impurity)指标来选择最优特征，同时决定该特征的最优二值切分点。设 n_{ik} 是结点 i 内第 k 类的样本数目，p_{ik} 是结点 i 内第 k 类中的样本比例，D_i 是结点 i 的不纯度的测度指标，有几种选择可用来计算不纯指标，其具体计算公式如下：

1) 偏差(deviance)：

$$D_i = -\sum_{k=1}^{K} n_{ik} \log p_{ik} \tag{3}$$

2) 熵(entropy)：

$$D_i = -\sum_{k=1}^{K} p_{ik} \log p_{ik} \tag{4}$$

3) 基尼指数(Gini index)：

$$D_i = 1 - \sum_{k=1}^{K} p_{ik}^2 \tag{5}$$

通常可同时使用这几种不纯指标来构建分类决策树，并采用误分率(misclassification rate)最小作为最好分类决策树的选择标准。

根据现有数据集 S，分类决策树的生成过程从根结点开始，递归地对每个结点进行如下操作，构建二叉决策树：

步骤 1：设结点的数据集为 S，计算现有特征对该数据集的不纯指标。此时，对每一个特征 A，对其可能的每个取值 a，根据样本点 $A = a$ 的测试为“是”或“否”将 S 分割成 S_1 和 S_2 两部分，利用公式(3)~(5)计算 $A = a$ 时的不纯指标；

步骤 2：在所有可能的特征 A 以及它们所可能的切分点 a 中，选择不纯指标最小的特征及其对应的切分点作为最优特征与最优切分点。依最优特征与最优切分点，从现结点生成两个子结点，将数据依特征分配到两个子结点中去；

步骤 3：对两个子结点递归地调用步骤 1 和步骤 2，直至满足停止条件；

步骤 4：生成分类决策树。

算法停止计算的条件是结点中的样本个数小于预定阈值，或样本集的不纯指标小于预定的阈值，或者没有更多特征。

构建的分类决策树提取了 DEA 评价的原始数据中的重要特征变量，并由此可形成直观的、可知识化的决策规则。

3. 实例分析

首先使用 DEA 模型对中国大陆 31 个区域创新系统(省、自治区和直辖市)的 2012~2013 年的科技创新的效率进行评价。吴和成、刘思峰(2007)运用主成分法和相关分析等统计方法，筛选并建立了区域 R&D 相对效率的 DEA 评价指标体系[9]。借鉴该研究成果，本文 DEA 评价模型所使用科研与创新活动的具体投入包括：

x_1：2012 年区域创新系统 R&D 支出(亿元)；

x_2：2012 年地方财政科技拨款(亿元)；

x_3：2012 年区域创新系统 R&D 人员(千人年)。

同时，在对区域创新系统进行 DEA 相对有效性评价时，应当考虑从投入到产出的延迟时间，本文假定该延迟时间为一年。因此，DEA 模型所使用科研与创新活动的具体产出考虑为：

y_1：2013 年国内中文期刊科技论文数(篇)；

y_2：2013 年高科技产业主营业务收入(亿元)；

y_3：2013 年高技术产品出口额(百万美元)；

y_4：2013 年发明专利申请授予量(项)。

具体的 DEA 投入与产出数据见表 1 的第 2 列至第 8 列，所有数据均来源于中国科技部发展计划司所发布的科技创新统计资料汇编[10] [11] [12]。

使用 DEA 软件 EMS 对我国 31 个区域创新系统的相对效率值 θ 进行计算，其结果列在表 1 的第 9 列。其中，有效决策单元集合为 **E** = {北京，上海，江苏，江西，广东，海南，重庆，陕西，甘肃，新疆}，

Table 1. DEA Input and Output Data and Efficiencies of regional innovation systems
表 1. 区域创新系统 DEA 投入与产出数据及相对效率

