专利价值评估与分类研究*
——基于自组织映射支持向量机

周　成　魏红芹
(东华大学旭日工商管理学院　上海 200051)

摘要: 【**目的**】充分利用专利数据, 研究专利价值评估和分类问题。【**方法**】根据专利的价值指标, 设计基于自组织映射(SOM)–支持向量机(SVM)的专利价值评估及分类模型, 使用自组织映射方法确定专利的价值类别, 采用随机森林(RF)对价值指标进行重要性排序, 并结合包裹式特征选择方法对价值指标进行约简, 以提高 SVM 的分类性能。【**结果**】通过 SOM 确定的价值标签能有效反映专利价值的高低; 同时, 约简后的指标由初始的 14 个减少到 10 个, 分类准确率由 76.28%提高到 86.89%。【**局限**】对每个类别中的专利价值没有细化, 专利价值指标存在进一步约减的可能。【**结论**】本文方法能够为专利研发活动提供支持, 避免过度依赖专家判断。

关键词: 专利价值评估　数据聚类　专利价值分类　特征选择　自组织映射　支持向量机
分类号: TP181　G306
DOI: 10.11925/infotech.2096-3467.2018.0674

1　引　言

拥有较多知识产权的企业或个人, 希望知道哪些专利值得投资, 并在后续过程中获得收益。目前专利的申请和授权数量每年都在增长, 专利体系中的审查与批准对不同行业, 特别是高新技术产业都有着重要作用。专利价值的研究也受到学术界和工业界越来越多的关注, 专利价值分析为企业决定是否进行新产品生产提供了一种便捷方法, 但是如何评估和预测新专利的价值给研究人员和企业人员带来新的挑战。现有的专利价值评估技术仍处于比较初级的阶段, 很多评估工作仍然由行业专家手工完成, 评估的准确性和效率不能保证。

目前, 学界及业界对专利的分析主要集中于专利价值的分析, 高价值的专利可以提高新产品研发的成功率[1]。专利价值评估包括对新颖性、诉讼、技术趋势等方面的分析[2]。然而, 传统的专利分析需要花费大量时间与人力, 分析成本也趋于高位, 市场对于缩短潜在高价值专利分析时间的需求愈发强烈。经验的专利分析方法一般是统计分析或指标计算。近期, 基于专利特征的聚类方法被广泛应用到专利集群分析中[3]。统计分析的方法虽然能帮助分析者了解专利的当下技术环境, 但不能为决策提供有效帮助。未来的专利评估主要是对初期专利(获得授权一年到两年的专利)的评估, 因为专利会对行业的发展产生影响。

知识产权局每年都会授权大量专利, 现有专利系统在评估这些专利价值时存在明显不足。传统的学者或分析人员更多地关注于各种专利价值指标, 这些指标大都来自专利数据库, 比如专利引用数量、IPC 分类数量。很多专利价值指标与专利投资、专利维持和专利诉讼有关[4-5], 这些是价值评估的基础。但是, 这些指标对新授权专利的价值没有更进一步的预测能力。因此, 数据挖掘逐渐被用来进行专利价值评估[6], 通过构建一个自动分类系统来快速响应专利价值评估的

通讯作者: 周成, ORCID: 0000-0002-6965-5908, E-mail: zhoucheng1017@163.com。
*本文系东华大学人文社会科学繁荣基金项目“互联网个性化定制用户需求多粒度模型研究”(项目编号: 108-10-0108076)的研究成果之一。

需要, 能够为投资者在新产业研发决策时提供帮助。

2 文献综述

2.1 专利价值指标及价值评估

专利价值受到众多因素的影响, 对于专利价值指标的选择已经受到学术界广泛关注。其中, Archontopouls[7]认为专利的法律状态能够体现专利不同阶段的价值; Dang 等[8]利用影响指数、技术生命周期、科技关联度等指标衡量专利价值; Chen 等[9]通过专利引用关系研究企业专利价值与市场价值的关系; Frietsch 等[10]认为专利价值可以通过出口来评估, 利用专利的引用关系评估专利价值是一种有效方式; Harhoff 等[11]认为专利被引次数、专利家族大小、专利诉讼情况能够有效代表专利的价值; 郑素丽等[12]总结出专利价值度影响因素, 包含专利生命周期、专利宽度、专利创新性、专利的功能、专利权人特征。可以看出, 对于专利价值的指标主要有两种评价标准: 技术性(专利引用指标、专利保护范围、专利家族大小等)和法定性(权利要求、专利诉讼等)。此外, 目前专利价值的指标越来越多, 关于如何从众多专利价值指标中筛选出一些重要的指标以直接体现专利价值高低的研究还较少。

