Statistics and Application 统计学与应用, 2024, 13(2), 430-436
Published Online April 2024 in Hans. https://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2024.132042

Hans汉斯

基于PSO-SOM算法的专利价值的分类研究

张美辰，金天颖，王可欣

燕山大学理学院，河北 秦皇岛

收稿日期：2024年3月25日；录用日期：2024年4月15日；发布日期：2024年4月23日

摘　要

根据世界知识产权组织统计的数据，我国的专利申请数量多年来名列前茅，专利授权数量也每年都在增长，因此学术界及社会各界越来越关注专利价值的研究。传统的专利价值评价方法大都是采用定性和定量结合的决策方法，比如：层次分析法、综合评价法等，已经不能满足当今大规模大体量数据的分析需求。利用大数据环境下的机器学习方法不但可以降低人力成本，还可以提高分类的准确率和效率。本文提出一种基于粒子群优化算法－自组织映射网络(PSO-SOM)的专利价值分类模型，依据专利价值指标，从incoPat专利数据库选取了5000条专利数据进行实证研究。通过PSO-SOM聚类得到了有效的专利价值标签，利用随机森林算法对初始专利价值进行指标重要性排序，并逐个依次将指标引入朴素贝叶斯模型中进行分类，能够有效提高朴素贝叶斯分类模型的准确率和效率。

关键词

粒子群优化算法，自组织映射网络，专利价值分类

Research on Patent Value Classification Based on PSO-SOM Algorithm

Meichen Zhang, Tianying Jin, Kexin Wang

College of Science, Yanshan University, Qinhuangdao Hebei

Received: Mar. 25th, 2024; accepted: Apr. 15th, 2024; published: Apr. 23rd, 2024

Abstract

According to statistics from the World Intellectual Property Organization, the number of patent applications in China has been among the top for many years, and the number of patent authorizations is also increasing every year. Therefore, the academic community and various sectors of society are increasingly concerned about the study of patent value. The traditional patent value evaluation methods mostly use a combination of qualitative and quantitative decision-making

文章引用：张美辰，金天颖，王可欣．基于 PSO-SOM 算法的专利价值的分类研究[J]．统计学与应用，2024，13(2): 430-436. DOI: 10.12677/sa.2024.132042

methods, such as the Analytic Hierarchy Process, Comprehensive Evaluation Method, and so on, which can no longer meet the analysis needs of large-scale and large-volume data today. The use of machine learning methods in the big data environment can not only reduce labor costs but also improve classification accuracy and efficiency. This article proposes a patent value classification model based on particle swarm optimization and self-organizing mapping network (PSO-SOM). Based on patent value indicators, 5000 pieces of patent data were selected from the incoPat patent database for empirical research. Effective patent value labels were obtained through PSO-SOM clustering, and the initial patent value was ranked in importance using the random forest algorithm. The indicators were introduced into the Naive Bayes model for classification one by one, which can effectively improve the accuracy and efficiency of the Naive Bayes classification model.

Keywords

Particle Swarm Optimization Algorithm, Self-Organizing Mapping Network, Patent Value Classification

1. 引言

专利作为一种知识产权，包含着丰富的信息。在知识经济背景下，知识产权竞争已逐渐成为全国乃至全球竞争的重要组成部分。专利作为知识产权的基本组成部分，已逐渐成为企业最重要的生产活动。专利价值的评估和分类可以为专利的申请、管理和维护提供重要的决策依据，包括政府投资、国家竞争力评估、企业技术转让和商业合并，这需要对专利的价值进行评估以支持决策。

为了对专利的价值进行分类，有必要采用一些科学的方法来分析专利的价值。当前人工智能驱动的技术革命和产业转型的新周期正在蓬勃发展，各领域正在加强大数据、云计算和机器学习等技术与原有行业的跨界融合。因此，数据挖掘和机器学习正逐渐被用于专利价值的评估和分类，构建一个自动分类系统以快速响应专利价值分类的需求，可以帮助投资者在新领域做出正确的决策选择。

由于 SOM 模型的聚类效果很大程度上取决于网络结构中参数设定是否合理，因此本文选用 PSO 算法来搜索 MiniSom 库中的两个关键参数：sigma 和 learning_rate，以最大化轮廓系数作为适应度函数，确定全局最优参数，以达到理想的聚类效果。本文的主要研究目的是尝试研究设计出一套基于 PSO-SOM 算法的专利价值分类策略，利用该机器学习算法构建并训练专利分类模型，以期实现高价值专利的计算机自动识别，并达到预期的分类效果，为专利资产管理和运营提供科学的决策保障。

