DOI:10.13878/j.cnki.jnuist.2019.03.011

幸嘉宝[1,2]　黄晓敏[1,2]　曹杨[1]　罗燎[1]　廖好[1,2]

数据驱动下的环境变迁与区域经济成长关联分析

摘要

由于经济发展的复杂性,本文旨在探索由环境变迁引发这一动态、复杂而又相互作用的过程,通过引入环境变迁与经济成长两方面因素分析其中的潜在关联性,并将区域稳定性作为环境变迁与经济成长相互作用后的衡量指标,来评估该过程.1) 通过使用衡量国家经济健康程度的健康性与复杂性(Fitness and Complexity)算法,获得了新的评估国家经济成长的国家经济健康性系数,该系数能在竞争激烈的动态国际贸易环境下有更好预测 GDP 的表现.随后建立机器学习模型,成功预测了不同国家的稳定性类别,且预测精度都在 90% 左右.2) 实现了基于数据的环境变迁和区域经济成长的关联性可视化分析,通过分析能够得到潜在关联性结论:一些发展中国家经济稳定性与水资源和二氧化碳排放呈强关联,而发达国家则与人均耕地面积有关联.3) 设立评估国家稳定性的新指标,与世界主流指标相比,构建的新指标更注重原始数据的量化,减少了概念抽象的指标对预测性能的影响,且在评估区域经济成长时能更符合当前国际的实际经济情况.本文提出的评估区域稳定性的新排名是完全基于量化指标的,因此更容易实现,说服力更强.通过实际的预测效果分析,该新排名在衡量区域稳定性时弥补了世界主流排名由抽象指标带来的预测失真缺陷,能够满足基本的区域稳定性预测功能,并且能够对预测结果造成影响的主要因素进行解释.

关键词

数据分析;机器学习;特征工程;经济复杂性

中图分类号 TP242

文献标志码 A

收稿日期 2019-05-24

作者简介

幸嘉宝,男,研究方向为数据挖掘研究和特征分析.garbohsing@ gmail.com

廖好(通信作者),男,博士,研究员,主要研究方向为大数据分析计算以及网络科学.haoliao@ szu.edu.cn

1　中电科大数据研究院有限公司,贵阳,550022

2　深圳大学 计算机与软件学院,深圳,518060

0　引言

环境的变迁给人类社会的发展带来的影响是巨大的,并在不同地区呈现出不同程度的影响.早在 20 世纪末,许多国家便已开始关注由环境的变迁给国家的经济和政局稳定所带来的影响的研究,并重视环境变化的影响及采取措施遏制环境变化的速度和应对后续影响.21 世纪初,美国国防部(DoD)将与环境有关的不稳定因素确定为一项基本的战略考量,证据表明,环境压力是造成当代冲突的一个重要因素.Krakowka 等[1]的研究表明:事实上,环境引发的冲突是由动态的、复杂的和相互作用的过程所激发的,而不仅仅只是一种简单的、确定性的关系.因此,需要建立一个分析框架来理解前因和后果,同时还提出一个综合的环境安全定义极其重要的两个要素.

对于国家政权内部而言,环境变迁效应能够显示各国应对自然灾害的能力;从国际层面来看,环境的剧烈变化会引起一系列人道主义灾难,招致政治暴力,使本已脆弱的国家摇摇欲坠[2].国际上的军事对抗形成的原因也可能从传统的意识形态或国家荣誉的冲突演变成对能源、食物和水等有限资源的迫切需要和争夺[3].对抗动机的转变将会使评判国家的脆弱性的指标以及现有的安全威胁警告标志发生改变,环境变迁引起的环境压力往往会结合治理的乏力和社会的分裂,从而加剧一个国家的脆弱性[4].历史上就曾发生过许多次因环境压力导致的暴力冲突,尤其表现在种族冲突和种族分裂上.与此同时,国家的经济发展程度,国家的贸易出口能力,以及一系列宏观经济因素也在一定程度上影响着国家的稳定性[5-7].因此,从已有数据中挖掘一定的关联规则[8],探究和衡量环境变迁对区域局势稳定性的影响这一工作就具有了重要而长远的意义.