对于专利价值的评估基于统计知识, 使用一系列技术或工具分析技术创新的模式和趋势, 其评估的对象主要包括:

(1) 专利计数分析, 对专利的技术生命周期以及专利数量的对比分析;

(2) 专利布局国家分析, 分析专利优先权国家以及其他申请国家或地区;

(3) 引用率分析, 对专利在其有效期内被其他专利引用的数量分析;

(4) 国际 IPC 分类, 对 IPC 的布局以及数量分析。

学术界对专利价值的评估方法已经有一定的成果: Harhoff 等[11]通过问卷调查的方式为专利拥有者提供专利的市场价值, 还发现专利价值和专利家族的大小存在相关性; Marco[13]尝试通过引用文本的边际价值来寻找专利被引用表现出来的更多价值; 苏健美[14]通过建立模糊综合评价评估体系, 运用灰色理论的相关方法消除因素间的相互影响, 建立专利权收益最小成本模型对专利的年限收益价值进行评估; 杨冠灿等[15]提出一种基于综合引用网络的专利价值评估方法, 利用矩阵转化方法对 4 种单一专利引用关系(直接引用、间接引用、耦合引用、共同引用)进行合并、筛选和重组, 提出适合专利价值评估的专利综合引用网络的构建方法。综上, 目前对于专利价值的评估大多集中于传统的 AHP 或借用资产评估中的实物期权等方法, 所采用的价值指标也多集中于引用、专利家族等少数指标, 对于专利价值的其他指标使用不多。另外, 由于专利的申请与授权量逐年增多, 专利数据也逐渐向大数据发展, 现有的一些价值评估方法(AHP、TOPSIS)在评估效率上已经明显不足, 寻找高效的评估方法成为专利价值评估的潜在方向。

2.2 机器学习在专利价值评估与分类中的应用

随着数据挖掘和机器学习技术的快速发展, 基于机器学习技术的专利价值评估方法逐渐得到广泛关注。赵蕴华等[16]提出将机器学习相关方法应用到专利价值评估中的可能性, 但对于具体如何构建评估框架没有给出说明; 邱一卉等[6]利用分类回归树算法尝试构建新的专利价值评估体系, 但进行专利价值分类的标签来自于第三方商业机构, 标签的可信度不能保证; 吕璐成等[17]利用决策树算法对可能影响专利价值的 12 个影响因素与专利是否被引的潜在关系进行分析; Chiu 等[18]使用支持向量机(Support Vector Machine, SVM)对专利文本信息进行分类, 以分析专利不同的价值表现。但吕璐成等和 Chiu 等都没有对专利价值指标进行选择, 也没有对分类模型进行一定的改进, 以提高分类的准确率。

可以看出, 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, 通过对专利的价值进行标注, 辅以相关价值表现指标, 使用相关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。因此, 如何对专利的价值进行标注以及如何科学合理地选择价值指标, 对于最终专利价值分类模型的效果有重要影响。目前一些商业专利检索系统从自身角度对专利进行价值标注, 但如何标注不得而知, 其可靠性不能保证, 也有相关研究使用专家判断或问卷调查的形式对专利价值进行标注。如 Wu 等[19]采用德尔菲法请相关专家对专利进行价值判断, 形成专利分类系统的参照标准; Harhoff 等[11]通过调查问卷为专利拥有者提供专利的市场价值, 但如果需要对大量专利

数据进行评估，这两种方式显然是不切实际的，不但耗费人力、时间，其可靠性也不能保证。

因此，本文利用可以公开获取的专利数据指标(如国家知识产权局专利数据库)，采用自组织映射(Self-Organizing Maps，SOM)聚类算法对专利进行价值评估，获得价值标签，利用随机森林对专利价值指标进行重要性排序，并在此基础上使用 SVM 对专利进行价值分类，进而验证本文所提评估和分类方法的有效性。

3　专利价值评估及分类整体框架

为对专利价值进行有效的评估与分类，本文提出混合自组织映射、随机森林以及支持向量机的评估分类模型，如图 1 所示。该模型整体上分为数据收集与整理、基于 SOM 的专利价值评估、基于随机森林的指标重要性排序、基于包裹式特征选择支持向量机的专利分类 4 个部分。

