Service Science and Management 服务科学和管理, 2019, 8(6), 227-233
Published Online November 2019 in Hans. http://www.hanspub.org/journal/ssem
https://doi.org/10.12677/ssem.2019.86034

Hans汉斯

Research of Patent Infringement Early Warning Based on Sample Equilibrium Technology

Yang Liu, Li Zhang

School of Information Technology and Management, University of International Business and Economics, Beijing
Email: zhangli_amy@uibe.edu.cn

Received: Oct. 14th, 2019; accepted: Oct. 24th, 2019; published: Oct. 31st, 2019

Abstract

The protection of intellectual property rights affects the creativity of researchers and development power of enterprises. It is an important driving force and legal guarantee for innovation. Patent infringement warning is an important component of intellectual property protection. A large number of patent infringement lawsuit data are used to analyze the model of patent infringement and to discover the risks of patent infringement. Based on the patent-related feature information, this paper constructs a patent early warning model based on sample equalization technology, and compares the performance of random forest, Bayesian network, neural network, decision tree model, logistic regression model and Support Vector Machine (SVM) algorithm. The experimental results show that the random forest model can obtain better early warning effect after sample equilibrium, and can better discover the infringement litigation relationship between the company and the patent, thus effectively realizing the function of patent infringement warning.

Keywords

Patent Early Warning, Intellectual Property, Sample Equilibrium, Random Forest

基于样本均衡技术的专利侵权预警研究

刘　洋，张　莉

对外经济贸易大学信息与管理学院，北京
Email: zhangli_amy@uibe.edu.cn

收稿日期：2019年10月14日；录用日期：2019年10月24日；发布日期：2019年10月31日

文章引用: 刘洋, 张莉. 基于样本均衡技术的专利侵权预警研究[J]. 服务科学和管理, 2019, 8(6): 227-233.
DOI: 10.12677/ssem.2019.86034

摘　要

知识产权的保护影响科研人员的创造力和企业的研发动力，是创新的重要驱动力和法律保障，专利侵权预警是知识产权保护的重要组成，而大量的专利侵权诉讼案件数据为挖掘、分析专利侵权行为模式，发现专利侵权的风险提供了基础。本文利用专利相关的特征信息，基于样本均衡技术构建了专利预警模型，对比分析了随机森林、贝叶斯网络、神经网络、决策树模型、逻辑回归模型和SVM算法的性能。实验结果表明，随机森林模型在样本均衡后可取得更好的预警效果，能够更好地发现公司与专利之间的侵权诉讼关系，从而有效地实现专利侵权预警的功能。

关键词

专利预警，知识产权，样本均衡，随机森林

1. 引言

随着社会经济的发展和科研创新能力的提高，创新驱动逐渐成为新时代经济发展的重要驱动力，自主研发的专利也越来越多，2013 年至 2018 年以超过 13%的年复合增长率在快速增加。但是，相应的专利侵权诉讼案件越来越多，2013 年至 2018 年全国专利侵权纠纷立案数量年复合增长率高达 34%。专利的侵权行为会侵害科研组织的权益，带来直接或间接的经济损失，进而伤害了专利拥有者的创新积极性，直接破坏企业的竞争力，进而破坏整体经济发展和创新环境。因此，对专利侵权行为预警可以减少企业专利保护和维权成本，促进创新驱动的发展，对企业的自身管理经营和国家新时代的结构优化型经济发展具有直接的帮助。

国内外学者在专利侵权的影响因素和专利预警方法方面进行了大量研究，取得了许多有价值的研究成果。崔胜男(2013)提出了专利预警系统的工作体系，通过指标和系统分析等定性方法进行专利预警，为专利预警系统框架的构建提供了理论参考[1]，另外一些学者开始将逻辑回归等数据分析方法应用到专利预警模型构建(2011) [2]。但是专利侵权数据存在样本不均衡的问题，侵权样本数据占比只有 2%左右，目前的研究缺少对本领域样本不均衡的问题的关注，并且构建预警模型的影响因素选取也没有统一。为此，本文借鉴已有的专利侵权影响因素的研究结论，构建了基于样本均衡技术的专利侵权预警模型，并通过实验对比分析了不同模型的预警性能。一方面本文构建的模型可以提高企业预警自身专利风险能力，推动知识产权保护技术的提高，另一方面，也为数据挖掘技术在本领域的应用提供了参考。