本文探索了由环境变迁引发冲突这一动态、复杂而又相互作用的过程,通过引入环境变迁与经济成长两方面因素分析其中的潜在关联性,并将区域稳定性作为环境变迁与经济成长相互作用后的结果,来评估该过程的影响程度.基于采集的数据,使用机器学习分类模型,对不同国家的脆弱性进行再分类,并衡量环境与经济因素在不同国家作用的相对脆弱程度,最终建立一个新的指数,该指数能在考虑更少指标,并只考虑纯量化数据的情况下达到与国际较为公认的脆弱国家指数(Fragile States Index,FSI)相近的结果.

1 资料与方法

1.1 资料来源

本文采用联合国公开提供的国际贸易统计数据库,通过 UN Comtrade 网站下载 CEPII 发布的 BACI 数据集(https://comtrade.un.org),其涵盖了 1995—2014 年间在产品级别细分的国家/地区之间的所有年度进出口交易量,产品主要以可达协调系统分类的 6 位数来代表.同时,本文采用了国际气候中心、World Bank 网站(https://data.worldbank.org)、GERMANWATCH 网站(https://germanwatch.org)及 FFP 网站(https://fundforpeace.org)下载的诸多气候/经济因子数据,如 FSI 指数、CRI 气候风险指数[9-15].将 BACI 国际贸易数据集、人均净化水占比数据集、二氧化碳排放量数据集、人均耕地面积数据集、CRI 气候风险指数数据集及 FSI 脆弱国家指数数据集进行整合清理,按照国家取交集,最终得到涵盖 2010—2014 年 122 个国家贸易进出口数据、人均净化水占比、二氧化碳排放量、人均耕地面积数、气候风险指数及脆弱国家指数的数据集.

1.2 研究方法

1.2.1 Fitness and Complexity 算法

Fitness and Complexity 算法旨在计量评估国家经济健康程度的指标,国家经济预测能力超过原有世界银行的预测模型,2017 年被世界银行开始采用.该算法的主要思想为:成功的国家几乎出口所有商品,无法从成功国家的出口数据中发现哪些商品是容易生产的,但对于生产能力较低的国家来说,它们能出口的商品,一定是易于生产与出口的.该算法由两个非线性耦合方程组成,最终通过迭代方法达到固定值[16-18].

F_i 与复杂度 Q_α 加权导出的乘积的总和成比例,Q_α 与出口产品的国家数量成反比,若一个国家的经济有较高的健康性,则该国家对于产品复杂性权重的贡献将减小,表达式如下:

$$F_i^n = \sum_{\alpha \in \xi_i} Q_\alpha^{n-1}, \tag{1}$$

$$Q_\alpha^n = \frac{1}{\sum_{i \in \xi_\alpha} 1/F_i^{n-1}}. \tag{2}$$

初始时 Q_α^0 和 F_i^0 的值均设为 1,使用迭代的方式计算中间变量 F_i^n 和 Q_α^n,最后将 F_i 和 Q_α 分别进行归一化.

国家贸易出口数据能直观地从量上反映国家的产品出口金额,但无法有效地评估国家的经济成长健康性,而使用 Fitness and Complexity 算法计算得到的国家经济健康性指数 Fitness 能有效地弥补这一弱点[19-20].

1.2.2 机器学习分类模型预测对比

本文为探寻进行区域稳定性分类最有效的机器学习模型,选取了当前主流且经典的分类预测模型:Decision Tree(决策树,DT)、Random Forest(随机森林)、GBDT(梯度提升树)和 XGBoost(提升树)模型,以对比方法的准确率、召回率和 F 值作为评估模型的衡量指标,将预测分类的结果进行对比.

1.2.3 机器学习分类模型多分类预测

本文采取的思路是仿照 FSI 指数将国家脆弱性分成脆弱、警告以及稳定 3 个类别,将数据集中 122 个国家的 FSI 指数排名作为分类模型的标注,即排名对应类别,而多分类模型的结果将产生涵盖 122 个国家的新排名,计算新排名与 FSI 排名的相似度及相关性.

2 气候因子与经济因子关联分析

2.1 关联性计算方法(Pearson 相关系数)

Pearson 相关系数,用于衡量数据的 2 个特征 α 和 β 之间的线性相关程度,该方法主要用于分析数据分布趋势、变化趋势一致性程度.其数学表达如式(3)所示:

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\mathrm{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}. \tag{3}$$

数据集的特征涵盖了气候变迁因子与经济成长因子,计算特征间的 Pearson 系数,结果如表 1 所示.