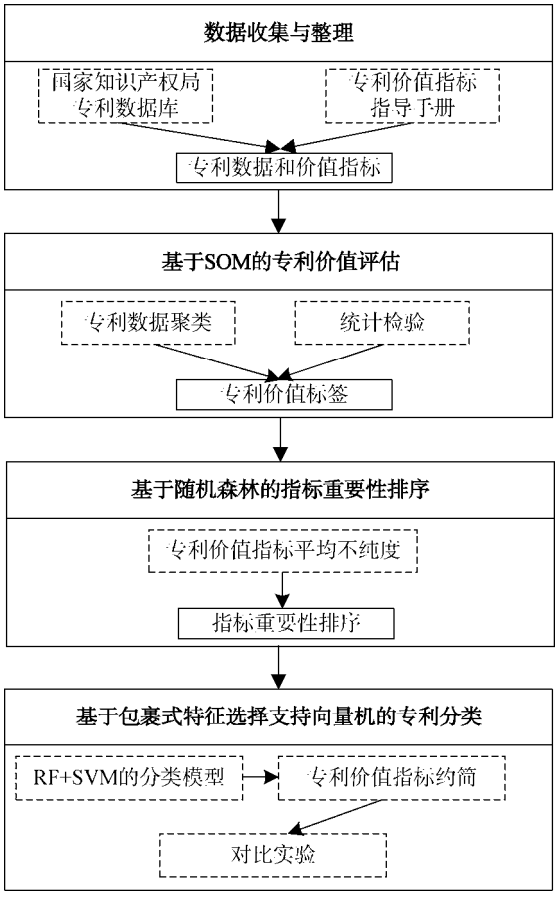

图 1　专利价值评估与分类模型的整体框架

(1) 数据收集与整理

从国家知识产权局专利数据库收集专利数据，并根据《专利价值指标指导手册》筛选专利价值指标，对数据进行规范化处理。

(2) 基于 SOM 的专利价值评估

目前，国家知识产权局数据库中并没有给出专利价值的类别标签，部分商业专利检索机构依据自身的经验判断给出专利价值的类别，但不同机构的知识结构或评估方法有所不同，这样的价值标签可信度有待证实。因此，本文采用 SOM 算法生成专利的价值标签，并将此标签作为分类模型构建的依据。

SOM 是一种两层神经网络，用于将多维数据映射为二维拓扑结构。映射过程中根据数据的相似性或按某种模式进行分组，比如欧式距离，所以 SOM 属于机器学习中的无监督学习范畴，允许数据自由组织，相较于 K-means 算法来说，SOM 倾向于实际的人脑结构，并在聚类过程中引入了竞争邻域的概念，通过某个节点及其邻居节点的竞争关系来动态调整节点的权重，在不断迭代过程中，形成最终的聚类结果，这种竞争理念相比单纯使用欧式距离作为聚类判定有一定的改进，是本文选取该聚类算法的主要原因。根据算法描述，本文的收敛准则依据输入变量进入 SOM 分析的迭代次数来确定。SOM 算法过程描述如下：

①设置 SOM 网络参数，如聚类个数和迭代次数。

②初始化每个神经元的权重 $w_i=[w_{i1},w_{i2},\cdots,w_{ij}]^{\mathrm{T}}$，其中 i 表示神经元个数，j 表示输入单元的个数。通过从输入数据集中随机选取样本来初始化权重。

③随机选取一个输入向量 $x=[x_1,x_2,\cdots,x_j]^{\mathrm{T}}$，判别函数定义为输入向量 x 和每个神经元 i 的权重向量 w_i 之间的欧式距离；对于竞争过程中优胜神经元的选取依据是该神经元到周围神经元欧式距离最短，如公式(1)所示。

$$\|x-w_c\|=\min_i\{d_i\} \tag{1}$$

其中，c 表示获胜神经元，d_i 表示欧式距离。

④更新获胜神经元 c 及其邻居的权重，调整方式如公式(2)所示。

$$w_i(t+1)=w_i(t)+h_{ci}(t)[x(t)-w_i(t)] \tag{2}$$

其中，i 是邻居神经元的索引，t 是一个表示训练时间的整数，$h_{ci}(t)$ 是获胜神经元和其邻居神经元 i 之间的时间距离函数，具体描述如公式(3)所示，它定义了输入变量在 SOM 上的影响区域，由邻域函数 $h(\|\cdot\|,t)$ 和学习率

