Statistics and Application 统计学与应用, 2022, 11(1), 76-85
Published Online February 2022 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2022.111009

Hans汉斯

基于聚类分析法的广东省区域科技竞争力的评价

徐凡然

南京信息工程大学，江苏 南京

收稿日期：2022年1月10日；录用日期：2022年1月23日；发布日期：2022年2月10日

摘 要

结合广东省科技竞争力发展水平，从R&D经费、发明专利授权量、财政科技支出占地方财政支出比重等9个维度出发，构建广东省区域科技竞争力评价指标体系。综合广东统计年鉴和广东科技年鉴的数据，用K-means++方法将科技发展水平相似的城市聚为同类，并对不同类别的区域科技竞争力进行主观评价与比较。对广东省21个城市这一数据集进行多次不同训练集和测试集的选取，进行多次对比实验。实验结果表明广东省科技发展水平总体较高，但区域之间科技竞争力差异较大，科技发展水平不平衡。最后通过对各指标结果综合评价并提出了针对加强广东省科技竞争力的对策建议。

关键词

科技竞争力，广东省指标体系，K-means++，聚类分析

Evaluation of Regional Science and Technology Competitiveness of Guangdong Province Based on Cluster Analysis Method

Fanran Xu

Nanjing University of Information Science and Technology, Nanjing Jiangsu

Received: Jan. 10th, 2022; accepted: Jan. 23rd, 2022; published: Feb. 10th, 2022

Abstract

Combined with the development level of Guangdong Province's science and technology competi-

文章引用: 徐凡然. 基于聚类分析法的广东省区域科技竞争力的评价[J]. 统计学与应用, 2022, 11(1): 76-85.
DOI: 10.12677/sa.2022.111009

tiveness, this paper constructs the evaluation index system of Guangdong Province's regional science and technology competitiveness from the following nine dimensions: R&D expenditure, invention patent authorization, and the proportion of financial science and technology expenditure in local financial expenditure etc. K-means++ method was utilized to cluster the cities in Guangdong Province through the data collected from Guangdong statistical yearbook and Guangdong science and technology yearbook. K-means++ clustered the cities with similar level of scientific and technological development into the same category, and make subjective evaluation and comparison of regional science and technology competitiveness of different categories. The dataset of 21 cities in Guangdong Province was divided into train dataset and test dataset several times. The experimental results show that the overall level of science and technology development in Guangdong province is high, but the differences in science and technology competitiveness between regions are large, and the level of science and technology development is unbalanced. Finally, through the comprehensive evaluation of the results of each index, the countermeasures and suggestions for strengthening the science and technology competitiveness of Guangdong province are put forward.

Keywords

Technological Competitiveness, Guangdong Province Index System, K-means++, Clustering Analysis

1. 引言

当今时代，城市与国家的发展差异很大程度上体现在科学技术领域。科技竞争力已经是衡量各个国家综合国力的重要指标。科技强则国家强，当下提升科技竞争力是每个国家、每个城市的重中之重。我们的主要任务是提高科技水平与其竞争力，使得国家或地区的发展水平得到提高。只有针对每个地区的科技水平和科技实力进行正确建模、分析，才能合理调配人才资源，因地制宜地发展科技水平，进而提升总体国力。

广东省是科技与经济大省，但不同区域发展水平参差不平。为了不断提高科技竞争力，缩小区域间科技水平的差距，要根据数据分析出每个城市的情况，进而针对各个城市的特点因地制宜地制定相关政策来促进城市综合发展。本文对科技竞争力评价模型的指标选取并构建评价指标，利用科技进步考核的指标和相关数据对广东省 21 个城市的科技竞争力进行综合分析，通过聚类方法并结合各个指标对广东省各城市的科技发展程度进行分级，通过分析科技发展较高的城市，从中总结经验然后对科技发展水平较低的城市提出发展建议。

2. 文献综述

当今时代是大数据的时代，复杂的数据只有被妥善的处理分析，才能转化为有价值的数字。聚类分析[1] [2] [3]是把分类对象按一定规则分成若干类。张鲲[4]利用了主成分分析的方法做了深圳市科技人才竞争力的评价研究，通过对初始数据的获得，运用 PCA 方法确定各指标权重分布，并计算出科技人才竞争力从而提出相应对策。傅春[5]等人使用熵权法和超效率 DEA 对中部六省进行了科技竞争力的评价。以科技指标为中心，同时综合经济、社会、环境等“软因素”选取相关指标，建立科技竞