数据集中的异常值将对 Pearson 系数的计算结果产生误差,为体现该误差,本文通过数据的排名代替数值计算 Pearson 系数,结果如表 2 所示.

表 2 与表 1 的对比体现了使用排名代替数值进行关联性分析的优势.结果显示,净化水占比与二氧化碳排放量及 FSI 排名之间存在极强关联度,与 Fitness 指数之间存在强关联度;二氧化碳排放量与 FSI 排名之间存在极强关联度;Fitness 指数与 FSI 排名之间存在中等关联度.

2.2 关联性可视化分析

本文为探究各个影响因子在不同联盟(如是否为“一带一路”国家)以及在不同地区(如所属的大洲)下呈现的差异性,对特征进行衍生分析,分别将国家按地区进行分类,并绘制热图以体现关联性差异,结果如图 1 和图 2 所示.

表 1　特征间的 Pearson 系数(数值)

Table 1　Pearson coefficient between features (value)

	净化水占比	人均耕地面积	二氧化碳排放量	气候风险指数	经济健康指数	脆弱国家指数
净化水占比	1.000	0.052	0.435	0.080	0.441	0.711
人均耕地面积	0.052	1.000	0.106	0.056	0.034	0.160
二氧化碳排放量	0.435	0.106	1.000	0.172	0.179	0.516
气候风险指数	0.080	0.056	0.172	1.000	-0.240	0.121
经济健康指数	0.441	0.034	0.179	-0.240	1.000	0.506
脆弱国家指数	0.711	0.160	0.516	0.121	0.506	1.000

表 2　特征间的 Pearson 系数(排名)

Table 2　Pearson coefficient between features (rank)

	净化水占比	人均耕地面积	二氧化碳排放量	气候风险指数	经济健康指数	脆弱国家指数
净化水占比	1.000	-0.022	0.822	-0.104	0.619	0.838
人均耕地面积	-0.220	1.000	-0.045	0.081	0.211	0.061
二氧化碳排放量	0.822	-0.045	1.000	-0.122	0.511	0.794
气候风险指数	-0.100	0.081	-0.122	1.000	0.281	-0.105
经济健康指数	0.619	0.211	0.511	0.281	1.000	0.560
脆弱国家指数	0.838	0.061	0.794	-0.105	0.560	1.000