$\alpha(t)$ 组成。

$$h_{ci}(t)=h(\|r_c-r_i\|,t)\alpha(t) \tag{3}$$

其中，r 表示神经元位置，$\alpha(t)$ 是一个单调减函数。引入高斯邻域函数对公式(3)进行改进，如公式(4)所示。

$$h_{ci}(t)=\exp\left(\frac{\|r_c-r_i\|}{2\sigma^2(t)}\right)\alpha(t) \tag{4}$$

⑤重复步骤③和步骤④直到满足收敛准则。

(3) 基于随机森林的指标重要性排序

随机森林(Random Forest, RF)是 Breiman[20]提出的一种组合分类算法，是对单分类回归树(Classification And Regression Tree, CART)的组合改进，该算法可以处理高维度数据集，对离散型和连续型数据的处理能力较好。由于随机森林的决策树是从所有特征中随机选择的特征子集，并非使用全部特征，因此RF被广泛地用于特征选择当中[21-22]，通过降低数据维度提高分类模型的性能。其中，指标重要性的依据是比较每个指标的平均不纯度，如公式(5)所示。

$$Gini(A)=1-\sum_{i=1}^{c}p_i^2 \tag{5}$$

其中，A 代表分类树的一个节点，p_i 表示属于第 i 类的概率，c 表示样本的类别数。

在这里对本文先进行专利价值评估，再进行指标重要性排序的原因进行如下几点说明：

①专利价值评估使用欧式距离和竞争邻域，而本文中指标重要性是通过平均不纯度来体现的，两者参照依据不同，指标重要性对于专利价值的评估结果不产生任何影响；

②本文将专利数据库中能够用于表征专利价值的所有指标都纳入到价值评估过程中，试图从一个全面的角度来考察专利价值，也为如何从专利数据库中选择专利价值指标提供参考；

③指标重要性与后续的特征选择和分类模型的构建密切相关，这样的顺序是为了保持专利价值指标筛选与分类模型的紧密性。

(4) 基于包裹式特征选择的支持向量机专利分类

目前，支持向量机(SVM)在分类问题中被广泛使用，其基本原理是：寻找一个最优的超平面，能够最大可能地将所有样本进行区分，通过线性或非线性转换，将输入变量映射到高维特征空间，一个有效的超平面能够高准确性地区分测试数据类型。该算法最初主要用于二分类，而专利价值分类问题往往不止两种，属于多分类问题。Cortes 等[23]对原始的二分类SVM 进行改进，使其可以支持多分类问题，本文借鉴他们的方法，构建多分类 SVM。经过随机森林过程，得到专利价值指标的重要性排序，也就是说，部分重要程度低的指标可能会对支持向量机的分类性能有负面影响，邱一卉等[6]也认为专利数据属于多维数据，过多的指标会影响专利分类的准确率，所以在专利价值评估时有必要进行特征约简，从而提高评估模型的分类性能。

机器学习中的特征选择方法主要有三种：过滤式、嵌入式以及包裹式，其中过滤式、嵌入式对于构建数据集的方式依赖性较大，不同的构建方式对最终的结果影响较大，包裹式则没有特定的构建方法，而是以最终模型的性能作为评价标准。此外，由于专利数据具有非线性的特性[24]，传统的主成分分析方法不适用于专利价值指标的特征选择，周成等[25]提出一种递归式特征搜索方法，通过排列组合的形式寻找最优指标集合，但递归搜索的时间复杂度较高，对于高维数据的搜索效率不高，也不易发掘那些重要指标。前向序列选择是机器学习中一种常见的特征约简方式，但传统的前向序列选择并没有预先考虑指标的重要性，不易发掘出重要指标。另外，如果采用递归组合的形式进行特征选择，就没有必要进行基于随机森林的指标重要性排序，本文采取随机森林重要性+前向序列选择进行特征筛选，目的在于避免递归搜索在处理多维度数据时效率不高的问题。综上，本文摒弃了传统排列组合的特征选择方法，在 SOM 和 RF 的基础上构建基于包裹式的前向序列特征选择 SVM 的分类模型，对专利价值指标进行一定的约简，筛选出重要的价值指标，从而提高特征选择效率和分类模型的准确率。此外，为验证该模型的分类性能，引入其他分类算法进行对比。

4 实证研究

4.1 数据收集

本文从国家知识产权局专利数据库(http://www.pss-system.gov.cn/)随机选取 2016 年获得授权的发明专利 10 000 个，并结合《专利价值指标指导手册》，得到 14 个评估指标，如表 1 所示。

表 1　专利价值指标

名称	描述
申请人	专利申请人的数量
扩展专利家族	与其他专利具有同一申请优先权下的专利数量
专利布局国家	专利同时申请保护的国家数量
专利家族	同一优先权和同一技术声明下的专利数量
专利引用次数	专利文献对该专利的引用情况
非专利引用次数	非专利文献对该专利的引用情况
代理人	专利申请时的代理人数量
专利施引	专利对已有专利引用数量
权利要求	专利的保护范围
发明人	专利发明人数量
授权时长	从申请到获得授权的时间
IPC 分类	专利的 IPC 分类数, 表示技术范围
优先权国家	优先权国家数量
优先权期限	从优先权日期到获得授权的时间