争力的综合评价指标体系。丁超勋[6]等人利用因子分析对我国科技人才区域进行了竞争力的评价。他们建立了科学严谨的关于科技人才区域竞争力的评价指标体系，综合分析并为科技发展提出可行性建议。

针对聚类算法而言，杨杰[7]在其 K 均值算法中对初始聚类中心的确定问题的研究中阐明了 K-means 聚类方法是一种简洁有效用途最广[8]的聚类方法，但它对初始聚类中心如何选取依赖较高，如果选取不当，就很容易陷入局部最优解的问题之中，无法得到全局最优解。所以一直以来国内外许多学者提出的改进方法层出不穷，K-means++ [9]算法有效解决了初始聚类中心相距较近的问题，K-means++与 K-means 相比较而言，在聚类精度上更胜一筹，K-means++是选择尽可能远的数据点作为初始聚类中心，但它依然没有解决由于随机选取第一个初始聚类点中心而导致的聚类结果不稳定问题。胥栋宽[10]等人在他们对聚类算法的综述中提到了传统聚类算法主要是针对数据本身固有特性进行分析，设计了基于损失函数，密度，数据分布情况等的聚类算法，现代聚类算法主要是通过核学习转为高维特征向量等。包志强[11]等人针对传统 K-means 聚类算法容易被孤立点影响的不足，提出一种新的对孤立点不敏感的 K-means 聚类算法。

3. 指标体系及模型的构建

3.1. 科技竞争力指标体系构建

为了全面反映广东省各区域的科技竞争力，本文以《广东统计年鉴(2020)》、《广东省经济综合竞争力评价分析报告(2017~2018)》[12]的数据作为依据，其中包括国民经济核算、就业和工资、区域主要经济指标、教育和科学等数据，把各个城市的技术输出、科研投入等纳入评价体系之中，构建科学严谨的广东省区域科技竞争力评价指标体系[13] [14]，见表 1。

Table 1. Evaluation index of regional science and technology competitiveness of Guangdong province
表 1. 广东省区域科技竞争力评价指标

科技竞争力评价指标	
X_1	R&D 人员
X_2	R&D 经费
X_3	科研机构数目
X_4	发明专利授权量
X_5	技术市场成交合同金额
X_6	科技经费支出占总经费支出比重
X_7	新产品销售收入
X_8	科技活动收入
X_9	高技术产品出口额

广东省各个城市的 R&D (实验研究与发展)人员和经费、科研机构数目、发明专利授权量都在统计年鉴中有准确的数据记录。而技术市场成交合同金额没有明确的数据显示，我们用已有的技术性收入的数据来代替原指标，依然可以很好地涵盖原指标中对技术开发、技术转让、技术咨询和技术服务类合同的

成交额的意义。由于原来的指标中高技术产业主营业务收入这一指标没有直接的数据来源，所以用新产品销售收入来代替原有指标，新产品是指采用新技术原理、设计构思研制生产的全新产品，并在结构、材质、工艺等某一方面比原有产品有明显改善，从而显著提高了产品性能。所以此指标可以直观的反映出科技发展水平。一个地区的专利授权量与其科技竞争力息息相关。上表中的 9 个指标对科技竞争力的影响十分重要。

3.2. 聚类模型的建立

数据分析是一种现代科学研究常用方法，这种方法综合了多个学科，如计算机科学、经济学、传播学、生物学等等。聚类分析作为数据分析的基本组成部分，具有重要作用。在本文的工作中，针对广东省区域科技竞争力的评价这一问题，主要研究 K-means 和 K-means++两种算法。距离衡量函数和相似性度量函数是聚类算法中不可缺少。研究表明，对于定量的数据特征，即每个样本的多元特征都可以用数字化指标来衡量，使用距离度量函数为好。对于定性的数据特征，则首选相似性度量函数。常见的距离度量函数有如下几种：