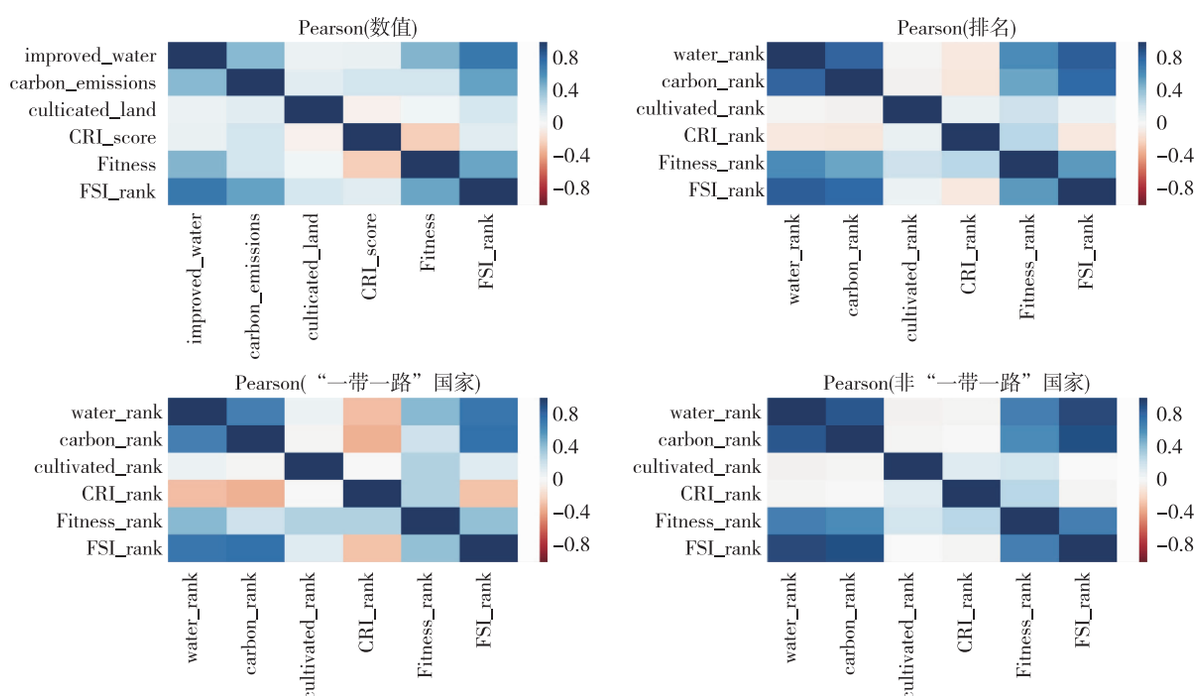


图 1　Pearson 相关系数热点图对比

Fig. 1　Pearson correlation coefficient hotspot map comparison

图 1 中 4 个子图分别代表用数值计算的 Pearson 系数、用排名计算的 Pearson 系数、“一带一路”沿线国家的特征间的 Pearson 系数和非“一带一路”国家特征间的 Pearson 系数.对比结果显示:对于非“一带一路”国家而言,净化水占比、二氧化碳排放量和 Fitness 指数与 FSI 排名的关联度比“一带一路”国家更强;对“一带一路”国家而言,区域稳定性与气候风险指数的关联度更高.

图 2 中 6 个子图分别代表了 6 个大洲的国家的特征间的 Pearson 相关系数.结果显示,对亚洲国家而言,气候风险指数与 Fitness 系数的值呈现了负相关的关系.在亚洲国家列表中,国土面积更大的国家

