基于知识挖掘的产业集群竞争力评价指标体系构建

蔡皎洁
（湖北工程学院，湖北 孝感 432000）

摘　要：文章提出了基于知识挖掘方法的产业集群竞争力评价指标体系构建思路，即基于K-中心点聚类法构建评价指标并赋予其影响力权值；基于关联规则挖掘构建评价指标间关系，并引入企业本体等语义处理技术计算指标间关系的权值，使指标体系上升到语义层面。实证研究证明了基于知识挖掘方法所构建的指标体系的有效性。

关键词：知识挖掘；产业集群；评价指标体系；K-中心点聚类；关联规则
中图分类号：F224　　**文献标识码**：A　　**文章编号**：1002-6487(2014)01-0060-03

0　引言

在产业集群竞争力评价研究中，评价指标体系的构建显得尤为重要，它的合理性及全面性直接影响了评价结果的准确率。国内外学者在这方面做出了深入研究，其中波特最早提出了钻石模型[1]，随后该模型被广泛应用于我国早期对产业集群竞争力定性评价中，该模型把社会需求状况、相关的支持企业、要素状况以及企业战略、结构和竞争者等要素作为产业集群评价的定性指标；Padmore和Gibson[2]在钻石模型的基础上提出了GEM模型，该模型把竞争力评价指标分为3组6个指标，分别是基础组，包括资源与设施；企业组，包括供应商和相关辅助产业、企业结构、战略和竞争；市场组，包括当地市场和外部市场。我国学者喻春光和刘友金[3]在GEM模型的基础上提出了GEMN模型，充分考虑了产业集群网络对评价的影响，因此添加了网络组，包括内网和外网两个因素。

从国内外有关产业集群竞争力评价指标体系构建研究现状总结发现：一是多数评价指标体系构建的原理建立在钻石模型和GEM模型基础上，有其合理性；二是多数评价指标体系构建仍缺乏定量化的科学指导和分析，构建方法主要采用人为的、经验性的定性分析，且指标要素繁杂而重复，增加了产业集群竞争力评价的复杂性和冗余性。本文提出了基于知识挖掘方法构建产业集群竞争力评价指标体系的思路，为定量研究开辟了新的研究途径。

1　基于K-中心点聚类挖掘构建竞争力评价指标

1.1　理论基础

K-中心点聚类法是典型的局域划分的聚类方法，其基本的处理流程如下[4]：

（1）对待聚类的文本集D，确定要生成的簇的数目K；

（2）按照某种原则（可随机）生成K个聚类中心作为聚类的初始中心点$S=\{s_1,...s_j,...,s_k\}$；

（3）对D中的每一个文本d_i，依次计算它与各个中心点s_j的相似度$sim(d_i,s_j)$；

（4）选取具有最大的相似度的中心点$\arg\max sim(d_i,s_j)$，将d_i归入以s_j为聚类中心的簇C_j，从而得到D的一个聚类$C=\{c_1,...,c_k\}$；

（5）重新确定每个簇的中心点；

（6）反复执行步骤（3）到（5），直到中心点不再改变，文本不再重新被分配为止。

1.2　构建思路

1.2.1　K的确立

K值的确定会影响聚类结果的精确度，本文将上述K-中心点文本聚类方法进行适当演绎用于产业集群竞争力评价指标的确定中，目的是针对目前大多数指标体系多而杂的情况，找出关键评价指标，确定产业集群边缘。K值确定的思路如下：

（1）将待研究的产业集群中的每个企业看成一个文本d_i，产业群看成文本集D；

（2）在对国内外学者构建产业集群竞争力评价指标体系研究成果总结的基础上，列出一组初始的竞争力评价指标，形成[资源，成长]二维模式，综合对比不同企业在最近5年中各种竞争资源的成长曲线平滑度，并按照这些资源对企业成长影响的大小赋予权值；

（3）计算每个企业在5年内各种资源成长量的加权平均，利用区间划分方法来确定K的值。

1.2.2　初始指标集的聚类

用K-中心点法对初始指标集进行聚类，即完成在K个初始企业簇群区间中，重新实施聚类，并找出影响企业重新聚类的关键因素，这些因素就是最终确定的竞争力评价指标。其基本思路如下：

（1）在K个初始企业簇群区间中，找出作为企业聚类的中心种子$S=\{s_1,...,s_j,...,s_k\}$，$V(s_j)=(W\ s_j(d_1),\cdots,W\ s_j(d_i),\cdots,W\ s_j(d_n))$，其中s代表资源，d代表企业，V代表资源—企业矩阵，W代表资源对企业影响的权值；