在中国, 由于专利从申请到公告至少需要 18 个月, 2017 年的专利数据可能不全, 因此, 选取 2016 年的专利; 由于所有专利均来自于 2016 年, 所以专利被引的概率是一致的, 消除了时间因素的影响; 已有部分研究指出, 发明专利的价值要高于实用新型和外观设计[26], 所以本文将发明专利作为实验对象。

4.2　基于 SOM 的专利价值评估

SOM 的目的是通过计算样本之间的欧式距离以动态调整邻近样本的权重, 使得相似的样本聚在一起。专利价值类别的确定如公式(6)所示。

$$Value(group_g)=\frac{1}{m\times n}\sum_{i=1,i\in group_g}^{m}\sum_{j=1}^{n}q_{ij} \qquad (6)$$

其中, q_{ij} 表示第 g 个专利类中第 i 个专利的第 j 个价值指标的价值得分, m 表示一个专利类中的专利个数, n 表示价值指标个数。

Wu 等[27]认为专利价值的类别划分数量不宜过多或过少: 过少, 专利之间的价值区分不明显; 过多, 区分过于细致, 不易于决策者的整体把握, 他们建议数量控制在 3-7 个。此外, 结合目前市场一些商业用途的专利检索系统(如 Innojoy、万象云等)给出的专利价值评估结果, 也都集中在 5-7 个。本文主要提出一种专利价值评估方法, 对于聚类数量应如何确定不做深究, 使用者可以依据自身情况确定, 另外, 考虑到奇数数量的类别较偶数数量类别在不同价值类别上有更好的区分度, 本文选取 3-7 的中间值将聚类个数设定为 5 个。经过 SOM 过程, 得到不同类别的平均价值得分, 如表 2 所示。本文认为价值得分越高的专利类别, 其内部专利的价值越高, 据此, 将专利的价值标签依次设定为 1, 2, 3, 4, 5, 标签数值越大的专利, 其价值也相应越高。此外, 利用单因素方差分析(Analysis of Variance, ANOVA)检验所提专利价值指标在上述价值分类情况下是否对专利价值有显著影响, 其中, 价值标签作为因变量, 价值指标作为自变量。分析结果如表 3 所示, 所有的价值指标在显著水平 5%下都具有统计学上的意义, 说明本文选取的价值指标有效。

表 2　专利价值类别得分

价值类别	价值得分
G1	0.0749
G2	0.4650
G3	0.6872
G4	1.0636
G5	1.1421

表 3　ANOVA 估计结果

指标	均方	F	指标	均方	F
PVI1	3313.964	87.619*	PVI8	138793.530	7976.321*
PVI2	841364.292	21387.040*	PVI9	100924.432	27890.653*
PVI3	106131.180	6606.425*	PVI10	3346.934	89.239*
PVI4	34256.478	2500.79*	PVI11	31236.481	2502.49*
PVI5	18978.022	39.879*	PVI12	18071.022	34.979*
PVI6	4144.875	66.068*	PVI13	4814.875	63.168*
PVI7	2978.577	76.432*	PVI14	3178.507	79.402*

(注: *表示 p<0.05。)

另外, 为验证专利价值类别的有效性, 结合 Harhoff 等[11]、孙玉涛等[28]、乔永忠等[29]的研究成果, 统计不同价值类别在 IPC 分类、权利要求、专利家族、专利引用次数 4 个重要价值指标的分布情况(取平均值), 结果如表 4 所示。从结果来看, 高价值的专利在这 4 个指标的数值更高, 意味着通过 SOM 方法得到的

专利价值类别是有效的, 相应的价值标签能够作为后续分类模型的参照点。

表 4　不同专利价值类别的指标分布

价值类别	IPC 分类	权利要求	专利家族	专利引用次数
G1	1.3543	5.3841	1.0231	0.0332
G2	1.3751	7.4034	1.4336	0.9764
G3	1.4072	7.9709	1.6503	1.2034
G4	1.4636	8.6721	2.0321	1.5431
G5	1.5421	9.8097	2.6534	2.0235

4.3　基于随机森林的指标重要性排序

采用随机森林对 14 个专利指标进行重要性排序, 并采用10折交叉验证进行训练, 每组实验重复100次, 最后计算 100 次实验结果的平均值。基于随机森林的指标重要性排序采用 Matlab 中的 RF 工具箱实现, 模型中二叉树数量设定为 100。实验结果如表 5 所示。