决策单元	x_1	x_2	x_3	y_1	y_2	y_3	y_4	θ
北京	1063.4	199.9	235.5	72,662	3826.1	20,354	20,695	1
天津	360.5	76.5	89.6	14,976	4243.5	19,289	3141	0.792
河北	245.8	44.7	78.5	21,543	1381.0	2811	2008	0.756
山西	132.4	33.3	47.0	8005	707.8	3228	1332	0.563
内蒙古	101.5	27.6	31.8	4217	344.8	110	549	0.348
辽宁	390.9	101.2	87．2	20,976	2362.4	5430	3830	0.692
吉林	109.8	25.0	50.0	9580	1431.3	386	1496	0.854
黑龙江	146.0	37.6	65.1	13,229	610.8	296	2238	0.812
上海	679.5	245.4	153.4	31,803	6823.4	88,710	10,644	1
江苏	1287.9	257.2	401.9	52,623	24,854.0	127,965	16,790	1
浙江	722.6	166.0	278.1	27,629	4360.1	14,276	11,139	0.787
安徽	281.8	96.0	103.0	14,310	1831.4	2826	4241	0.698
福建	271.0	48.5	114.5	9820	3545.0	15,527	2941	0.720
江西	113.7	27.5	38.2	7182	2289.6	3442	923	1
山东	1020.3	125.0	254.0	25,849	8946.5	17,394	8913	0.784
河南	310.8	69.6	128.3	21,466	4284.4	20,726	3173	0.792
湖北	384.5	54.4	122.7	27,779	2445.3	5209	4052	0.735
湖南	287.7	48.2	100.0	16,814	2564.9	1660	3613	0.785
广东	1236.2	246.7	492.3	36,277	27,871.1	256,431	20,084	1
广西	97.2	42.8	41.3	11,512	1126.2	1942	1295	0.827
海南	13.7	12.1	6.8	3284	121.4	571	449	1
重庆	159.8	29.8	46.1	14,809	2624.2	24,836	2360	1
四川	350.9	59.4	98.0	24,843	5160.5	19,217	4566	0.953
贵州	41.7	29.0	18.7	6037	372.0	154	776	0.756
云南	68.8	32.7	27.8	8502	291.1	2019	1312	0.804
西藏	1.8	5.1	1.2	302	11.8	56	44	0.746
陕西	287.2	34.9	82.4	30,055	1374.0	4739	4133	1
甘肃	60.5	16.2	24.3	9485	140.9	242	785	1
青海	13.1	7.2	5.2	1504	50.7	24	91	0.644
宁夏	18.2	9.6	8.1	2250	31.8	129	184	0.645
新疆	39.7	33.0	15.7	8034	20.7	331	540	1

数据来源：x_1、x_2 和 x_3 来源于文献[10]；y_1 和 y_4 来源于文献[11]；y_2 和 y_3 来源于文献[12]。

其构成了 DEA 模型中的有效性前沿，而无效决策单元集合 **NE** = {天津，河北，山西，内蒙古，辽宁，吉林，黑龙江，浙江，安徽，福建，山东，河南，湖北，湖南，广西，四川，贵州，云南，西藏，青海，宁夏}，是 DEA 模型识别出的无效区域创新系统。

为了从 DEA 评价的原始数据中识别出重要的特征变量，并产生直观的、有意义的决策规则，有必要进一步构建分类决策树模型。把 DEA 相对效率值 θ 转换为取值为“有效($\theta=1$)”和“无效($\theta<1$)”的分类变量，并将 DEA 模型的投入与产出变量作为特征，使用 R 语言 tree 软件包，采用最小偏差作为分割结点的准则，建立分类决策树如图 1 所示。

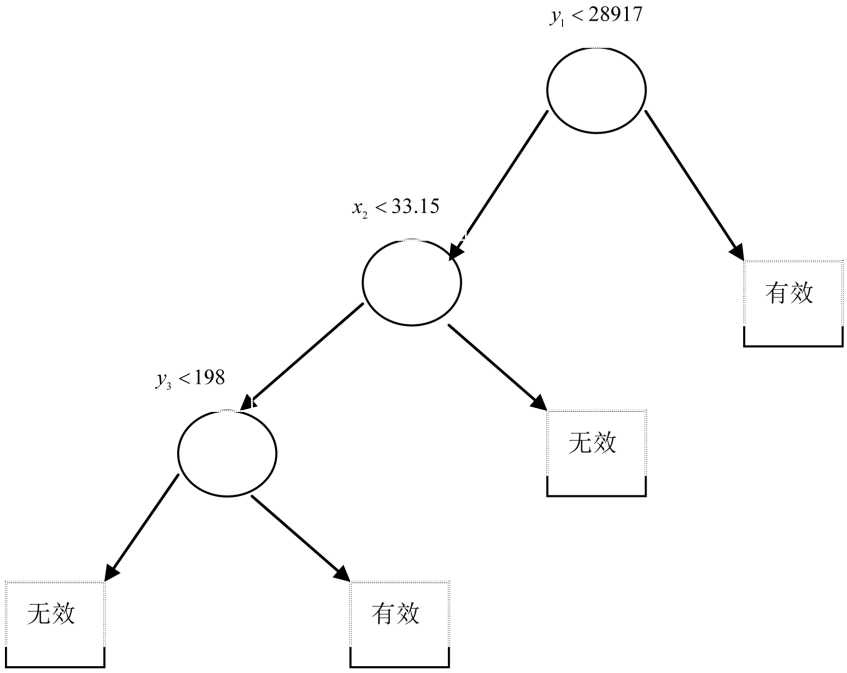

Figure 1. Classification tree
图 1. 分类决策树

从图 1 的分类决策树可见，该决策树抽取了 DEA 投入与产出 7 个指标中的 3 个重要的特征 y_1、x_2 和 y_3，并产生如下 4 条判别决策单元是否 DEA 有效的决策规则。