2. 相关工作

目前，国内外学者在专利侵权预警方面的研究可以划分为两个分支，一是专利侵权的影响因素，二是专利侵权预警模型，为此我们从这两方面对相关研究进行分析。

2.1. 专利侵权的影响因素

学者们从不同的视角，分析了专利侵权的影响因素。尹志锋(2018)等人通过实证研究发现被专利侵权

的组织通常拥有较多专利且专利质量较高，专利实施水平较高[3]。漆苏(2014)从专利价值评估的视角，通过实证研究发现专利维持费、后向引用、权利要求数量、向前引用等特征对侵权风险影响较大[4]。张世玉(2016)将影响企业技术威胁的技术差距、技术成熟度和技术研发重点等因素融入技术威胁预警模型中[5]。贺宁馨(2016)以我国 605 件专利侵权诉讼案为样本，基于多元回归分析方法研究我国专利侵权赔偿额的主要影响因素，发现专利侵权时间、侵权人数等 12 个变量对赔偿额有显著影响[6]。张军荣(2018)以行为主义为理论视角，运用多元回归方法对 198 份判决进行了实证研究，回归分析结果表明，专利技术复杂度、被告经济实力等非法定事实是影响法官裁判的重要因素[7]。Lanjouw 和 Schankerman (2001)研究了涉及诉讼的专利以及专利拥有者的特征，发现不用的专利和拥有者在专利侵权的风险上有很大的差异[8]。Chien (2011)研究了影响专利诉讼的影响因素，没有关注专利的内在方面，而是专注于专利申请成功后才产生的专利信息，其中包括专利所有权、专利的后续投资、专利保护和专利引用[2]。Cremers (2004)统计分析了前引数量、权利要求数量和家族规模大小来确定最容易发生专利侵权的相关特征[9]。Lim (2014)更详细地分析了原告与被告公司之间的引用与专利侵权之间的关系，发现间接引用和潜在引用对专利诉讼的影响要大于直接引用[10]。Marco (2017)等对专利审查质量与专利诉讼之间的关系进行了研究，发现一些审查特征也可以用来预测专利侵权的风险[11]。Jin (2016)等开发了一个协同过滤框架来预测某一特定行业类别的高科技企业的诉讼风险[12]。有关学者的研究总结如表 1。

Table 1. Characteristic factor study comparison table
表 1. 特征因素研究对照表

作者(年份)	影响因素	主要结论
尹志锋(2018)	拥有的专利数量，专利质量，专利实施水平	被专利侵权的组织通常拥有较多专利且专利质量较高，专利实施水平较高
漆苏(2014)	专利规模大小、后向引用和专利家族，专利权维持、专利授权后审查、向前引用、专利权交易和专利权融资担保	通过实证研究发现专利维持费、后向引用、权利要求数量、向前引用等特征对侵权风险影响较大
张世玉(2016)	技术差距、技术成熟度和技术研发重点等	通过技术差距等三个影响因素实现专利侵权预警
张军荣(2018)	专利技术复杂度、被告经济实力等非法定事实	专利技术复杂度、被告经济实力等非法定事实是影响法官裁判的重要因素
Chien (2011)	专利所有权、专利的后续投资、专利保护和专利引用	检验了其模型的可行性
Cremers (2004)	前引数量、权利要求数量和家族规模大小	从统计的角度尝试确定最容易发生专利侵权的相关特征
Lim (2014)	直接引用，间接引用和潜在引用	间接引用和潜在引用对专利诉讼的影响要大于直接引用