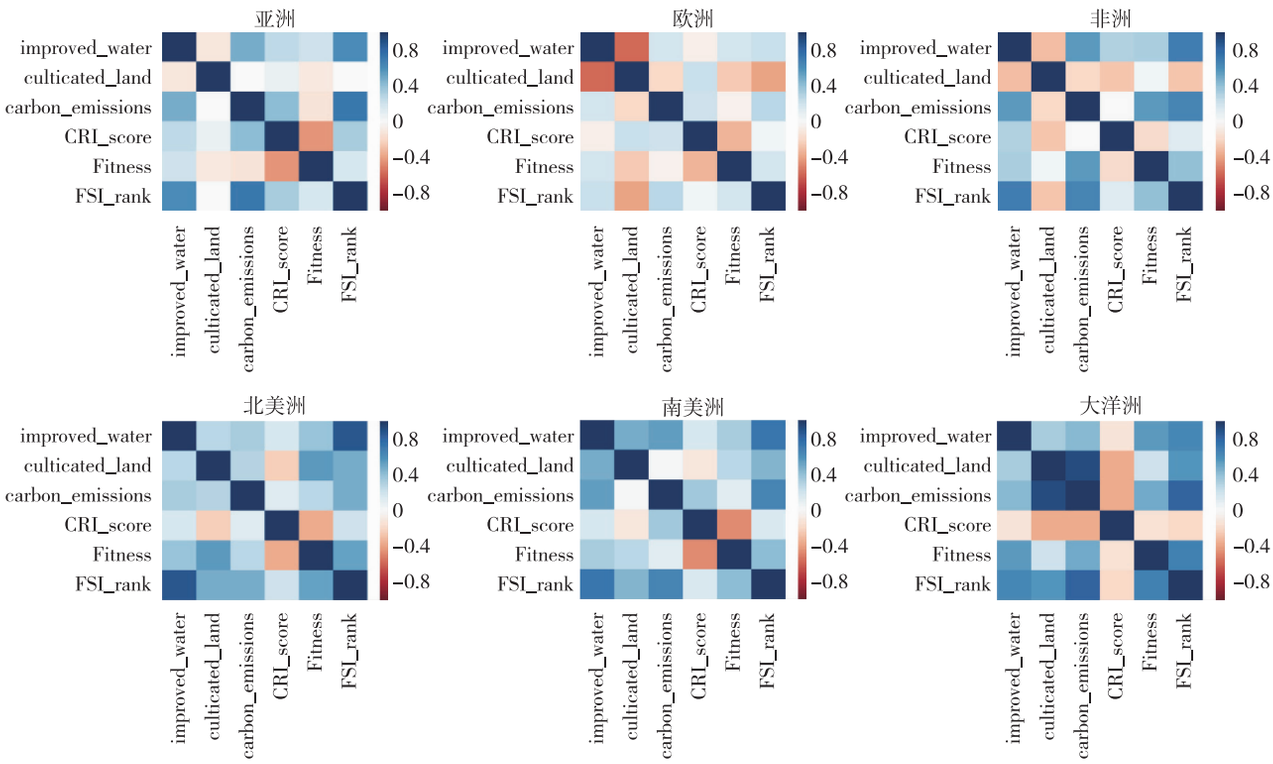


图 2　不同大洲的国家的 Pearson 系数热点图对比

Fig. 2　Comparison of Pearson coefficient heat maps of countries for different continents

对应了更高 Fitness 系数,同时伴随着更高的气候风险.对欧洲国家而言,区域稳定性受净化水占比和二氧化碳排放量的影响程度甚微.区域稳定性受净化水占比影响最大的是北美洲.查阅北美洲的国家列表发现,除美国和加拿大外,其余国家基本属于落后国家,尤其是拉丁美洲的国家,这些国家若要提高区域的稳定性,最好的办法是加大水质改良的投入.对大洋洲的国家而言,发展农业、增大人均耕地面积是提高区域稳定性的一种有效方法.

3　结果分析

3.1　分类模型的预测结果对比

本文采用 4 个分类模型分别是 Decision Tree、Random Forest、GBDT 和 XGBoost,其对比分析结果如表 3 和图 3 所示.

分析表 3 和图 3 可知,Random Forest 模型的分类预测效果最好,对国家的稳定性分类预测准确率最高能达 93%.

3.2　多分类预测结果对比

上述 4 种分类模型多分类所得的国家稳定性新排名与 FSI 排名对比结果如表 4 和图 4 所示.

表 3　不同模型的分类预测结果对比

Table 3　Comparison of classification prediction results of different models

模型	准确率	精确率	召回率	*F*1 值
决策树(ID-3)	0.885 8	0.892 7	0.885 8	0.887 2
随机森林	0.931 2	0.933 5	0.931 7	0.932 4
梯度提升树(GBDT)	0.905 4	0.911 6	0.909 9	0.910 7
提升树(XGBoost)	0.911 5	0.914 8	0.913 4	0.913 8

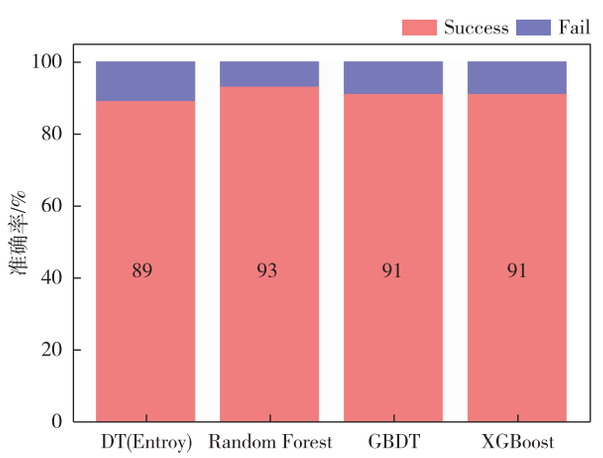


图 3　不同模型的分类预测结果对比

Fig. 3　Comparison of classification prediction results of different models

表 4　新排名与 FSI 排名的对比

Table 4　Comparison of new rankings with FSI rankings

模型	准确率	Pearson 相关系数
决策树(ID-3)	0.123	0.91
随机森林	0.200	0.98
梯度提升树(GBDT)	0.090	0.72
提升树(XGBoost)	0.150	0.92