表 5　专利价值指标重要性排序

排名	指标	排名	指标
1	权利要求	8	授权时长
2	IPC 分类	9	优先权国家
3	专利布局国家	10	专利施引
4	专利家族	11	发明人
5	专利引用次数	12	优先权期限
6	非专利引用次数	13	申请人
7	扩展专利家族	14	代理人

从表 5 可以看出, 权利要求、IPC 分类、专利布局国家、专利家族、专利引用次数对于专利价值的影响程度较高, 这也印证了邱一卉等[6]、Chen 等[9]、乔永忠等[29]相关学者的研究成果, 而申请人、发明人这类指标对专利价值的重要程度不明显, 意味着对专利价值的判断重点可以从排序靠前的指标入手。

4.4　基于包裹式特征选择的支持向量机专利分类

(1) 指标约简

由于本文专利价值指标较多, 可能部分指标对于专利的价值分类是多余的。因此, 在指标重要性排序的基础上, 尝试对 14 个指标进行适当删减, 以提高支持向量机的分类准确率。使用 Matlab 中的 LibSVM 工具包, 构建多分类支持向量机模型, 采用4.2节中的专利价值类别为分类标签, 按照随机森林计算出的指标重要性排序表逐个加入到 SVM 分类模型中, 从而得到不同指标个数下的分类准确率, 如表 6 所示。可知, 在全指标下的分类准确率为 76.28%, 依次加入指标时, 准确率呈现先升再降的趋势, 当加入指标为 10 个时支持向量机的分类准确率达到最大 86.89%, 优于所有指标, 说明初始的专利价值指标存在冗余, 会影响最终分类模型的性能, 因此本文将表 5 中的前 10 个指标作为最终用于专利价值评估及分类的指标。

表 6　支持向量机模型分类性能

指标个数	分类准确率(%)	指标个数	分类准确率(%)
1	61.54	8	83.11
2	77.76	9	86.81
3	81.07	10	86.89
4	81.72	11	85.09
5	81.31	12	85.06
6	82.45	13	84.93
7	82.46	14	76.28

(2) 对比实验

为验证本文专利价值标签和约简后价值指标的有效性, 选择线性判别分析(Linear Discriminant Analysis, LDA)、决策树(Decision Tree, DT)和神经网络(Neural Network, NN)三种分类模型进行对比验证, 各模型的参数设置均使用默认参数, 结果如表 7 所示。可以看出, 利用自组织映射得到专利价值类别和价值指标在不同分类模型上均有良好的效果(分类准确率高于 60%), 同时本文采用的 SVM 分类模型分类准确率要优于其他三种模型, 说明本文构建的支持向量机分类模型有良好的泛化能力。

表 7　不同分类模型性能对比

分类器模型	SVM	LDA	DT	NN
准确率(%)	86.89	79.10	77.26	64.34

5　结　语

本文提出一种基于自组织映射支持向量机的专利价值评估与分类模型, 从国家知识产权局专利数据库随机选取发明专利数据进行实证研究。实验结果表明通过 SOM 得到的专利价值标签能够有效反映其价值; 利用 RF 和包裹式特征选择方法对初始专利价值指标约简后, 能够提高 SVM 的分类准确率和效率。通过该模型可以对专利的价值进行科学合理的评估, 为企业研发工作提供支持。另外, 本文使用的数据量符合大数据分析的基本要求, 避免了从少量数据中对专利进

行价值评估所带来的决策错误。

未来将对每个专利价值类别组内的专利进行更进一步的价值度考量，因为即使同属于一个价值类别，每个专利的价值还是存在一定的差异，通过更为精确的组内价值度分析，可以提高专利价值分类系统的精确度。对于专利价值指标的约简，只是做了简单筛选，可能还存在进一步约简的可能性；暂时只讨论了 5 个专利价值类别的情况，未来可以讨论更多类别的情况。

参考文献:

[1] 阮敏. 企业所有权性质、环境规制与发明专利的研发效率[J]. 软科学, 2016, 30(2): 55-59. (Ruan Min. Corporate Ownership Nature, Environmental Regulation and R&D Efficiencies of the Invention Patent[J]. Soft Science, 2016, 30(2): 55-59.)

[2] 张耀天，杜慰纯，贾明顺，等. 基于自适应层次分析法的企业专利质量评价研究[J]. 图书情报工作，2016，60(7): 110-115. (Zhang Yaotian, Du Weichun, Jia Mingshun, et al. Research on the Evaluation of Enterprise Patents Quality Based on the Adaptive Analytic Hierarchy Process[J]. Library and Information Service, 2016, 60(7): 110-115.)

