安徽省高新技术统计关键指标关联性研究

王 俊
（安徽省科学技术情报研究所，安徽合肥 230011）

摘要：在国内外的研究基础上，结合安徽特有的基本情况，根据安徽省“1+6”政策体系，建立了一套高新技术统计指标体系。以最大依赖性、最大相关性和最小冗余为准则建立模型，选择过滤式特征选择方法的代表算法之一mRMR来选择特征子集，在众多指标中抽取关键指标，并利用数据挖掘中聚类分析方法挖掘指标间潜在的关联性，提出高新技术产业增加值和高新技术企业培育情况是影响一个地区高新技术产业运行情况的重要指标。

关键词：高新技术；数据挖掘；关键指标；相关度；安徽省
中图分类号：C813；TP181 **文献标识码：**A **DOI：**10.3772/j.issn.1674-1544.2017.02.013

Research for Correlation with the Statistics Key Indexes of New and High Technology in Anhui Province

WANG Jun
(Scientific and Technological Information Institute of Anhui Province, Hefei 230011)

Abstract: Firstly, on the basis of research at home and abroad, and combining the basic situation of Anhui characteristics, this article establishes a set of index system of new and high technology industries. Secondly, to maximize the dependency, maximum correlation and minimum redundancy for the guidelines, this article establishes a model, chooses mRMR to select feature subset which is one of the representative algorithms of the filter, and extract the key indexes in many indexes. Thirdly, data mining the potential correlation between excavated index using the method of clustering analysis. At last, put forward that it is the added value of new and high technology industries and the enterprises which affect the high and new technology industry.

Keywords: new and high technology, data mining, key indexes, relativity, Anhui province

高新技术产业是在高强度研究开发基础上发展起来的最具活力和潜力的知识和技术高度密集的产业群体，它的崛起和迅猛发展对经济和社会发展产生了深刻的影响[1]，是经济发展的动力。随着高新技术产业的逐步兴起，相关国际组织以及国家政府部门和科研机构为高新技术产业统计工作的开展做了大量工作[2]，取得了一定成效。近年来，安徽省积极实施创新驱动发展工程，高新技术产业一直处于稳中有进的发展态势。高新技术产业数据统计工作也同样得到了省政府的高度关注与重视，为推进安徽省高新技术产业发展和产业结构调整提供了重要决策依据。2014 年，安徽省出台了“1+6”政策，提出了安徽省创新能力评价指标体系，明确将高新技术产业中的两个相关指标（即“高新技术产品进出口总额占地方进出口总额的比重”和“高新技术产业增加值

作者简介：王俊(1985—)，女，安徽省科学技术情报研究所助理研究员，硕士，主要研究方向：科技统计。
基金项目：安徽省科技攻关计划项目“高新技术统计关键指标挖掘研究”（1301023012）；国家创新发展司委托项目子课题“安徽省企业创新情况调查分析与研究”（ZLY2015123）。
收稿日期：2016 年 11 月 22 日。

占GDP比重”）列入了考核内容。但是，在日常工作中高新技术产业相关的统计指标却多达十几个，存在指标体系不够健全、关键指标不突出、缺少指标间关联分析等问题。本文将以安徽省创新能力评价中的高新技术指标为基础，兼顾指标数据的可获取性，选取了科技统计日常工作中使用的17个高新技术相关指标，并将数据挖掘和统计分析技术引入高新技术统计工作中，建立一套安徽省高新技术统计关键指标体系，形成一套高新技术关键指标分析框架和模型及可视化系统，再利用数据挖掘技术，深入分析和评价安徽省高新技术产业发展现状。

1　统计关键指标的抽取

在高新技术产业统计工作中，数据本身庞大高维，且往往掺杂着大量无关、冗余特征，影响数据信息的有效挖掘[3]。因此，要在多个指标中进行关键指标抽取。关键指标抽取研究适用于机器学习领域中的关键特征选择和特征提取[4−6]。这里分别研究监督学习条件下的特征选择方法和无监督学习条件下的特征提取方法对问题的适用性。

1.1　特征选择的算法

从是否使用了目标变量的角度，可以将特征选择算法分为有监督、无监督和半监督的特征选择方法。其中，有监督的特征选择方法是在数据具有标签的前提下，通过评估特征和目标变量之间的相关性，选择有判别性特征的指标，即得到哪些指标具有较强的标签指示性。在实际应用中，很难得到有标签的数据，因此相比于有监督的特征选择方法，无监督的特征选择方法的研究受到更多的关注。而半监督的特征选择，即“小标记样本问题”，使用目标变量的信息以及对应于标签数据和无标签数据之间的流形结构[7−8]。