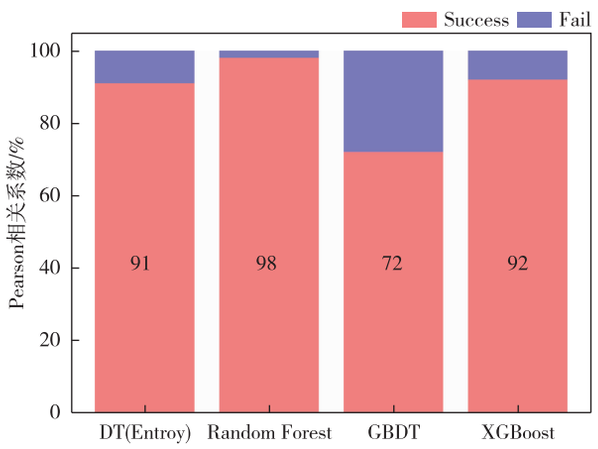


图 4　新指标排名与 FSI 排名的对比

Fig. 4　Comparison of new indicator rankings with FSI rankings

分析表 4 和图 4,结果显示,使用随机森林(Random Forest)模型进行多分类得到的新排名最佳,该新排名与 FSI 排名的相似率达到 0.2,且 Pearson 系数达到了 0.98,说明新排名结果与 FSI 排名呈现出高度一致,仅有细微的排名波动.通过 GBDT 模型进行多分类得到的新排名并不理想,与 FSI 排名之间的 Pearson 系数在 4 个模型中排名最低,仅有 0.72.通过 XGBoost 模型和信息增益决策树模型进行多分类得到的新排名虽然无法达到随机森林模型的效果,但新排名与 FSI 排名的 Pearson 系数也超过 0.9,说明新排名与 FSI 排名有高度的趋同程度.

不同的特征在不同的分类模型中的重要性程度存在差异性.为了探究不同特征在不同的分类模型中对新排名的影响程度,本文使用机器学习模型的特征评分函数(Feature-importance)来计算所有特征在各个分类模型中的重要性分数,计算结果如表 5 所示.

表 5 结果显示,新排名效果最好的随机森林模型中,二氧化碳排放量是影响排名最重要的因素,紧接着是人均耕地面积和 Fitness 系数.对 ID-3 决策树和 XGBoost 模型而言,人均净化水占比被视为影响稳定性排名的最重要因素.在 GBDT 模型中,各特征的重要性都不明显,验证了该模型在预测排名时表现不佳的结果.综合而言,净化水占比、二氧化碳排放量及 Fitness 指数是影响区域稳定性的 3 大因素.

表 5　不同特征在不同模型中的重要性

Table 5　The importance of different features in different models

特征	决策树(ID-3)	随机森林	梯度提升树(GBDT)	提升树(XGBoost)
净化水占比	0.30	0.15	0.07	0.16
二氧化碳排放量	0.23	0.21	0.12	0.15
人均耕地面积	0.14	0.18	0.16	0.12
气候风险指数	0.10	0.15	0.06	0.10
经济健康指数	0.13	0.16	0.11	0.13
“一带一路”政策	0.05	0.02	0.01	0.15
所属大洲	0.02	0.04	0.01	0.12
所属年份	0.04	0.08	0.01	0.07

4　结论

本文基于 2010—2014 年间 122 个国家的各项环境变迁与经济成长相关数据观察,在初步分析数据间关联性的基础上,采用多种分类模型对国家的区域稳定性进行分类预测,并对比预测结果的差异性,预测结果与脆弱国家指数的分类结果基本吻合.随后基于分类模型进行多分类得到国家稳定性新排名,新排名与脆弱国家指数排名对比,更加注重数据的量化,减小抽象化因素(如政治、军事)对结果的影响,并且所使用的原始数据量更少,使得提出的新排名具有一定的可解释性和可量化性,为进一步细化研究区域环境经济与区域稳定性提供了一个可参考的结果.