根据前人的研究来看，在专利侵权的相关研究中已经涵盖了许多的影响因素，比如专利的前引数量、后引数量、专利家族规模大小、专利质量、专利技术含量和专利寿命等众多角度的影响因素。但是专利技术含量等影响因素主要依靠主观判断而难以有一个统一的衡量标准，而专利发明人数和专利后续投资等可明确量化的影响因素在专利相关的其他领域极少有相关研究，并不能有效地描述专利的特征，因此本研究选取有明确量化计算标准的、在其他专利相关研究领域也经常用来描述专利特征的专利前引数量、专利后引数量和专利家族规模大小三个影响因素来描述专利的特征。

2.2. 专利预警模型的研究

Su H. N. (2012)等发现侵权专利和非侵权专利在申请人数量、申请国家、专利引用等方面存在显著不

同，利用 USPTO 的实际数据进行逻辑回归，得到相关函数曲线，以此来进行专利侵权的预测和专利价值的评估[13]。Changyong Lee (2013)等采用分层关键词向量来表示索赔元素之间的依赖关系，采用树匹配算法来比较专利索赔要素，通过语义专利索赔分析来进行专利侵权的风险测量[14]。Ganlu Su (2015)等提出了有关专利技术危机特征的四个维度：技术稳定，技术垄断，技术安全和技术前景，并采用层次分析法进行权重分析，其中技术稳定性权重为 0.0507，技术垄断权重为 0.1781，技术安全权重为 0.5539，技术前景权重为 0.2172，并采用多目标线性加权函数来进行专利侵权风险的评分以衡量企业面临的技术危机程度[15]。

Chien (2011)使用逻辑回归、时间序列分析对诉讼专利和非诉讼专利进行分析，通过分析二者的不同，认为专利的内在特点可以对专利侵权有一定的预警性，可以缩小企业重点观测可能被侵权的专利范围，但 Chien 的模型只是一个十分粗略的模型[2]。Petherbridge (2012)通过使用贝叶斯理论，认为 Chien 通过回归分析进行专利预警存在着较大偏差，由于存在假阳性结果，在准确性方面存在着局限性，因此可能会造成“安全错觉”，且 Chien 回归中的信息可能存在不可得的问题，因此在实用性方面也存在一定的局限性[16]。Kesan (2012)为 Chien 的工作提供了后续研究，认为 Chien 使用的数据仅限于 1990 年的专利，不具有代表性，并提出在专利所有权、重新检查、专利所有权转移时间等数据的数据优化方法，同时阐述之后的研究中应当对不同规模和类型的专利进行区分研究，也应考虑多重专利侵权的情况[17]。

崔胜男(2013)通过总结前人的研究提出了专利预警系统的工作体系，通过指标和系统分析等定性方法进行专利预警[1]。漆苏(2014)通过使用回归分析的方法，发现专利侵权不仅与专利授权时的品质有关，也和专利运行时的品质有关[4]。张世玉(2018)以战略三角模型为理论基础，从企业顾客、企业自身和竞争对手三个维度确定了企业专利威胁的评估维度，构建了企业专利威胁预警的 P-L 模型，从产品的需求，周期和份额来研究专利威胁预警，通过实证研究证明了其模型的实用性[18]。曹亚莎(2018)使用 ANP-SWOT 方法分析了湖南省粮油行业专利发展现状，明晰行业优势以及机遇，发现可能面临的专利风险[19]。马治国(2019)构建了区域知识产权保护环境评价体系，包括知识产权立法保护、执法保护、司法保护、公众意识和区域发展五个维度的一级指标和 27 项二级指标，采用因子分析法提取出 7 个影响程度较大的综合评价因子，结果发现我国不同地区差别很大，并给出了提升的建议[20]。