明式距离(Minkowski distance)：

$$\left(\sum_{l=1}^{d}\left|x_{il}-x_{jl}\right|^{n}\right)^{\frac{1}{n}} \tag{1}$$

其中，x_{il} 是第 i 个样本的第 l 维特征的数值。

标准欧式距离(Standardized Euclidean distance)

$$\left(\sum_{l=1}^{d}\left|\frac{x_{il}-x_{jl}}{s_l}\right|^{2}\right)^{\frac{1}{2}} \tag{2}$$

其中，s_l 是第 l 维特征向量的标准差。

马氏距离(Mahalanobis distance)

$$\sqrt{\left(x_i-x_j\right)^{\mathrm{T}} S^{-1}\left(x_i-x_j\right)} \tag{3}$$

其中，S 是样本的协方差矩阵。

4. 广东省区域科技竞争力评价的问题求解

4.1. 数据预处理

表 2 给出 21 个城市的未经过归一化的各项指标的数据。

Table 2. Index values of technological competitiveness of major cities (raw data)
表 2. 几大城市的科技竞争力指标数值(原始数据)

科技竞争力指标数值				
广州	X_1	X_2	X_3	X_4
	95562	267.27	102	796
X_5	X_6	X_7	X_8	X_9
3177834	0.7431	48522236	9887133	6386057

Continued

珠海	X_1	X_2	X_3	X_4
	30808	82.77	7	12
X_5	X_6	X_7	X_8	X_9
0	0.8390	14180174	148514	4592020
中山	X_1	X_2	X_3	X_4
	36620	59.28	3	0
X_5	X_6	X_7	X_8	X_9
0	0.7681	11307042	21608	3632154

聚类模型的输入需要保证高维特征向量的各个维度一致，也就是需要对数据进行归一化，或称为去量纲化。这样做的意义是在多指标的评价体系中，由于评价指标选取的不同，单位的差异会带来数值上的巨大差异，因此需要将数据进行归一化处理，使得不同维度的特征在数值上有可比性，提高聚类模型的准确度。我们使用常用的数据归一化方式——Min-max normalization：

$$x^* = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{4}$$

这是对原始数据进行了线性变换，对所有的特征向量 $\boldsymbol{x}$ 去量纲化和归一化，使得所有指标的数值全部落在[0-1]之间，下面给出表 2 中的城市归一化之后的数据分布，如表 3 所示。

Table 3. Index values of science and technology competitiveness of several major cities (after data normalization)
表 3. 几大城市的科技竞争力指标数值(数据归一化后)

	科技竞争力指标数值			
广州	X_1	X_2	X_3	X_4
	0.32627	0.27453	1	1
X_5	X_6	X_7	X_8	X_9
1	0.65861	0.40013	1	0.13456
珠海	X_1	X_2	X_3	X_4
	0.10127	0.08318	0.0404	0.01508
X_5	X_6	X_7	X_8	X_9
0	0.78754	0.11398	0.01502	0.09653
中山	X_1	X_2	X_3	X_4
	0.12147	0.05881	0	0
X_5	X_6	X_7	X_8	X_9
0	0.69227	0.09003	0.00219	0.07618

4.2. 聚类模型的实验结果分析

将科技竞争力归一化后的指标数据代入聚类模型，对数据集(21 个城市分成训练集和测试集)进行 3 次随机分，每次选择 18 个城市的指标数据用于聚类模型的输入，3 个城市作为测试。下面给出 3 次实验的划分：