参考文献

References

[1] Krakowka A R,Heimel N,Galgano F A.Modeling environmental security in sub-saharan africa [J].Geographical Bulletin,2012,53(1):21-38

[2] Busby J W.Climate change and national security:an agenda for action [M].Washington D.C.:Georgetown University Press,2007:93-95

[3] Vorosmarty C J.Global water resources:vulnerability from climate change and population growth[J].Science,2000,289(5477):284-288

[4] Görg C,Brand U,Haberl H,et al.Challenges for social-ecological transformations:contributions from social and political ecology[J].Sustainability,2017,9(7):1045

[5] Portes A, Sensenbrenner J. Embeddedness and immigration:notes on the social determinants of economic action[J].American Journal of Sociology,1993,98(6):1320-1350

[6] Hechter M,Wallerstein I.The modern world-system:capitalist agriculture and the origins of the European world-economy in the sixteenth century[J].Contemporary Sociology,1975,4(3):217
[7] Sassen S.Cities in a world economy[M].London:Sage Publications,2018
[8] 王玮,陈恩红.关联规则的相关性研究[J].计算机工程,2000,26(7):6-8
WANG Wei,CHEN Enhong.Research on correlation of association rules[J].Computer Engineering,2000,26(7):6-8
[9] Harmeling S.Global climate risk index 2008[R].Berlin and Bonn:Germanwatch,2007
[10] Harmeling S.Global climate risk index 2009[R].Berlin and Bonn:Germanwatch,2008
[11] Harmeling S.Global climate risk index 2010[R].Berlin and Bonn:Germanwatch,2009
[12] Harmeling S.Global climate risk index 2011[R].Berlin and Bonn:Germanwatch,2010
[13] Harmeling S.Global climate risk index 2012[R].Berlin and Bonn:Germanwatch,2011
[14] Harmeling S,Eckstein D.Global climate risk index 2013[R].Berlin and Bonn:Germanwatch,2012
[15] Kreft S,Eckstein D,Junghans L,et al.Global climate risk index 2014[R].Berlin and Bonn:Germanwatch,2013
[16] Tacchella A,Cristelli M,Caldarelli G,et al.A new metrics for countries' fitness and products' complexity[J].Scientific Reports,2012,2:723
[17] Cristelli M,Gabrielli A,Tacchella A,et al.Measuring the intangibles:a metrics for the economic complexity of countries and products[J].PLoS One,2013,8(8):e70726
[18] Liao H,Huang X M,Vidmer A,et al.Economic complexity based recommendation enhance the efficiency of the belt and road initiative[J].Entropy,2018,20(9):718
[19] Tacchella A,Mazzilli D,Pietronero L.A dynamical systems approach to gross domestic product forecasting[J].Nature Physics,2018,14(8):861-865
[20] Cristelli M,Tacchella A,Cader M,et al.On the predictability of growth[M].The World Bank,2017.DOI:10.1596/1813-9450-8117

Data-driven analysis of the correlation between climate change and regional economic growth

XING Jiabao[1,2] HUANG Xiaomin[1,2] CAO Yang[1] LUO Liao[1] LIAO Hao[1,2]
1 CETC Big Data Research Institute Co.,Ltd,Guiyang 550022
2 School of Computer and Software,Shenzhen University,Shenzhen 518060

Abstract Regional economic development not only impacts regional politicization and economic construction but also improves national comprehensive competitiveness.How to predict the relationship between regional stability and economic development is an important problem.Accurately and quantitatively explaining development trends by using historical regional economic development data to analyze future development of the region can be difficult owing to the complexity of economic development.The goal of this work is to explore the dynamic,complex,and interactive process of the conflict caused by environmental change and analyze the potential correlation between environmental change and economic growth by regarding regional stability as a measure of the relationship between environmental change and economic growth.The main aspects of this work are as follows:1)Using the Fitness and Complexity algorithm achieves better performance in predicting national GDP growth.By applying machine learning models,we can predict stability categories of different countries with a prediction accuracy of 90%.2)We perform a correlation visualization analysis of data-based environmental change and regional economic growth.We find that some developing countries have strong economic stability associated with water resources and carbon dioxide emissions,whereas the economic stability of developed countries is associated with per capita arable land.3)We propose new indicators.Compared with the current mainstream indicators,the new indicators are more focused on quantification of raw data,reducing the impact of conceptual abstraction indicators on forecast performance,which makes them more responsive to assessing a country's stabilization.Through actual forecasting effect analysis,the new ranking compensates for the prediction distortion defects caused by the abstract indicators in the world mainstream ranking when measuring regional stability and can satisfy the basic regional stability prediction function.Moreover,it may help us to better understand the prediction results with factor explanation.

Key words data analysis; machine learning;feature engineering;economic complexity