（2）对每个资源因子t_i，依次计算它与各个种子s_j的相

似度 $sim(t_i,s_j)$。其中 t_i 和 s_j 之间的相似度可以用向量 $V(t_i)$ 和 $V(s_j)$ 的余弦来计算，其公式[5]描述如下：

$$Sim(t_i,s_j)=\frac{\sum_{i}^{n}W_{t_i}(d_i)*W_{s_j}(di)}{\sqrt{\sum_{i=1}^{n}W_{t_i}(d_i)^2}*\sqrt{\sum_{i=1}^{n}W_{s_j}(d_i)^2}} \quad (1)$$

（3）选取具有最大相似度的种子 $\arg\max sim(t_i,s_j)$，将 t_i 归入以 s_j 为聚类中心的簇 c_j，从而得到初始指标集的聚类集 $C=\{c_1,\cdots,c_j,\cdots c_k\}$；

（4）重新确定每个簇的中心点；

（5）重复步骤（2）、（3）、（4），直到中心点不再改变为止。

1.2.3 关键指标集的确定

该过程其实是对初始指标集聚类不断反复实施的过程。其基本思路如下：

（1）抽取簇 c_j 的中心点和孤立点，并存入关键指标集数据库中；

（2）打乱已形成的聚类集C，重新进行[资源，成长]二维模式的数值评估，重新确定K值；

（3）利用K-中心点法重新实施聚类，把新生成的聚类集中心点和孤立点再抽取到关键指标集数据库中；

（4）重复（1）、（2）、（3）步，直到企业簇群C稳定为止。

2 基于关联规则挖掘构建竞争力评价指标间关系

2.1 理论基础

关联规则挖掘是知识挖掘中最为重要的方法，主要在大量数据项集中发现有意义的关联。其描述如下[6]：

设I是一个项集，其中的元素称为项。设D是事务集，其中每个事务T是项的集合，即 $T\subseteq I$。关联规则的形式可表达为 $X\Rightarrow Y$ 的蕴含式，其中 $X\subset I$，$Y\subset I$，且 $X\cap Y=\Phi$。其中 $support(X\Rightarrow Y)=P(X\cup Y)$ 表示规则 $X\Rightarrow Y$ 在事务集D中的支持度，即D中包含 $X\cup Y$ 的百分比；$confidence(X\Rightarrow Y)=P(Y|X)$ 表示规则 $X\Rightarrow Y$ 在事务集D中的置信度，即D中包含 $X\cup Y$ 的事务与包含X的事务数之比。同时满足大于最小支持度阈值和最小置信度阈值的规则称为强规则；如果项集的出现频率大于或等于最小支持度阈值与D中事务总数的乘积时，则称该项集为频繁项集。

Apriori算法是关联规则挖掘技术的核心方法。其基本思想是将关联规则挖掘分解为两步：

（1）找出所有支持度大于最小支持度的项集，即频繁项集；

（2）使用第（1）步找到的频繁项集产生所期望的规则。

2.2 构建思路

2.2.1 基于Apriori算法发现关键指标间的关系

（1）设I是关键指标集，其中的元素为影响企业成长的关键资源；D为企业集，其中每个事务是由I中元素的子集构成；

（2）设定最小支持度阈值min_sup，找出所有支持度大于min_sup值的项集，确定频繁项集；

（3）在频繁项集中产生有意义的规则，即企业间关键指标间的关联规则。

2.2.2 基于企业本体赋予关键指标间关系权值

指标间关系权值的大小代表着企业相互影响量的多少，其总和代表了产业集群的凝聚力的大小。在计算两指标之间关系权值时，我们需要参照企业本体全局概念体系，考量两指标之间的语义相似度，从纯客观的角度来考虑两指标之间互相影响的大小。本体作为共享的概念模型的形式化的规范说明，已普遍应用于计算机领域。企业本体描述了企业各方面的资源及活动、及其之间复杂网络关系的概念体系。基于企业本体赋予关键指标间关系权值P的方法如下：

（1）统计频繁项集中两指标发生关联的频度N；

（2）参照企业本体概念体系，计算指标 t_1 和 t_2 之间的语义相似度，其公式[7]描述如下：

$$Sim_{Word}(t_1,t_2)=\frac{\alpha\times(l_1+l_2)}{(Dis(t_1,t_2)+\max((l_1-l_2),1)} \quad (2)$$