实验一) 训练集：梅州、广州、云浮、深圳、珠海、揭阳、佛山、韶关、河源、惠州、湛江、中山、江门、阳江、茂名、肇庆、清远、汕尾。

测试集：东莞、汕头、潮州。

实验二) 训练集：深圳、珠海、汕头、东莞、韶关、河源、梅州、惠州、汕尾、中山、江门、阳江、湛江、茂名、潮州、清远、揭阳、云浮。

测试集：广州、佛山、肇庆。

实验三) 训练集：广州、珠海、佛山、肇庆、汕头、东莞、韶关、河源、江门、惠州、汕尾、中山、阳江、茂名、潮州、清远、揭阳、云浮。

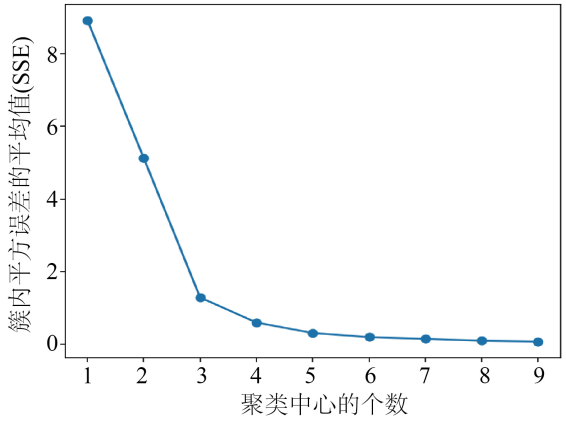
测试集：深圳、梅州、湛江。

K-means++选择初始聚类中心是随机选择的，因此需要增加随机次数来提高模型的分割精度，我们将随机选择初始点次数记做变量 n_init，在实验中我们设置其默认值为 10。这样做的意义在运行K-means++的聚类算法中，会进行 n_init 次的随机试验，最终返回簇内平方误差和(SSE)最小的一次聚类结果，在单次 K-means++实验中，最大迭代步数(max_iter)也是一个重要的参数，我们设置其默认值为 300。聚类模型的重要参数设置整理为表 4。

Table 4. Important superparameters in clustering model
表 4. 聚类模型中的重要超参数

变量名	变量含义	默认值
n_init	随机选择初始点次数	10
max_iter	最大迭代步数	300
tol	跳出迭代次数的条件	簇内平方误差和小于 10^{-4}
backbone	骨干模型	K-means++

聚类中心个数的选取是 K-means++模型中另一个重要的参数，针对这个参数，我们采用有限步长迭代尝试法，由于广东省仅有 21 个城市，训练样本为 18 个城市，我们将聚类中心设置为[1,10]，循环利用表 4 中的模型参数进行聚类分析，通过分析聚类中心的个数和簇内平方误差和之间的关系来确定合适的 k 值。本次实验使用的编程语言为 Python，主要用到了其中的 sklearn、numpy、matplotlib 等库函数。按照之前划分的数据集，我们做了三次尝试，分别得到了一张聚类中心的个数和簇内平方误差和的关系图，如下图所示：



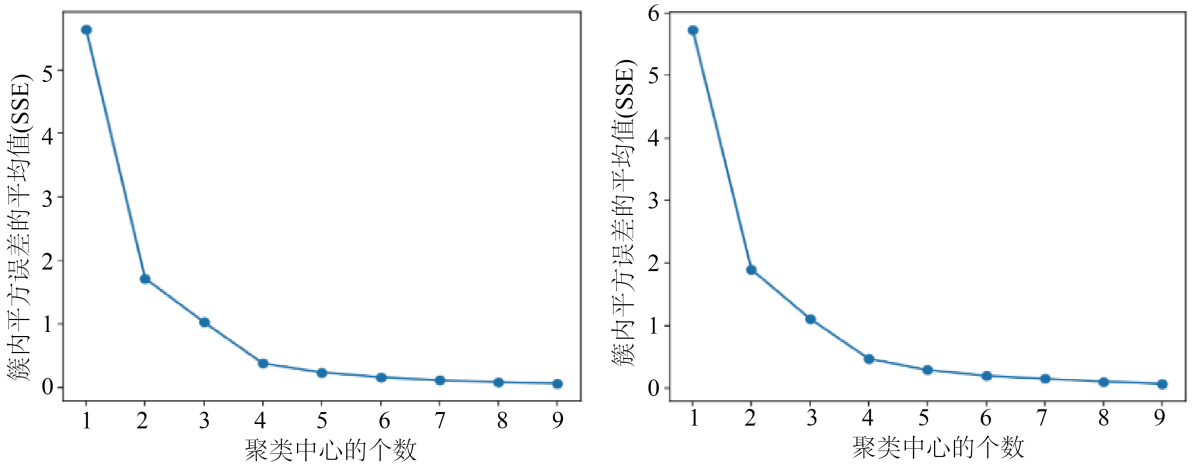


Figure 1. Results of three experiments
图 1. 三次实验的结果图

上图第一行是在第一组实验数据集测到的结果，第二行的两幅图分别是在第二和第三组实验数据集上测到的结果，可以看到不同数据集的划分并不影响整体的趋势，从图 1 可以直观地看出，在用 K-means++ 聚类模型分析广东省区域科技竞争力的时候，聚类中心的个数应该选为 4，即 k 值为 4，因此，之后的实验都是默认为 k 值为 4 进行的实验。下面，我们对划分好的三个数据集进行聚类实验，并且给出实验中训练集和测试集的聚类归属类别和聚类中心的值，并且对实验聚类的结果进行直观上的分析。