2. 文献综述

2.1. 专利价值指标

影响专利价值的因素是多样性的[1]，学术各界对于专利价值指标的选择广泛关注。其中，Archontopouls [2]认为专利的法律状态，包括向公众开放的日期、其合规性等，能够体现专利不同阶段的价值；Dang 和 Motohashi [3]利用专利引用次数、影响指数、技术生命周期、科技关联度等指标衡量专利价值；Chen 和 Chang [4]选择具有代表性的医药公司，采用多元回归分析的方法，通过专利引用分析、专利维持年限来分析阐述企业专利价值与市场价值的关系；Frietsch 等[5]指出出口量在经济增长中贡献比很

大，因此专利的产品或技术的出口量可以反映专利价值，但是存在很大的局限性；Harhoff 等[6]认为专利文献引用次数、专利被引次数、专利技术的创新性和实用性、专利权利要求范围能有效代表专利的价值；Fischer 和 Leidinger [7]研究表明了专利被引数量与专利价值存在显著的正相关关系，李春燕和石荣[8]较完整地阐述了专利质量评价指标，认为可以从专利引用指标、专利引用科技文献数量、权利要求数、专利技术强度以及其他指标构建专利质量指标体系。

2.2. 专利价值评估方法

初期，大多数专利价值的评估侧重于层次分析法或指数期权等传统方法，选取的价值指标也大多集中在引用和专利家族等一些指标上，较少使用其他专利价值指标。此外，由于专利申请和授权量逐年在增加，专利数据逐渐向大数据发展演变，专利增长情况呈现出一种强劲的势头。一些现有的价值评估方法，比如层次分析法、综合评价法，在评估效率上明显不够，因此在专利价值评估的方法方面，存在着巨大的潜能。

伴随着机器学习和数据挖掘技术的蓬勃发展，以机器学习技术为基础的专利价值评估方法逐渐得到广泛关注和使用。近年来，机器学习在专利价值评估中的应用主要集中在分类问题领域，通过选取相关价值表现指标，使用分类算法构建分类模型，经过训练和学习，得到最终的专利价值评估模型。刘夏等[9]利用 85 万余条专利信息特征搭建了随机森林模型来预测专利的后续引用，但是特征变量仅包括专利的文档的基本信息和引证信息；李欣等[10]系统考虑了专利价值指标体系的多方面，提出基于机器学习的专利质量评估模型，以专利 3 年、5 年、10 年被引次数以及专利转让次数为质量衡量标准，分别用支持向量机、人工神经网络和随机森林构建了评价模型，发现基于机器学习的评估方法能够有效的识别专利质量；王思培和韩涛[11]构建了基于随机森林算法的潜在高价值专利预测模型；马鑫[12]探索了机器学习算法应用于专利创造性辅助判断的可能性；符川川等[13]提出基于机器学习的组合模型用于专利质量的分析与分类预测，该模型由自组织映射、核主成分分析以及支持向量机 3 种方法组成，并以新兴产业的区块链技术专利为例展开研究，但是自组织映射算法存在一定的局限性，如该算法需要神经元的权重足够必要且充分来对输入进行聚类，如果 SOM 提供在权重层中的信息太少或者是太多都会对结果产生较大的影响：容易陷入局部最优；可能会出现空簇现象等。

3. 变量指标与数据的获取

本文将专利价值和影响价值的相关因素作为研究变量，该变量的集合为一个价值评价指标体系。指标的选取对模型的训练和结果预测至关重要，因此本文秉承着构建的专利指标能够客观、准确、真实的反映出专利价值的真实状况这一宗旨，在选择专利价值指标时必须遵循一定的构建原则。

利用专家和学者们的研究经验，结合本文的研究特点，提出以下准则：第一，客观性。在选择专利价值指标时，需要尽可能确保指标的客观性。本文选择可以直接从专利数据库中提取、具有原始数据、不会产生异议的指标，以确保专利价值分类的准确性。第二，科学性。在选择专利价值的影响指标时，要确保指标的有效性、合理性和科学性。每个指标的数据来源都应该可靠准确，每个指标都应该能够提供一个对专利价值合理的解释。第三，经济性。选择的专利价值评估预测指标必须注重过程的经济性，指标数据的收集和分析不应导致过多的时间和经济负担。