其中，两指标 t_1 和 t_2 的相似度记为 $Sim_{Word}(t_1,t_2)$，两指标 t_1 和 t_2 在企业本体全局概念体系的语义距离记为 $Dis(t_1,t_2)$，L_1 和 L_2 是 t_1 和 t_2 分别所处的层次，α 是相似度为0.5时 t_1 和 t_2 之间的距离，α 是一个可调节的参数，一般 $\alpha>0$。

（3）两指标间关系权值P记为以下公式：

$$P=N+Sim_{word}(t_1,t_2) \quad (3)$$

3 实证分析

武汉城市圈是指以武汉市为核心、半径为100公里的城市群落，包括武汉市、黄石、孝感、黄冈、咸宁、仙桃、潜江及天门等8个城市。武汉城市圈是湖北品牌服装、医用纺织品、中高档汽车内饰面料、工业用布、家用纺织品等的集中产地和销售地，各项经济指标占全省纺织服务业收益的2/3以上。其中汉川市马口镇形成了纺织、染纱、制线、织布、服装、纺织机械和纺机配件等一条龙产业链；仙桃市澎湖镇聚集了无纺布制品及其相关企业116家；武汉市江汉区和桥口区依托汉正街批发市场形成服装产业集群等[8]。

实证分析的目的是随机从武汉城市圈纺织服务产业集群中抽取130家企业为研究对象，其中包括30家龙头企业，100家涉及供应、生产、配件等中小企业，运用上述提出的K-中心点聚类法、关联规则挖掘等知识挖掘方法和技术，找出影响该产业集群竞争力的关键评价指标和指标间的关系，构建评价指标体系，并与GEM模型进行对比验证该评价指标体系的有效性。

3.1 基于K-中心点聚类挖掘竞争力评价指标

3.1.1 K的确立

首先，在国内外学者关于产业集群竞争力评价指标构建的研究成果基础上，我们选择以下指标作为初始的竞争力评价指标，并形成[资源，成长]二维数据模式，如表1所示：

表1　初始竞争力评价指标

资源		成长
硬性资源	自然资源、劳动力资源、资金资源、公共基础设施、机构设施、企业规模和数量、企业流动率、内地市场规模、国际市场规模、企业净利润、企业销售收入等	增长量
软性资源	人才、企业家素养、供应链的稳定度、企业文化凝聚力、品牌影响力、政府支持度、知识创新力、管理能力等	提升量

其次，根据[资源，成长]二维数据模式，依次画出爱帝、红人、佐美尔等企业在最近5年内的[资源，成长]曲线图，综合评价每个资源对不同企业成长影响度，即曲线平滑度，并赋予影响权值。图1表示的是某企业最近5年内净利润指标对成长影响的曲线图，其权值的大小为曲线中每个顶点夹角余玄之和，即$\cos45^0+\cos15^0+\cos30^0+\cos45^0+\cos(-15^0)=0.87$。同理可以计算其它相关资源对该企业的影响大小，即权值。

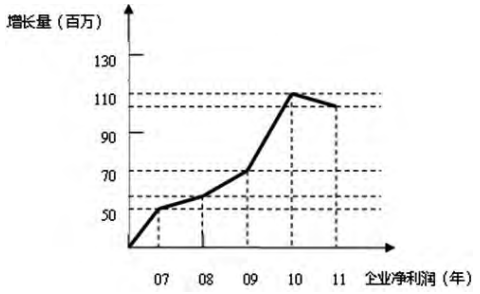

图1 净利润对企业影响力曲线图

最后，计算最近5年内每个企业各种资源增长量的加权平均，经计算分析得出武汉城市圈纺织服装产业大致形成5个数值区间，即定K值为5。

3.1.2　初始指标集的聚类和关键指标集的确立

根据K值，我们按照上述所提出的方法对企业群进行初始聚类，然后按照[资源，成长]曲线平滑度不断移出对聚类结果中心点和边缘点影响较大的资源因素，即为要找的关键指标，直到聚类结果形状不再改变为止。图2是经过CARROT2聚类工具三次重新聚类后形成的结果，可见武汉城市圈的纺织服装产业集群已相当成熟，资源中心的改变对产业集群的影响不大。