[3] Dereli T, Durmuşoğlu A. Classifying Technology Patents to Identify Trends: Applying a Fuzzy-Based Clustering Approach in the Turkish Textile Industry[J]. Technology in Society, 2009, 31(3): 263-272.

[4] Narin F. Patent Bibliometrics[J]. Scientometrics, 1994, 30(1): 147-155.

[5] Ashtor J H. Redefining “Valuable Patents”: Analysis of the Enforcement Value of U.S. Patents[J]. Social Science Electronic Publishing, 2015. https://ssrn.com/abstract=2628198.

[6] 邱一卉，张驰雨，陈水宣. 基于分类回归树算法的专利价值评估指标体系研究[J]. 厦门大学学报：自然科学版，2017，56(2): 244-251. (Qiu Yihui, Zhang Chiyu, Chen Shuixuan. Research of Patent-Value Assessment Indictor System Based on Classification and Regression Tree Algorithm[J]. Journal of Xiamen University: Natural Science, 2017, 56(2): 244-251.)

[7] Archontopoulos E. Prior Art Search Tools on the Internet and Legal Status of the Results: A European Patent Office Perspective[J]. World Patent Information, 2004, 26(2): 113-121.

[8] Dang J, Motohashi K. Patent Statistics: A Good Indicator for Innovation in China? Patent Subsidy Program Impacts on Patent Quality[J]. China Economic Review, 2015, 35: 137-155.

[9] Chen Y S, Chang K C. The Relationship Between a Firm’s Patent Quality and Its Market Value—The Case of US Pharmaceutical Industry[J]. Technological Forecasting & Social Change, 2010, 77(1): 20-33.

[10] Frietsch R, Neuhäusler P, Jung T, et al. Patent Indicators for Macroeconomic Growth—The Value of Patents Estimated by Export Volume[J]. Technovation, 2014, 34(9): 546-558.

[11] Harhoff D, Scherer F M, Vopel K. Citations, Family Size, Opposition and the Value of Patent Rights[J]. Research Policy, 2003, 32(8): 1343-1363.

[12] 郑素丽，宋明顺. 专利价值由何决定?——基于文献综述的整合性框架[J]. 科学学研究，2012，30(9): 1316-1323. (Zheng Suli, Song Mingshun. Review on the Determinants of Patent Value: An Integrate Framework[J]. Studies in Science of Science, 2012, 30(9): 1316-1323.)

[13] Marco A C. The Dynamics of Patent Citations[J]. Economics Letters, 2006, 94(2): 290-296.

[14] 苏健美. 基于收益法的专利权价值评估研究[D]. 昆明：云南大学，2014. (Su Jianmei. Research on Patent Value Evaluation Based on Income Law[D]. Kunming: Yunnan University, 2014.)

[15] 杨冠灿，刘彤，李纲，等. 基于综合引用网络的专利价值评价研究[J]. 情报学报，2013，32(12):1265-1277.(Yang Guancan, Liu Tong, Li Gang, et al. Research on Patent Value Evaluation Based on Comprehensive Citation Network[J]. Journal of the China Society for Scientific and Technical Information, 2013, 32(12): 1265-1277.)

[16] 赵蕴华，张静，李岩，等. 基于机器学习的专利价值评估方法研究[J]. 情报科学, 2013, 31(12): 15-18. (Zhao Yunhua, Zhang Jing, Li Yan, et al. Study on Evaluation for Patent Value Based on Machine Learning[J]. Information Science, 2013, 31(12): 15-18.)

[17] 吕璐成，刘娅，杨冠灿. 基于决策树方法的专利被引影响因素研究[J]. 情报理论与实践，2015，38(2): 28-32. (Lv Lucheng, Liu Ya, Yang Guancan. Research on the Influencing Factors of Patent Citation Based on Decision Tree Method[J]. Information Studies: Theory & Application, 2015, 38(2): 28-32.)

[18] Chiu C Y, Huang P T. Application of the Honeybee Mating Optimization Algorithm to Patent Document Classification in Combination with the Support Vector Machine[J]. International Journal of Automation and Smart Technology, 2013, 3(3): 179-191.

[19] Wu C H, Ken Y, Huang T. Patent Classification System Using a New Hybrid Genetic Algorithm Support Vector Machine[J]. Applied Soft Computing, 2010, 10(4): 1164-1177.

[20] Breiman L. Random Forests[J]. Machine Learning, 2001, 45(1): 5-32.