因此，本文从已有文献研究中梳理出可以量化的专利价值影响因素，基于 incoPat 专利数据库搜集可以获得的指标数据，拟选取影响专利价值的 13 个因素。同时，在中国，由于专利从申请到公告至少需要 18 个月，2022 年的专利数据可能不全，因此，本文选取了 2021 年的专利；由于所有专利均来自于 2021 年，所以专利被引的概率是一致的，消除了时间因素的影响。本文下载了获得授权的 5000 条数据，具体的指标名称及描述说明如表 1 所示。

Table 1. The name and description of the patent value indicators
表 1. 专利价值指标名称及说明

指标名称	描述说明
文献页数	专利文献的页数
专利施引数	专利对已有专利引用的数量
发明人数	专利发明人的数量
专利引用数	其他专利对该专利的引用次数
权利要求数	表示专利的保护范围
独立权利要求数	指无需用其他权利要求来确定其范围和含义的完整权利要求
授权时长	从申请到获得授权的时间
简单同族个数	指有完全相同的优先权的专利文献构成的一族专利的数量
首权字数	首项权利要求的字数
IPC 分类号数	专利的 IPC 分类数
优先权期限	从优先权到获得授权的时间
同族国家地区数	指的是与某一专利申请或专利相关的一组专利申请或专利所覆盖的国家或地区的数量
扩展同族个数	指的是与某一专利相关的所有同族专利的数量

4. 模型的建立与研究

4.1. 数据预处理

本文利用了 Python 中的 Pandas 库的 dropna ()、fillna ()以及 astype ()等方法对原始数据集进行处理和清洗，删除掉了重复数据、使用均值填补了数据集中的残缺值；然后，对数据进行标准化处理，在 python 中安装 Sklearn (Scikit-learn)工具包后，直接调用下 preprocessing 模块中的 MinMaxScaler 函数，设置 MinMaxScaler 函数中 feature_range 参数，可以手动设置或者选择默认值，默认值为 0，1。输入数据，将数据收敛到 0 到 1 之间。在对专利价值指标进行后续的分类时，对原始数据进行归一化处理能够减弱或消除量纲对评估结果的影响，提高模型运行速度，避免过于拟合，提升模型的性能。

4.2. PSO-SOM 聚类分析

PSO 算法的基本思想是在搜索的过程中不断进行自我改进，在个体与种群的交互中，寻求高质量的解。当使用粒子群优化(PSO)算法来调整 MiniSom 模型的参数时，需要选择适当的参数作为优化目标。通常，MiniSom 模型有两个关键参数需要调整。第一个，sigma：SOM 模型的半径参数，它控制了神经元的邻域大小。较大的 sigma 值会导致更广泛的邻域，较小的值会导致更小的邻域。这个参数影响 SOM 模型的拓扑结构。第二个，learning_rate：学习率参数，控制了神经元权重的更新速度。较大的 learning_rate 值会导致更快的权重更新，较小的值会导致更慢的权重更新。

使用 PSO 算法来搜索合适的 sigma 和 learning_rate 的值，以最大化轮廓系数作为适应度函数。此时，得到全局最优参数 sigma 和 learning_rate 分别为 0.4 和 0.8，然后，使用全局最优位置创建并训练 MiniSom 模型，我们可以得到聚类数为 3 时，聚类效果最好，距离地图如图 1 所示。