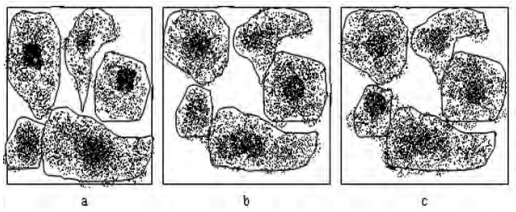

图2 关键指标集确立过程图示

2.1.3　关键指标间关系的确立和权值的赋予

首先，将带有权值的关键指标集整理为有利于关联规则挖掘算法实施的形式，为了表达方便将事务集进行标号，其部分数据集如表2所示：

其次，设min_sup为2，运用Apriori算法从上述事务—项集数据集中发现频繁项集，并在领域专家的指导下选择出有意义的指标间关联规则模式，从中可见指标间关系的产生可来自于企业内部，也可以来自于不同企业中。其部分关联规则模式集如表3所示：

表2　部分事务—项集数据集示例

事务	项集
飞银纺织服装（001号）	厂房(0.83)，供应商(0.54)，机械(0.93)，固定资金(0.97)，劳动力(0.76)，设计者(0.87)，互联网平台(0.32)，知识创新力(0.45)，固定顾客群(0.97)，.....
华成纺织服装（002号）	厂房(0.64)，机械(0.53)，劳动力(0.87)，设计者(0.94)，企业净利润(0.74)，企业家素养(0.91)，供应商(0.32)，互联网平台(0.87)知识创新力(0.97)，品牌影响力(0.93)，.....
惠丰纺织服装（003号）	厂房(0.71)，供应商(0.83)，机械(0.54)，资金周转率(0.78)，劳动力(0.65)，设计者(0.83)，互联网平台(0.67)，知识创新力(0.73)，品牌影响力(0.73)，.....

表3　部分指标间关联规则模式集

指标间关联规则模式集
固定资金(001)→厂房(001)，固定资金(001)→机械(001)，固定资金(001)→互联网平台(001)，固定资金(001)→劳动力(001)，设计者(001)→知识创新力(001)，设计者(001)→固定顾客群(001)，.....
设计者(002)→知识创新力(002)，知识创新力(002)→企业净利润(002)，知识创新力(002)→品牌影响力(002)，企业家素养(002)→知识创新力(002)，互联网平台(002)→知识创新力(002)，.....
资金周转率(003)→机械(003)，资金周转率(003)→厂房(003)，资金周转率(003)→劳动力(003)，设计者(003)→知识创新力(003)，设计者(003)→品牌影响力(003)，互联网平台(003)→品牌影响力(003)，.....
供应商(001)→供应商(002)，供应商(003)→供应商(001)，固定客户群(001)→固定客户群(002)，固定客户群(001)→固定客户群(003)，互联网平台(002)→设计者(001)，品牌影响力(002)→固定客户群(001)，.....

最后，参照企业本体全局概念体系，按照公式3计算指标间关系权值，并用基于本体的知识表示方式来表示，形成一个复杂的带有语义权值标签的网络结构。部分竞争力评价指标体系如图3所示：

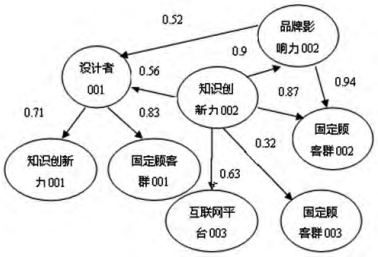

图3 竞争力评价指标间关系示意图

3.2　竞争力评价指标体系有效性检验

本文对基于上述知识挖掘方法构建的竞争力评价指标体系与基于GEM模型的评价指标体系进行有效性对比研究，其基本思路如下：

（1）从武汉城市圈纺织服装产业集群中随机抽取60家企业作为评价数据集，避免与生成评价指标集的企业群训练数据集重复；

（2）分别根据基于上述知识挖掘方法构建的竞争力评价指标体系与基于GEM模型的评价指标体系，对60家企业相对应的数据进行预处理，整理为(0,1)的数据形式；

（3）针对不同的评价指标集，分别对这60家企业评价数据集的相应数据进行神经网络分类，对比分类结果的准确率，并在领域专家的指导下，评价不同指标集对评价集影响的有效性。其准确率表示为分类正确的企业簇占所有分类评价企业簇的大小，其计算公式[9]如下：

$$\Pr ecision=\frac{correct_{clustering}}{all_{clustering}} \quad (4)$$

基于主成分分析法的区域文化产业效率评价

鲁小伟，毕功兵
(中国科学技术大学 管理学院，合肥 230026)