陶新民(2013)提出一种基于样本特性欠取样的不均衡 SVM 分类算法，通过选择最具有代表性的样本点来提高实验效果，能够提高 SVM 算法在失衡数据中少数类的分类性能、总体分类性能和鲁棒性[21]。徐剑(2019)为了解决 SVM 算法难以处理样本不均衡数据的问题，SSIC 不均衡分类框架，自适应从大小类中选取有价值样本，并结合代价敏感学习构建不均衡数据分类器，并且通过实证验证了该框架的实用性[22]。

国内外的学者已经在专利侵权预警及相关领域做了大量的研究，使用了逻辑回归、系统分析、指标分析和语义分析等多种方法，取得了许多研究成果。但是在现有的研究中，并没有解决样本不均衡的问题。因此，本研究尝试在专利预警研究中引入随机森林等多种数据分析方法，并且融入在其他领域中开始应用的样本选择和虚拟样本构造的样本均衡方法来解决专利侵权预警领域的实际问题。

3. 专利侵权预警分析

本研究所使用的数据来源于 Unifiedpatents (https://www.unifiedpatents.com)，Unified Patents 是一个由 200 余个组织机构共同组成的专利组织，目的是为了提升专利质量与专利分析能力。从网站上获取了 120 家公司，209 个专利，共 25080 专利样本数据，其中专利侵权数据 542 个，占比 2.16%，对获取的原始数据进行清洗、匹配和归一化等数据预处理后，得到的数据描述性特征如表 2 所示。

Table 2. Data descriptive statistics
表 2. 数据描述性统计

影响因素	最小值	最大值	平均值	方差	偏度	峰度	中位数
家族规模大小	0	1	0.067	0.034	4.351	18.849	0.000
引用个数	0	1	0.044	0.013	6.073	43.781	0.021
被引用次数	0	1	0.070	0.037	4.096	16.002	0.012
公司被侵权次数	0	1	0.192	0.047	2.394	6.367	0.119
公司侵权次数	0	1	0.213	0.055	1.463	1.601	0.113
公司涉及专利数	0	1	0.057	0.010	4.960	37.493	0.021

3.1. 样本均衡

由于本数据集存在较为明显的样本不均衡问题，因此本研究进行样本均衡操作，以求得到更好的实验效果。对所有样本进行训练集和测试集划分后，首先对训练集样本较多的一类进行随机的样本选择，并通过两部聚类和样本随机选择筛选排除掉一部分特征非常类似的样本，然后对样本较少的一类进行虚拟样本构造操作和和进行虚拟样本构造操作，最终对不同操作的实验结果进行对比。详细的步骤如表 3。

Table 3. Sample equalization operation detailed steps
表 3. 样本均衡操作详细步骤

STEP1	对所有数据进行区分，30%的数据进入测试集，其余 70%进入待定训练集
STEP2	对待定数据集中的数据按照分类结果的值进行区分，样本较少的部分进入预备训练集，样本较多的部分则进入临时训练集
STEP3	对临时训练集中的数据进行两步聚类，聚类结果数量较少的数据划为 A 类训练集，其他样本数据为 B 类训练集
STEP4	随机取 B 类训练集中的 30%数据，与所有 A 类训练集、预备的数据集共同构成最终的训练集
STEP5	对上述最终训练集和测试集，分别采用预备训练集部分的数据权重增加四倍的样本均衡操作和不进行样本均衡操作

3.2. 实验过程与性能分析

本实验采用决策树、随机森林、类神经网络、贝叶斯网络、SVM 和逻辑回归数据分析方法，采用专利前引数量、后引数量、家族规模数量专利影响因素，采用公司侵权专利次数，公司的专利被侵权次数和公司涉案总数等统计性因素作为影响因素，通过以下实验步骤进行专利预警实验研究。

首先，选择在精确率，召回率和 F 值作为实验结果的评价指标，精确率常用来衡量模型的效率，在本问题中可以用来衡量专利侵权相关利益的收益和成本的关系；召回率常用来衡量模型的捕获能力，在本问题中用来衡量模型对潜在或已经发生的侵权关系的捕获能力，能够衡量模型的实际意义。而单考虑以上两个指标有一定的局限性，因此要引入 F 值来平衡以上两个指标来综合衡量模型的实用价值。