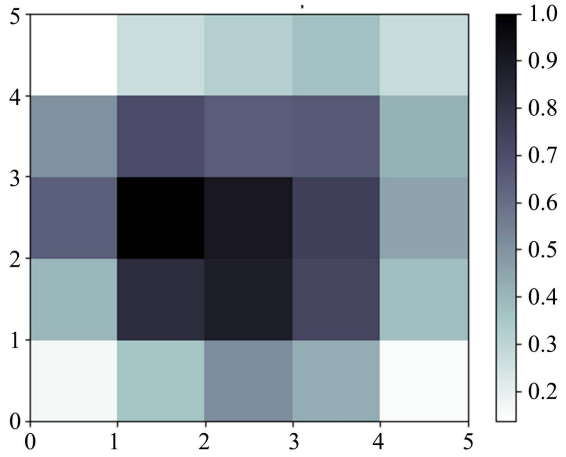

Figure 1. Distance map
图 1. 距离地图

为了区分专利价值的高低，计算各专利价值类别在全指标下的均值，同时为方便专利价值聚类结果易于被理解，将标准化后的均值结果进行离散化处理，将数据集中所有指标离散成高、中、低三类。依据已有相关文献中对于专利价值影响指标的研究，认为 IPC 分类号数、专利引用次数、同族个数、同族国家地区数等对专利价值存在显著的积极影响，因此对离散化后的聚类结果进行分析，形成了三个专利价值类别，如表 2 所示。其中，1 代表低价值专利，2 代表高价值专利。专利价值指标及符号表示依次为：授权时长(A_1)，优先权期限(A_2)，权利要求数量(A_3)，独立权利要求数量(A_4)，文献页数(A_5)，首权字数(A_6)，IPC 分类号数(A_7)，发明人数量(A_8)，专利引用次数(A_9)，专利施引数(A_{10})，简单同族个数(A_{11})，扩展同族个数(A_{12})，同族国家地区数(A_{13})。

Table 2. The discretization result of patent value category
表 2. 专利价值类别离散化结果

类别	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}	A_{11}	A_{12}	A_{13}
0	中	中	中	中	高	低	中	高	中	高	中	中	中
1	中	中	低	低	中	中	中	中	中	中	中	中	中
2	高	高	中	中	中	中	高	中	高	中	高	高	高

4.3. 指标重要性排序

运用随机森林算法对 13 个专利价值指标进行重要性排序，调用 Python 中的 RandomForestClassifier 来实现，模型中树的数量定为 300，实验结果如下表 3 所示。

Table 3. The importance ranking of patent value indicators
表 3. 专利价值指标重要性排序

排名	指标名称	排名	指标名称
1	文献页数	8	专利施引数
2	IPC 分类号数	9	优先权期限
3	独立权利要求数	10	发明人数

续表

4	权利要求数	11	授权时长
5	扩展同族个数	12	首权字数
6	简单同族个数	13	专利引用数
7	同族国家地区数		

4.4. 构建朴素贝叶斯分类模型

由于本文选取的专利价值指标数量较多，可能有部分指标对于专利价值的分类是多余的。所以，尝试对价值指标进行合理适当地删减，以便能提高分类器的准确率。依据表 3 的指标重要性排名，将专利价值指标依次逐个引入，从而得到不同指标个数下的分类准确率，如图 2 所示。

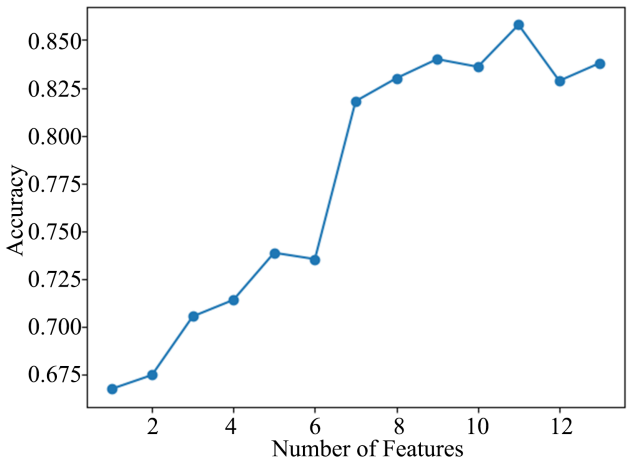

Figure 2. The classification performance of Naive Bayes model
图 2. 朴素贝叶斯模型的分类性能

可以看出，在引入所有指标训练模型时，分类准确率为 83.8%，逐个依次加入指标时，分类准确率呈现出了先增后降的趋势。当引入指标为 11 个时，分类准确率达到最大 85.8%，优于引入所有指标的情况，说明最初的专利价值指标内存在冗余，降低了分类的准确率，影响分类器的分类性能。因此本文将表 3 的前 11 个指标引入模型作为最终专利价值分类模型的指标。

依据上述结果得到的专利价值指标的最优特征集，用全部 5000 条数据对朴素贝叶斯模型进行训练，按 7:3 的比例划分训练集和测试集，得到样本训练集 3500 条以及测试集 1500 条。依据指标和数据集训练模型，得到该模型产生的混淆矩阵结果如表 4 所示。

Table 4. Confusion matrix of Naive Bayes classification model
表 4. 朴素贝叶斯分类模型的混淆矩阵

	预测值	预测值	预测值
实际值	0	1	2
0	300	25	158
1	7	63	0
2	43	4	900

依据表 4 可以得出，朴素贝叶斯模型的分类准确率为 84.2%。

为测试朴素贝叶斯分类模型的准确性，对该分类模型进行十折交叉验证，将 5000 条数据均分成 10 个随机选择的子集，每个子集包含 500 条数据，轮流将其中 9 份作为训练数据，1 份作为测试数据，进行试验，共计进行 10 次试验，得到的朴素贝叶斯分类模型十折交叉验证的结果如表 5 所示。