特征选择算法有过滤器、包装器和嵌入式3种。其中，过滤器算法是指定义一些准则，对特征进行评估，得到评估值，再对这些值进行排序，从而选出最好的若干个特征。相关的特征评估准则包括互信息、最大间距准则、内核对齐和希尔伯特−施密特独立性准则。过滤器采用多种准则来避免冗余，而mRMR（min−Redundancy and Max−relevance）是最具有代表性的算法，是以最大依赖性、最大相关性和最小冗余为准则。mRMR是为了找到一个特征子集，与目标变量具有最大的相关性，而特征子集中的特征之间具有最小的冗余[9−10]。

1.2　评估指标与标签的相关性

特征的重要性依靠互信息熵来计算，即通过计算特征$x\left(x\in S_N\right)$与类标签C的关系，也就是该特征的重要度，可以表示为式（1）：

$$MI\left(x,c\right)=\sum_{c\in C}\sum_{j=1}^{\left|v^x\right|},\ P\left(v_j^x,c\right)\log\frac{P\left(v_j^x,c\right)}{P\left(v_j^x\right)P\left(c\right)} \tag{1}$$

其中，

$$P\left(v_j^x,c\right)=\frac{N_D\left(v_j^x,c\right)}{\sum_{j=1}^{\left|v^x\right|}N_D\left(v_j^x,c\right)},\quad P\left(v_j^x\right)=\frac{N_D\left(v_j^x\right)}{\sum_{j=1}^{\left|v^x\right|}N_D\left(v_j^x\right)},$$

$$P(c)=\frac{N_D(c)}{\sum_{j=1}^{\left|v^x\right|}N_D(c)}$$

其中，v_j^x是特征x的第j个属性值，$\left|v^x\right|$是特征x所有属性值的个数，c是类标签集合C中某一取值，$N_D\left(v_j^x,c\right)$是v_j^x与c在数据集D中出现的次数，$N_D\left(v_j^X\right)$与$N_D\left(c\right)$则分别表示v_j^x与c在数据集D中出现的次数。

根据式(1)，得到S_N特征集中所有特征的排序，即S'。在此排序的基础上，我们选择前k个特征，表示为：$S'=\left\{S'_1,S'_2,\cdots,S'_k\right\}$。

1.3　评估指标间的相关性

考虑到mRMR算法所选出的特征子集能在使特征子集与类标签之间的相关性最大化的同时，还能保证特征子集内部冗余最小化，可以有效提升分类器的性能。因此，本文选择了过滤式特征选择算法的代表算法之一mRMR来选择特征子集。首先给出mRMR的相关定义：

定义1　最小冗余：特征子集S内部的冗余最小化

$$mR(S)=\min\left\{\frac{1}{|S|^2}\sum_{f_i,f_j\in S}I(f_i;f_j)\right\} \tag{2}$$

定义2　最大相关：特征子集S与类标签L保持最大的相关

$$MR(S,L)=\max\left\{\frac{1}{|S|}\sum_{f_i\in S}R(f_i;L)\right\} \quad (3)$$

综合式（2）和式（3）就是最小冗余最大相关准则：$\phi=\max(MR-mR)$。在实际操作中，采用增量式搜索法寻找能够使ϕ最大化的最优（或近似最优）特征子集S。

假设已经找到含$n-1$个特征的子集S_{n-1}，则查找第n个特征的过程是：（1）在集合$\{F-S_{n-1}\}$中查找使ϕ最大的特征f；（2）将f添加到特征子集中$\{S_{n-1}\cup f\}$，并把f从集合$\{F-S_{n-1}\}$中去除$\{(F-S_{n-1})\setminus f\}$；（3）重复步骤（1）和步骤（2）查找其他特征直到满足停止条件，从而找到最优特征子集。其中，步骤（1）的目标优化式可以换成ϕ的等价形式

$$\phi=\max_{f\in F-S}\left\{R(f;L)-\frac{1}{|S|}\sum_{f_i\in S}I(f;f_i)\right\}$$

2　评价指标体系

根据日常统计经验，选取了和高新技术产业相关的17个指标，涉及高新技术产业、高新技术企业、高新技术产业开发区、科技企业孵化器、高新技术产业基地和生产力促进中心等众多方面，如表1所示。

3　综合评价

通过查阅年鉴和相关公报，收集了2005—2014年的相关统计数据。对2005—2014年的原始数据进行离散化处理。离散化是将一组连续的数据值放入存储桶的过程，以便得到可能状态的离散数目，表2中显示的就是通过离散化处理后，把原本“连续的”变量变成“1-5”5个离散的变量。然后再对两两指标进行相关性计算，得出结果如表3所示。