其次，对原始数据区分训练集和测试集，随后进行样本选择、聚类筛选等操作，得到最终训练集，使用此训练集进行训练，并用最初的测试集进行测试，得到实验结果。

最后，对原始数据区分训练集和测试集，随后进行样本选择、聚类筛选等操作，并加上虚拟样本构造操作完成训练集的样本均衡操作，使用此训练集进行训练，并用最初的测试集进行测试，得到实验结果，实验结果如表 4 所示。

Table 4. Experimental result
表 4. 实验结果

	样本均衡前			样本均衡后		
算法	Precision	Recall	F-score	Precision	Recall	F-score
逻辑回归	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
SVM	0.00%	0.00%	0.00%	60.81%	2.41%	4.63%
类神经网络	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
贝叶斯网络	0.00%	0.00%	0.00%	16.67%	0.60%	1.15%
决策树	0.00%	0.00%	0.00%	4.94%	25.00%	8.25%
随机森林	4.93%	56.55%	9.06%	4.77%	66.07%	8.90%

从表 4 可以看出：未做样本均衡的一组实验结果很差，几乎不能找到真实侵权的样本；在这一组实验结果中，只有随机森林可以准确找到一部分真实侵权的样本数据，因此在样本不均衡的专利侵权预警问题中，未进行完整样本均衡操作的算法很难找到目标侵权关系，对实际问题的解决难有贡献。而在进行了样本均衡操作之后，SVM、贝叶斯网络、决策树和随机森林都能够找到一部分真实侵权样本数据，两组实验结果对比发现，样本均衡了提升专利侵权预警的效果，对实际问题的解决提供了行之有效的方案。同时，在各类算法的对比中，随机森林样本均衡前后都有最好的表现，即集成分类器比单分类器会有更好的表现。最后，从召回率来看表现最好的是随机森林算法，从精确率来看最好的是 SVM 算法，两者的最好表现都出现在样本均衡之后，因此在实际问题的解决中，根据专利保护和维权的成本，以及专利所带来的经济利益和法律效益综合考虑可以根据企业不同的情况采取不同的侵权预警方法，进而提升企业的创新动力和自我保护机制。

4. 结论

本研究通过使用随机森林，决策树，类神经网路、贝叶斯网络、SVM 和逻辑回归算法对实际专利侵权数据进行了实验。实验结果表明，随机森林最适用于此领域的实际问题，在样本均衡后会取得更好的效果。同时，较高的召回率可以在实际问题中更有效地排查出潜在的专利侵权危险，进而提升专利侵权预警能力。

本研究的主要意义有二：首先，本研究将以上多种分类方法引入专利侵权预警领域，并且发现随机森林算法较适应本应用领域的实际问题，为后续的研究提出了新的思路。其次，本研究的最优实验结果现实，可以做到超过 60%的召回率，这对于实际生活中的专利侵权预警问题有十分重要的意义，可以避免大部分的专利侵权损失和伤害。

本研究目前存在以下几点不足：首先，本研究所选取的专利特征都是结构化的专利特征，尚且不能够非常有效地表征专利侵权的具体特征；其次，本研究使用的算法虽然开始进行量化的研究，但是尚且不能完全解决本领域数据稀疏性的问题，希望在未来的研究中可以找到适当的算法来解决数据稀疏性问题。

参考文献

[1] 崔胜男, 田玲. 我国专利预警理论研究概述[J]. 科技情报开发与经济, 2013, 23(14): 148-152.

[2] Chien, C.V. (2011) Predicting Patent Litigation. *Texas Law Review*, **90**, 283-329.

[3] 尹志锋, 邓仪友. 中国企业的专利侵权特征及维权策略研究[J]. 经济管理, 2018, 40(3): 5-21.

[4] 漆苏. 企业国际化经营专利风险因素——基于专利属性的实证研究[J]. 科研管理, 2014, 35(11): 139-145.