第一组实验中训练集的聚类结果如下表 5 所示。

Table 5. City category table
表 5. 城市归属类别表

肇庆	揭阳	珠海	佛山	韶关	河源
1	1	1	1	1	1
梅州	惠州	茂名	中山	江门	阳江
1	1	1	1	1	1
湛江	云浮	深圳	广州	清远	汕尾
1	1	2	3	4	4

四个聚类中心(保留了小数点后四位)的值分别为：

$$\begin{bmatrix} K_1 \\ K_2 \\ K_3 \\ K_4 \end{bmatrix} = \begin{bmatrix} 0.0648 & 0.0425 & 0.0866 & 0.0073 & 0.0008 & 0.6824 & 0.0693 & 0.0056 & 0.0489 \\ 1.0000 & 1.0000 & 0.0303 & 0.6533 & 0.0453 & 1.0000 & 1.0000 & 0.0625 & 1.0000 \\ 0.3263 & 0.2745 & 1.0000 & 1.0000 & 1.0000 & 0.6586 & 0.4001 & 1.0000 & 0.1346 \\ 0.0079 & 0.0063 & 0.0556 & 0.0000 & 0.0000 & 0.6216 & 0.0210 & 0.0001 & 0.0098 \end{bmatrix}$$

将测试集汕头、东莞、潮州的特征数据与四个聚类中心进行距离判别，结果显示这三个城市都是属于第一类别。通过这组测试数据，可以发现广州和深圳两个城市被单独归为一类，根据客观条件来看应该是属于科技较为发达的城市，而汕尾和清远被归为一类，根据客观条件来看应该是属于科技较为落后的城市，其余的城市被归为一类，属于科技水平中等的城市，实验结果符合预期所想，因为汕头、东莞、

潮州等地的科技水平确实比广州、深圳还是相差较远，只能属于科技发展中等的城市。第二组、第三组实验中训练集的聚类结果略，从三组对照实验可以观察出，深圳，广州是比较极端的城市，就是说在21个城市样本中，没有和这两个城市科技水平较为接近的城市，因此，这两个城市放在训练集中是比较好的。

5. 实验与分析

5.1. K-means 与 K-means++方法对比

根据实验一中的数据集划分(深圳市和广州市均在训练集中)，保持其他参数不变，将聚类模型设置为K-means 和 Kmeans++来验证实验结果。如下表 6 所示：

Table 6. City category table of different methods
表 6. 不同方法的城市类别表

城市	K-means++	K-means
广州	3	3
深圳	2	2
珠海	1	1
佛山	1	1
韶关	1	1
河源	1	1
梅州	1	1
惠州	1	1
汕尾	4	4
中山	1	1
江门	1	1
阳江	1	1
湛江	1	1
茂名	1	1
肇庆	1	1
清远	4	4
揭阳	1	1
云浮	1	1

从上表可以看到，K-means 算法和 K-means++算法在样本容量小的情况下，经过充分迭代后得到的结果一致，并且测试集中的汕头、东莞、潮州也都被归为第一类别。

5.2. 迭代次数的影响

采取实验一中的数据集划分(深圳市和广州市均在训练集中)，选用 K-means++和 K-means 模型，其他参数与表 4 中保持一致，并且聚类中心的个数为 4。调整最大迭代次数的参数值，比较聚类结果的 SSE 可以分析出最大迭代次数对于聚类模型的影响。如图 2 所示：