[21] 姚登举，杨静，詹晓娟. 基于随机森林的特征选择算法[J]. 吉林大学学报：工学版, 2014, 44(1): 137-141. (Yao Dengju, Yang Jing, Zhan Xiaojuan. Feature Selection Algorithm Based on Random Forest[J]. Journal of Jilin University: Engineering and Technology Edition, 2014, 44(1): 137-141.)

[22] Strobl C, Boulesteix A L, Kneib T, et al. Conditional Variable Importance for Random Forests[J]. BMC Bioinformatics, 2008, 9: 307.
[23] Cortes C, Vapnik V. Support Vector Network[J]. Machine Learning, 1995, 20(3): 273-297.
[24] 裴云龙, 蔡虹, 王晓南. 中外科学文献对中国高新产业技术创新质量的影响——基于专利的科学引文的分析[J]. 情报学报, 2013, 32(12): 1333-1344. (Pei Yunlong, Cai Hong, Wang Xiaonan. How Does Domestic and Foreign Scientific Literature Affect Technology Innovation Quality of High-tech Industries in China: An Analysis Based on Patent Citations Scientific Publications[J]. Journalof the China Society for Scientific and Technical Information, 2013, 32(12): 1333-1344.)
[25] 周成, 魏红芹. 基于随机森林属性约简的众包竞赛参与者识别体系研究[J]. 数据分析与知识发现, 2018, 2(7): 46-54. (Zhou Cheng, Wei Hongqin. Identifying Crowd Participants with Modified Random Forests Algorithm[J]. Data Analysis and Knowledge Discovery, 2018, 2(7): 46-54)
[26] 徐庆富, 康旭东, 杨中楷, 等. 基于专利权转让的我国省际技术转移特征研究[J]. 情报杂志, 2017, 36(7): 66-72. (Xu Qingfu, Kang Xudong, Yang Zhongkai, et al. Research on the Characteristics of Inter-Provincial Technology Transfer in China Based on Patent Right Transfer[J]. Journal of Intelligence, 2017, 36(7): 66-72.)
[27] Wu J L, Chang P C, Tsao C C, et al. A Patent Quality Analysis and Classification System Using Self-Organizing Maps with Support Vector Machine[J]. Applied Soft Computing, 2016, 41: 305-316.
[28] 孙玉涛, 栾倩. 专利质量测度"三阶段—两维度"模型及实证研究——以 C9 联盟高校为例[J]. 科学学与科学技术管理, 2016, 37(6): 23-32. (Sun Yutao, Luan Qian. A 'Three Stages-Two Dimensions' Model of Patent Quality Measuring and Its Empirical Study: A Case Study of C9 League[J]. Science of Science and Management of S.& T., 2016, 37(6): 23-32.)
[29] 乔永忠, 谭婉琳. 专利权利要求数与维持时间关系实证研究——以中日授权专利数据为例[J]. 科学学与科学技术管理, 2017, 38(2): 77-86. (Qiao Yongzhong, Tan Wanlin. Empirical Studies of the Relationship of the Claims Number and the Maintenance Time of Patents: Based on the Data of Patents Granted by China and Japan[J]. Science of Science and Management of S.& T., 2017, 38(2): 77-86.)

作者贡献声明:

周成: 提出研究思路, 采集数据, 设计并实验算法, 起草论文;
魏红芹: 分析实验结果, 修订论文。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储, E-mail: zhoucheng1017@163.com。
[1] 周成, 魏红芹. 专利数据. xlsx. 实验原始数据及结果数据.

收稿日期: 2018-06-25
收修改稿日期: 2018-09-24

Evaluating and Classifying Patent Values Based on Self-Organizing Maps and Support Vector Machine

Zhou Cheng Wei Hongqin
(Glorious Sun School of Business and Management, Donghua University, Shanghai 200051, China)

Abstract: [Objective] This paper proposes a new method for evaluating and classifying patent values. **[Methods]** With the help of value indicators, we designed a patent value analysis and classification system based on self-organizing maps (SOM) and support vector machine (SVM) techniques. We used the SOM to determine value categories, and then applied the random forest (RF) algorithm to rank value indictors based on their significance. Finally, we improved classification performance with the wrapped feature reduction method. **[Results]** The value tags determined by SOM effectively represented the patent values. Meanwhile, the value indictors were reduced from 14 to 10, and the classification accuracy was increased from 76.28% to 86.89%. **[Limitations]** Further refinement of patent values in each category is needed, which might reduce the patent value indicators. **[Conclusions]** The proposed SOM-RF-SVM method could support research and development activities as well as reduce the dependence on human factors.

Keywords: Evaluation of Patent Values Data Clustering Classification of Patent Values Feature Selection Self-Organizing Maps Support Vector Machine