[5] 张世玉, 王伟, 陶成琳, 刘思蓓. 企业技术威胁预警模型构建——基于专利组合分析的视角[J]. 情报杂志, 2016, 35(11): 70-74.

[6] 贺宁馨, 李黎明. 我国专利侵权赔偿额的影响因素及预测研究[J]. 科研管理, 2016, 37(10): 137-145.

[7] 张军荣. 专利复杂度、被告实力与侵权赔偿责任承担[J]. 科研管理, 2018, 39(11): 116-121.

[8] Schankerman, L.M. (2001) Characteristics of Patent Litigation: A Window on Competition. *The RAND Journal of Economics*, **32**, 129-151. https://doi.org/10.2307/2696401

[9] Cremers, K. (2004) Determinants of Patent Litigation in Germany. ZEW Discussion Papers 04-72. https://doi.org/10.2139/ssrn.604467

[10] Lim, J. (2014) Analysis of the Relationship between Patent Litigation and Citation: Subdivision of Citations. *Applied Mathematics & Information Sciences*, **8**, 2515-2522. https://doi.org/10.12785/amis/080549

[11] Marco, A.C. and Miller, R. (2017) Patent Examination Quality and Litigation: Is There a Link? Social Science Electronic Publishing, Rochester. https://doi.org/10.2139/ssrn.2995698

[12] Jin, B., Che, C., Yu, K.F., Qu, Y., Guo, L., Yao, C.L., Yu, R.Y. and Zhang, Q. (2016) Minimizing Legal Exposure of High-Tech Companies through Collaborative Filtering Methods. 22*nd ACM SIGKDD International Conference*, San Francisco, 13-17 August 2016, 127-136. https://doi.org/10.1145/2939672.2939708

[13] Su, H.N., Chen, M.L. and Lee, P.C.P. (2012) Patent Litigation Precaution Method: Analyzing Characteristics of US Litigated and Non-Litigated Patents from 1976 to 2010. *Scientometrics*, **92**, 181-195. https://doi.org/10.1007/s11192-012-0716-7

[14] Lee, C., Song, B. and Park, Y. (2013) How to Assess Patent Infringement Risks: A Semantic Patent Claim Analysis Using Dependency Relationships. *Technology Analysis and Strategic Management*, **25**, 23-38. https://doi.org/10.1080/09537325.2012.748893

[15] Sun, G.L., Guo, Y. and Yang, F. (2015) Technology Early Warning Model: A New Approach Based on Patent Data.

[16] Petherbridge, L. (2012) On Predicting Patent Litigation. *SSRN Electronic Journal*, **90**, 283-329. https://doi.org/10.2139/ssrn.1981798

[17] Kesan, J.P., Schwartz, D.L. and Sichelman, T.M. (2012) Paving the Path to Accurately Predicting Legal Outcomes: A Comment on Professor Chien's Predicting Patent Litigation. Social Science Electronic Publishing, Rochester.

[18] 张世玉, 王伟, 贾宇希, 姜钰莹. 企业专利威胁预警的 P-L 模型构建[J]. 情报理论与实践, 2018, 41(5): 5-10.

[19] 曹亚莎, 谭洁, 王奎武, 伍小松. 基于 ANP-SWOT 模型的湖南粮油产业发展战略[J]. 湖南农业科学, 2017(10): 86-90+94.

[20] 马治国, 秦倩. 中美贸易摩擦背景下中国区域知识产权保护环境的评价与优化[J]. 西安交通大学学报(社会科学版), 2019, 39(5): 1-13.

[21] 陶新民, 郝思媛, 张冬雪, 李震. 基于样本特性欠取样的不均衡支持向量机[J]. 控制与决策, 2013, 28(7): 978-984.

[22] 徐剑, 王馨月, 才子昕, 沈启航, 景丽萍. 价值样本选取的不均衡分类[J]. 计算机科学与探索, 1-11.