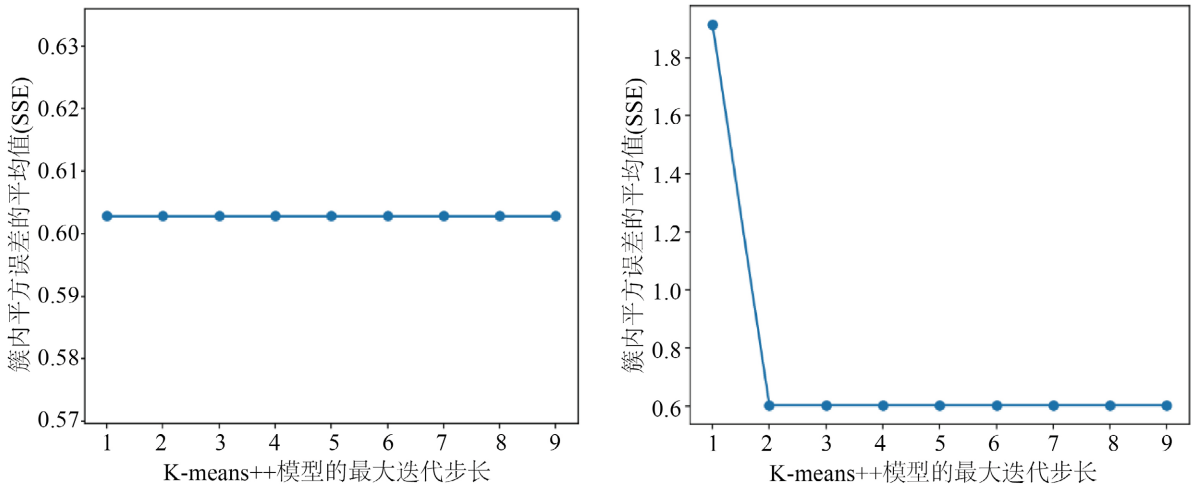


Figure 2. Maximum iteration step size
图 2. 最大迭代步长

从上图可以直观的看出，利用 K-means 算法对实验一的数据集只需要一次迭代就能完成广东省不同区域科技竞争力的聚类，而 K-means 算法则需要三次迭代才能完成，我们可以推断，如果做全国的区域科技竞争力的评价模型，K-means++算法的优良性会得到更为显著的体现。

6. 总结

近几年来，广东省认真贯彻落实党的十八大、十九大的指导思想，调集了全社会很多人力物力财力对创新驱动发展策略做全面部署协调。习近平总书记强调应加强对创新工作的技术与经费投入，深入实施创新驱动发展战略，推动科学技术发展。但广东省各区域之间发展水平差异较为明显，其中深圳市不断加快建设现代化国际创新型城市和国际科技、产业创新中心。广州市也紧追其后，不断加强对科技企业的管理，提高科技企业的产出效率和质量。截至 2016 年，珠海市(R&D)高达 55.23 亿元，占区域 GDP 比重约为 2.48%。更为惊讶的是，珠海市的有效发明专利量高达 33.5 件(按万人计)，仅次于深圳市。伴随着科研水平的发展，珠海市的高新技术产品也呈现了较快的增长速度，同比增长 9.3%，产值达 2446.56 亿元，并且仍然呈现出不断加快的发展速度的趋势。中山市不断强化柔性引进，放宽团队引进门槛，加大创新创业高端人才的政策虹吸效应。韶关、河源、梅州、惠州、江门等地虽然高端创新资源较为匮乏、与社会经济发展不是十分适应，但他们积极听从省科技厅的指导，坚决执行创新驱动发展的科技强国战略，积极调配国内外的科技资源，为提升科技竞争力，这些城市还在软科技方面下足了功夫，比如改善科研环境，提升科技服务的效率和质量，发挥科技创新在加快型升级中的支撑引领作用。汕尾市河源市在深圳市的帮扶机制下，他们的科技创新政策环境和社会氛围得到了良好的改善。阳江、湛江、江门、清远等地在技术领域实现多个零突破，初步形成“1 + N”科技创新政策体系。茂名、肇庆、揭阳、云浮等地通过组织实施国家重点的科技类创新的发展项目，并且这些城市把握住了机会，在一批产业的关键技术上取得了突破，提高了自主创新水平，促进了产业优化升级，大力推动当地科技创新能力的发展。