（1）相关度较高的指标有4对，其相关度大约在2.12，分别为：高新技术产业产值（亿元）—累计毕业企业数（家）；高新技术产业增加值占GDP的比重（%）—高新技术产业基地数（个）；高新技术企业数（家）—高新技术产业产值（亿元）；高新技术企业数占规模以上企业比重（%）—高新技术产业增加值占全省工业增加值的比重（%）。

（2）相关度次之的指标有4对，其相关度大约在1.952，分别为：上市高新技术企业数（家）—营业总收入（亿元）；高新技术产业基

表1　基础数据指标情况一览表

序号	指标	单位
1	安徽省高新技术产业产值	亿元
2	高新技术产业增加值	亿元
3	高新技术产业增加值占全省工业增加值的比重	%
4	高新技术产业增加值占GDP的比重	%
5	高新技术企业数	家
6	高新技术企业数占规模以上企业比重	%
7	营业总收入	亿元
8	国家重点高新技术企业数	家
9	上市高新技术企业数	家
10	孵化器数	家
11	孵化场面积	万平方米
12	在孵企业数	家
13	累计毕业企业数	家
14	高新技术产业基地数	个
15	基地内企业总收入	亿元
16	生产力促进中心数	个
17	生产力促进中心收入	亿元

表 2　离散化结果

年份/年	高新技术产业产值	高新技术产业增加值	占全省工业增加值的比重	占GDP的比重	高新技术企业数	占规模以上企业比重	营业总收入	国家重点高新技术企业数	上市高新技术企业数	孵化器数
2014	5	5	5	5	5	5	5	5	5	5
2013	5	5	5	5	5	5	5	5	5	4
2012	4	4	5	5	4	5	5	5	5	4
2011	3	3	4	5	3	4	5	3	5	3
2010	2	3	4	4	2	3	4	1	4	3
2009	2	2	3	3	2	3	3	2	2	2
2008	1	1	3	2	2	3	3	2	2	2
2007	1	1	2	2	2	1	2	2	2	1
2006	1	1	2	2	2	1	1	1	1	1
2005	1	1	1	1	1	1	1	1	1	1

表 3　相关度计算结果

互信息（相关性）	2.1219	2.1219	2.1219	2.1219	1.971	1.971	1.9219	1.9219
特征编号1	1	4	5	6	9	14	13	16
特征编号2	13	14	1	3	7	15	1	1

地数（个）—基地内企业总收入（亿元）；生产力促进中心数（个）—高新技术产业产值（亿元）；累计毕业企业数（家）—高新技术产业产值（亿元）。

相关度越大的指标，说明指标统计冗余度越高。由此可见，选取的 17 个高新技术指标中有一定的冗余度，可以进行筛选。

通过对 2009—2014 年合肥累计认定高新技术企业数、当年认定高新技术企业、高新技术产业总产值、高新技术产业增加值和高新技术产业增加值占 GDP 比重等指标及处于中等位次排名信息进行分析，以中等位次排名为类标签，基于互信息模型分析其余各指标对中等位次排名的影响程度，提取其中的关键指标。

分析结果显示，对中等位次排名影响度从大到小的指标分别为高新技术产业增加值占 GDP 比重、高新技术产业增加值、累计认定高新技术企业数、高新技术产业总产值、当年认定高新技术企业数，其重要度指标分别为 1.45、1.12、1、1、0.46。从中可以看出，当年认定的高新技术企业数对中等位次排名的影响度不大，而高新技术产业增加值占 GDP 比重对排名影响较大。

此外，对合肥、淮北、亳州等 16 个地市近年来的统计指标（累计认定高新技术企业数、当年认定高新技术企业数、高新技术产业增加值和高新技术企业数与规模以上工业企业数之比等指标）进行分析，分析哪些指标对高新技术产业总产值上升具有重要影响。

基于互信息模型进行相关性分析，结果显示，对高新技术产业总产值上升影响力的重要度从大到小依次为：高新技术产业增加值占 GDP 比重、高新技术企业数与规模以上工业企业数之比、高新技术产业增加值、累计认定高新技术企业数、当年认定高新技术企业数，其影响度分别为 0.087、0.053、0.0403、0.0194、0.0194。由此可见，高新技术产业增加值和高新技术企业培育情况是衡量一个地区高新技术产业运行情况的重要指标。

4　结语

本文建立了一套高新技术统计体系指标，选取 2005—2014 年统计数据作为研究的原始数据，抽取统计关键指标分析安徽省高新技术产业的发展情况。统计分析表明，“十二五”以来，安徽省高新技术产业处于稳中有进的发展态势，截至 2016 年年底，全省拥有高新技术企业 3863 家，占全省规模以上工业企业数的 19.9%；全省高新技术产业实现增加值 4094.9 亿元，占全省 GDP 的 17%。由于研究初期兼顾统计指标的可获取性，研究结果可能存在一定的局限性，但对高新

技术日常统计工作仍然具有一定的指导作用。研究结果表明，影响一个地区高新技术产业运行情况的重要指标有高新技术产业增加值和高新技术企业培育情况，从而提高了日常统计工作中高新技术产业数据的有效性，可更深层次地分析全省高新技术产业的发展。

参考文献

[1] 张珍花,路正南.高新技术产业统计指标体系的构建[J].统计与决策, 2015,187(4):13−14.