摘　要：文章针对当前应用数据包络分析进行文化产业效率评价指标选取方面的局限性，通过主成分分析法得到文化产业DEA投入产出指标体系，再利用DEA相关模型获得广东等13省市文化产业投入产出效率值，最后对所得结果进行科学的分析和比较研究，以期为各省市提高文化产业效率提供决策支持。

关键词：文化产业；主成分分析法；DEA；效率评价
中图分类号：F224　　**文献标识码**：A　　**文章编号**：1002-6487(2014)01-0063-03

0　引言

近年来，对文化产业效率及其区域竞争力的评价研究越来越受到学者们的重视，他们所使用的评价方法包括了ANP法，结构模型法等。作为研究多投入与多产出系统效率的DEA方法，在文化产业效率评价研究中也逐步得到推广。但学者们对文化产业投入产出指标体系的建立尚未取得统一的意见，这种现状已成为制约DEA在文化产业效率评价方面进一步发展的重要因素。如何建立合理的投入产出指标体系，成为当前这一研究领域重要问题。本文将根据主成分分析法，从统计的角度科学建立文化产业效率评价的指标体系，进而应用DEA模型对中国区域文化产业的运营效率进行评价，以期为文化产业的科学发展提供决策依据。

1.1　数据来源及主成分分析过程

本文沿用文献[1]中2007年广东等13省市文化产业投入产出原始指标和数据，利用主成分分析方法确定新的投入产出指标。在文献[1]中作者具体选用了广东、湖南、广西、福建、上海、河北、吉林、新疆、云南、四川、重庆、山东和安徽等13省市2007年投入产出数据。作者选取的投入产出指标分别是文化、文物单位数(个)，从业人员数量(万人)，财政投入(亿元)，固定资产投入额(亿元)；产出增加值(亿元)，总产出(亿元)，生产税净额(亿元)，营业盈余

基金项目：国家自然科学基金资助项目(71171181)
作者简介：鲁小伟(1988-)，男，安徽安庆人，硕士研究生，研究方向：产业效率评价。
毕功兵(1966-)，男，安徽无为人，博士，副教授，研究方向：管理科学。

经过上述步骤验证了基于知识挖掘方法所构建的竞争力评价指标体系要比基于GEM模型的评价指标体系更为有效。其准确率对比如表4所示：

表4　　基于不同评价指标体系的神经网络分类结果对比

基于不同评价指标体系的神经网络分类	准确率
基于知识挖掘构建的评价指标体系	0.854
基于GEM模型的评价指标体系	0.561

4　结论

(1)本文在国内外研究成果的基础上，运用知识挖掘等先进技术和方法构建指标体系，在定量研究方面有一定的理论创新性。

(2)在指标间关系权值计算中，引入了参照企业本体全局概念体系计算指标间语义相似度，使指标体系的构建上升到语义层次，提高了指标体系构建的准确度。

(3)以武汉城市圈纺织服装产业集群为研究对象进行实证分析，验证了本文所提出基于知识挖掘的竞争力评价指标构建方法，并运用神经网络分类方法与基于GEM模型的指标体系对比了其有效性。

参考文献：
[1]迈克·波特著，郑海燕译.簇群与新竞争经济学[J].经济社会体制比较，2000，(2).
[2]Padmore T.，Gibson H.Modeling Systems of the Innovation: a Framework for Industrial Cluster Analysis in Region[J].Research Policy，1998，(26).
[3]喻春光，刘友金.产业集群竞争力定量评价GEMN模型及其应用[J].系统工程，2008，5(26).
[4]苏新宁等.数据仓库和数据挖掘[M].北京：清华大学出版社，2006.
[5]张玉峰，蔡皎洁.基于数据挖掘的Web文本语义分析与标注研究[J].情报理论与实践，2010，2(33).
[6]唐涛.基于文本挖掘的领域本体学习研究[D].武汉大学博士论文，2009.
[7]刘群，李素键.基于《知网》的词汇语义相似度计算[C].Processing of Computer Linguistics and Chinese Language Processing，2002，(2).
[8]彭继汉.武汉城市圈纺织服装业布局新思路[N].中国纺织报，2012-3-11(10).
[9]杨学明.基于本体学习的个性化网页推荐[J].情报杂志，2009，18(3).

(责任编辑/亦　民)