规则 1：IF ($y_1 > 28917$)，THEN (决策单元 DEA 有效)。

规则 2：IF ($y_1 < 28917$ 且 $x_2 > 33.15$)，THEN (决策单元 DEA 无效)。

规则 3：IF ($y_1 < 28917$ ， $x_2 < 33.15$ 且 $y_3 > 198$)，THEN (决策单元 DEA 有效)。

规则 4：IF ($y_1 < 28917$ ， $x_2 < 33.15$ 且 $y_3 < 198$)，THEN (决策单元 DEA 无效)。

该分类决策树的误分率为 $0.65\left(=2/31\right)$，也即正确的分类比例达到 93.5%，表明提取的决策规则是相当有效的。

相对于 DEA 是高维度(多投入、多产出)抽象的数学模型，决策树提供了直观和意义明确的决策规则，这些决策规则对区域创新系统的政策制定者提供了有用的信息，帮助他们进行区域创新系统有效性的影响因素分析及定位决策。只有明确了影响区域创新系统有效性的最重要因素，各区域创新系统才能根据自身的资源条件，制定出正确的区域创新政策，充分发挥自身的优势，努力克服或改善自身的不足，在我国各地生机勃勃的科技创新活动中不断改进，力争上游。

4. 结论

区域创新系统是我国创新系统的重要组成部分、我国区域经济增长和科技发展的重要基础。DEA 是评价区域创新系统(决策单元)相对有效性的最常用评价方法之一。由于 DEA 是高维度(多投入、多产出)抽象的数学模型，在具体识别出有效决策单元和无效决策单元之后，有必要使用决策树方法对 DEA 的分析结果进行进一步的数据挖掘，构建分类决策树并产生直观和意义明确的决策规则，为区域创新系统的政策制定者提供决策上有价值的信息。实例分析结果表明，本文提出的综合运用 DEA 模型与决策树方法的评价分析模式是有效的。

参考文献 (References)

[1] 李海基, 周霞, 李红. 区域创新系统评价综述[J]. 科技管理研究, 2010, 30(1): 13-14.

[2] Charnes, A., Cooper, W.W. and Rhodes, E. (1978) Measuring the Efficiency of Decision Making Units. *European Journal of Operational Research*, **2**, 429-444. https://doi.org/10.1016/0377-2217(78)90138-8

[3] 刘顺忠, 官建成. 区域创新系统创新绩效的评价[J].中国管理科学, 2002, 10(1): 75-78.

[4] 池仁勇, 唐根年. 基于投入与绩效评价的区域技术创新效率研究[J]. 科研管理, 2004, 25(4): 23-27.

[5] 白俊红, 江可申, 李靖. 中国区域创新系统创新效率综合评价及分析[J]. 科技管理与科技政策, 2009, 21(9): 2-8.

[6] 李航. 统计学习方法[M]. 北京: 清华大学出版社, 2012.

[7] 杨会志, 张雅洁. 基于 DEA 与决策树方法的企业规模收益问题研究[J]. 计算机工程, 2002, 28(1): 220-221.

[8] Wu, D. (2006) Detecting Information Technology Impact on Firm Performance Using DEA and Decision Tree. *International Journal of Information Technology and Management*, **5**, 162-174. https://doi.org/10.1504/IJITM.2006.010116

[9] 吴和成, 刘思峰. 基于改进 DEA 的地域 R&D 相对效率评价[J]. 研究与发展管理, 2007, 19(2): 108-112.

[10] 中国科技部. 中国科技统计数据 2013 [R]. http://www.sts.org.cn/sjkl/kjtjdt/data2013/

[11] 中国科技部. 中国高技术产业数据 2014 [R]. http://www.sts.org.cn/sjkl/gjscy/data2014/data14.pdf

[12] 中国科技信息研究所. 中国主要科技指标数据库——省市主要指标(2008-2013) [EB/OL]. http://www.sts.org.cn/kjnew/maintitle/sub.asp?Main=15.