[2] 沈艳华,赵振宁.高新技术产业统计调研报告[J].商业研究, 2006, 346(14):204−206.

[3] SPOLAOR N, CHENRMAN E A, MONARD E A, et al. A comparison of multi−label feature selection methods using the problem transformation approach[J]. Electronic Notes in Theoretical Computer Science,2013,209:135−151.

[4] ZHANG M L, ZHOU Z H. A review on multi−label learning algorithms[J].IEEE Transactions on Knowledge and Data Engineering,2014,26(8):1819−1837.

[5] DENDAMRONGVIT S, VATEEKUL P, KUBAT M. Irrelevant attributes and imbalanced classes in multi−label text−categorization domains[J].Intelligent Data Anvlysis,2011,15(6):843−859.

[6] 周国静, 李云.基于最小最大策略的集成特征选择[J].南京大学学报(自然科学), 2014,50(4):457−465.

[7] 王婧.面向在线环境的数据编码问题研究[D].合肥:合肥工业大学, 2015.

[8] 许尧.过滤式特征选择算法研究[D].合肥:合肥工业大学, 2015.

[9] 姚明海, 王娜, 齐妙,等.改进的最大相关最小冗余特征选择方法研究[J].计算机工程与应用, 2014,50(9):116−122.

[10] 胡学钢, 许尧,李培培,等.一种过滤式多标签特征选择算法[J].南京大学学报(自然科学版),2015, 51(4):723−730.

（上接第 67 页）

业得到跨越式发展，销售收入从几百万元增长突破亿元大关，并成功在新三板上市。

6　结语

多年来，湖南省通过竞争情报精准服务企业新产品开发，形成了基于企业生命周期理论的竞争情报精准服务产品体系，初步探索了为企业产品开发决策提供有效支撑的竞争情报采集分析工作方法与体系。但是，企业产品开发与技术创新任重道远。为企业的产品创新提供全流程的竞争情报服务尚不完善，尤其是在产品开发中后段的模块化服务有待进一步深入研究与实践操作。笔者将在以后的研究中不断探索与优化企业竞争情报服务新模式。

参考文献

[1] ADIZES Ichak. 企业生命周期[M]. 北京: 中国社会科学出版社, 1997:9−10.

[2] 陈佳贵. 关于企业生命周期与企业蜕变的探讨[J].中国工业经济, 1995(11):5−13.

[3] WUYTS S,DUTTA S,STREMERSH S.Portfolios of interfirm agreements in technology: intensive markets: consequences for innovation and profitability[J].Journal of Marketing, 2004,68(2):88−100.

[4] ROBERT Buzzell D, BRADLEY T Gale. The PIMS principles: linking strategy to performance[J]. Journal of Marketing, 1987(8): 37−48

[5] 黄永春,姚山季. 产品创新与绩效:基于元分析的直接效应研究[J]. 管理学报, 2010(7):1027−1031.

[6] 屈鹏.基于企业生命周期的竞争优势及战略研究[D]天津: 南开大学,2007:20−45.

[7] 晏文胜,陈述. 企业生命周期与超循环理论探析[J].现代管理科学,2004(12):56−58.

[8] 叶琳琳,史博文.虚拟企业生命周期各阶段竞争力管理分析[J]. 经济研究导刊,2010(18):31−33.

[9] 甘胜军,王玉.基于企业生命周期视角的竞争战略转换规律研究[J].科技管理研究,2014(5):209−216.

[10] 潘杏梅,方红,张培峰.面向小微企业的竞争情报服务模式研究: 基于浙江的分析[J].科技管理研究, 2015(21):135−140.

[11] 刘建,刘玉照.面向产品创新模糊前端的竞争情报系统研究[J].情报杂志,2009(5):95−98.

[12] 王晓慧.基于内部供应链的企业竞争情报分析模型及实施方法研究[J].情报学报,2014(2):140−147.

[13] 史敏,李维思,刘素华,等.中小企业竞争情报供给方法研究:以2012年长沙市中小企业“协同创新”示范活动为例[J].情报理论与实践,2013(12):59−63.