Table 5. Cross-validation results of Naive Bayes classification model
表 5. 朴素贝叶斯分类模型交叉验证结果

组别	准确率	召回率	F1 值	组别	准确率	召回率	F1 值
Fold1	83.80%	83.80%	82.86%	Fold6	83.00%	83.00%	82.14%
Fold2	83.20%	83.20%	82.07%	Fold7	83.40%	83.40%	82.45%
Fold3	83.80%	83.80%	82.73%	Fold8	82.60%	82.60%	81.55%
Fold4	84.60%	84.60%	83.59%	Fold9	83.40%	83.40%	82.44%
Fold5	83.40%	83.40%	82.97%	Fold10	82.80%	82.80%	82.31%

依据表 5 可以得出朴素贝叶斯算法在本专利价值数据集上的平均分类准确率为 83.40%，召回率的平均值为 83.40%，F1 值的平均值为 82.51%。

5. 结论

本文提出一种基于粒子群优化算法－自组织映射网络(PSO-SOM)的专利价值分类模型，从 incoPat 专利数据库随机选取 5000 条专利数据进行实证研究。实验结果表明通过 PSO-SOM 聚类得到了有效的专利价值标签，利用随机森林算法对初始专利价值指标约简后，提高了朴素贝叶斯分类模型的准确率和效率。通过该模型可以对专利的价值进行科学合理的分类，为企业研发工作提供支持。

参考文献

[1] Narin, F. (1994) Patent Bibliometrics. *Scientometrics*, **30**, 147-155. https://doi.org/10.1007/BF02017219

[2] Archontopoulos, E. (2004) Prior Art Search Tools on the Internet and Legal Status of the Results: A European Patent Office Perspective. *World Patent Information*, **26**, 113-121. https://doi.org/10.1016/j.wpi.2003.08.004

[3] Dang, J.W. and Motohashi, K. (2015) Patent Statistics: A Good Indicator for Innovation in China? Patent Subsidy Program Impacts on Patent Quality. *China Economic Review*, **35**, 137-155. https://doi.org/10.1016/j.chieco.2015.03.012

[4] Chen, Y.S. and Chang, K.C. (2010) The Relationship between a Firm's Patent Quality and Its Market Value—The Case of US Pharmaceutical Industry. *Technological Forecasting & Social Change*, **77**, 20-33. https://doi.org/10.1016/j.techfore.2009.06.003

[5] Frietsch, R., Neuhäusler, P., Jung, T., *et al*. (2014) Patent Indicators for Macroeconomic Growth—The Value of Patents Estimated by Export Volume. *Technovation*, **34**, 546-558. https://doi.org/10.1016/j.technovation.2014.05.007

[6] Harhoff, D., Scherer, F.M. and Vopel, K. (2003) Citations, Family Size, Opposition and the Value of Patent Rights. *Research Policy*, **32**, 1343-1363. https://doi.org/10.1016/S0048-7333(02)00124-5

[7] Fischer, T. and Leidinger, J. (2014) Testing Patent Value Indicators on Directly Observed Patent Value—An Empirical Analysis of Ocean Tomo Patent Auctions. *Research Policy*, **43**, 519-529. https://doi.org/10.1016/j.respol.2013.07.013

[8] 李春燕, 石荣. 专利质量指标评价探索[J]. 现代情报, 2008, 28(2): 146-149.

[9] 刘夏, 黄灿, 余骁锋. 基于机器学习模型的专利质量预测初探[J]. 情报学报, 2019, 38(4): 402-410.

[10] 李欣, 范明姐, 黄鲁成. 基于机器学习的专利质量评价研究[J]. 科技进步与对策, 2020, 37(24): 116-124.

[11] 王思培, 韩涛. 基于随机森林算法的潜在高价值专利预测方法研究[J]. 情报科学, 2020, 38(5): 120-125.

[12] 马鑫. 机器学习算法在专利创造性辅助判断中的应用研究[J]. 中国发明与专利, 2021, 18(9): 70-79.

[13] 符川川, 陈国华, 袁勤俭. 基于机器学习的专利质量分析与分类预测研究——以区块链技术专利为例[J]. 现代情报, 2021, 41(7): 110-120.