通过对比广东省 21 个市区域科技竞争力的相关情况，发现广州、深圳综合竞争力较强，汕尾、清远的科技发展水平较为落后。其余城市的科技发展水平较为相似。广东省区域科技发展水平明显差异较大，发展不平衡现象较为严重。本研究进一步提出提高广东省科技竞争力的对策如下：

1) 加速发展珠三角地区，使得其创新驱动发展能力明显增强。扎实推进创新创业人才队伍建设，对于科技发展速度较为缓慢的地区要积极实施人才引育“一揽子”计划，积极落实人才子女入学、特色人

才住房补贴等政策，实现优秀人才队伍引育全覆盖。

2) 加速科技金融互帮互助融合发展、相互促进局面形成。将科技金融作为加快科技成果转化成新兴产业的重要推动力，打造服务创新驱动金融服务体系，推进科技与金融深度融合创新。

3) 缩小区域科技发展差异。虽然广州、深圳等地的科技水平非常发达，但广东省其余城市的科技竞争力较弱，从大局观出发，为了提高广东省整体的科技竞争力，就务必要将缩小区域科技发展差异作为重要任务，近年来，专业镇正在蓬勃发展中，专业镇的设立初衷是为了更好地协调区域间的科技、金融等行业，促进区域间的人才和知识流通，只有更多“专业镇”的出现，才能实现不同区域的产业互补，互惠共赢，进而缩小区域间的发展隔阂与差距，提高广东省总体的平均科技竞争力。

参考文献

[1] Zhao, D., Hu, X., Xiong, S., *et al*. (2021) K-Means Clustering and kNN Classification Based on Negative Databases. *Applied Soft Computing*, **110**, Article ID: 107732. https://doi.org/10.1016/j.asoc.2021.107732

[2] Ji, J., Bai, T., Zhou, C., *et al*. (2013) An Improved k-Prototypes Clustering Algorithm for Mixed Numeric and Categorical Data. *Neurocomputing*, **120**, 590-596. https://doi.org/10.1016/j.neucom.2013.04.011

[3] 谢娟英, 周颖. 一种新聚类评价指标[J]. 陕西师范大学学报(自然科学版), 2015, 43(6): 1-8.

[4] 张鲲. 基于 PCA 分析法的深圳市科技人才竞争力评价研究[J]. 经济研究导刊, 2019(8): 84-90.

[5] 傅春, 杨丽. 基于熵权法和超效率 DEA 的中部六省科技竞争力评价[J]. 科技通报, 2017, 33(7): 258-263.

[6] 丁超勋. 基于因子分析的我国科技人才区域竞争力评价[J]. 创新科技, 2018, 18(2): 40-43.

[7] 杨杰. K 均值算法中初始聚类中心的确定问题研究[D]: [硕士学位论文]. 上海: 上海师范大学, 2018.

[8] 张宇, 朱凝秀. 数据挖掘中的模糊聚类分析[J]. 工业设计, 2012(3): 71-72.

[9] Arthur, D. and Vassilvitskii, S. (2007) K-Means++: The Advantages of Careful Seeding. *Proceedings of the* 18*th Annual ACM-SIAM Symposium on Discrete Algorithms*, New Orleans, Louisiana, 1027-1035.

[10] Xu, D.K. and Tian, Y.J. (2015) A Comprehensive Survey of Clustering Algorithms. *Annals of Data Science*, **2**, 165-193. https://doi.org/10.1007/s40745-015-0040-1

[11] 包志强, 赵媛媛, 胡啸天, 赵研. 一种对孤立点不敏感的新的 K-Means 聚类算法[J]. 现代电子技术, 2020(5): 109-112.

[12] 尹成祥, 张宏军, 张睿, 綦秀利, 王彬. 一种改进的 K-Means 算法[J]. 计算机技术与发展, 2014, 24(10): 30-33.

[13] 李良成, 杨国栋. 广东省创新型科技人才竞争力指标体系构建及评价[J]. 科技进步与对策, 2012, 29(19): 130-135.

[14] 何忠葵, 张峰, 刘颖莹. 山东省科技竞争力指标分析及与苏、浙、粤比较——基于《中国省域经济综合竞争力发展报告(2017~2018)》[J]. 科技经济导刊, 2020, 28(2): 174-175.