Use of the Industrial Property System in Colombia (2018): A Supervised Learning Application

Jenny-Paola Lis-Gutiérrez[1,2,5](✉), Melissa Lis-Gutiérrez[2,3], Adriana Patricia Gallego-Torres[4], Vladimir Alfonso Ballesteros Ballesteros[5], and Manuel Francisco Romero Ospina[5]

[1] Universidad de la Costa, Barranquilla, Colombia
jlis@cuc.edu.co
[2] Universidad Nacional de Colombia, Bogota, Colombia
{jplisg,mlisg}@unal.edu.co
[3] Corporación Universitaria del Meta, Villavicencio, Colombia
melissa.lis@unimeta.edu.co
[4] Universidad Distrital Francisco José de Caldas, Bogota, Colombia
adpgallegot@udistrital.edu.co
[5] Fundación Universitaria Los Libertadores, Bogota, Colombia
{jplisg,vladimir.ballesteros,mfromeroo}@libertadores.edu.co

Abstract. The purpose of this paper is to establish ways to predict the spatial distribution of the use of the intellectual property system from information on industrial property applications and grants (distinctive signs and new creations) and copyright registrations in 2018. This will be done using supervised learning algorithms applied to information on industrial property applications and grants (trademarks and new creations) and copyright registrations in 2018. Within the findings, 4 algorithms were identified with a level of explanation higher than 80%: (i) Linear Regression, with an elastic network regularization; (ii) Stochastic Gradient Descent, with Hinge loss function, Ringe regularization (L2) and a constant learning rate; (iii) Neural Networks, with 1,000 layers, with Adam's solution algorithm and 2,000 iterations; (iv) Random Forest, with 10 trees.

Keywords: Spatial distribution · Distinctive signs · New creations · Supervised learning · Machine learning

1 Introduction

According to the World Intellectual Property Organization (WIPO, 2016), intellectual property (IP) refers to all creations of the intellect, which include (i) inventions, (ii) literary, artistic and scientific works, (iii) symbols, names and images used in commerce. Under the traditional protection scheme, IP is divided into three branches: industrial property, copyright and related rights, and plant varieties [1].

Y. Tan et al. (Eds.): ICSI 2020, LNCS 12145, pp. 506–514, 2020.
https://doi.org/10.1007/978-3-030-53956-6_46

Specifically, Colombia is characterized by low levels of use of copyright and industrial property protection systems [2]. According to the World Intellectual Property Organization [3, 4], in 2017 Colombia ranked 48th out of 129 countries in terms of applications for patent registrations, 36th in terms of trademark applications and 69th in terms of industrial registrations.

The use of the industrial property system by residents in Colombia is significantly lower than the use given by foreigners. Therefore, the purpose of this paper is to establish how to predict the spatial distribution of the use of the intellectual property system from information on industrial property applications and grants (distinctive signs and new creations) and copyright registrations in 2018.

The data used correspond to records from the Superintendence of Industry and Commerce [5], on new creations and distinctive signs requested and granted at the departmental level in 2018, and information from the different types of copyright records in 2018 from the National Copyright Directorate [6].

In order to achieve this purpose, several supervised learning algorithms are applied [7], such as (Random Forest, KNN, Support Vector Machines, linear regression, neural networks, among others [8]).

2 Literature Review

2.1 Intellectual Property

Some of the most recent studies on intellectual property include the following. [9] found that reforms in IP systems in some countries led to a significant reduction in the cost of debt in IP-intensive industries, through lower rates of borrowing. [10] employed a dynamic panel, for 70 countries with data between 1965 and 2009, establishing that patent rights have no effect on productivity growth.

There is also the study of [11]. These authors found that the enforcement of intellectual property rights is positively related to increased exports to advanced economies [12], but has negative effects in developing countries, associated with reduced speed of technology transfers and incentives to invest in R&D. [13] established that rich countries and small or poor countries apply intellectual property rights for different reasons. The former to protect innovations, the latter seeking access to foreign markets. According to these authors, emerging countries have greater flexibility in the application of IP systems. For their part, [14] showed that most large developers employ a combination of informal mechanisms and formal intellectual property rights (copyright, patents and trademarks).

With respect to the use of trademarks, the study by [15], who were able to establish that trademarks allow for the measurement of the degree of innovation and the capacity of response in foreign markets, stands out. [16], based on 712 observations from a cross-sector sample of European companies, applied a panel data model. The results revealed a positive relationship between the companies' international performance and the company's attitude towards enriching its portfolio with externally developed trademarks. [17], through a survey of 48 Portuguese companies located in S&T parks and incubators, analyzed the direct and indirect effects of intellectual property rights protection mechanisms. Contrary to most studies, they found that formal protection of intellectual property rights is detrimental to the competitiveness of firms, but other non-formal mechanisms such as trade secrets do encourage it.

The paper by [18], uses data from 14,065 Chinese companies between 2007 and 2013. The authors use a discrete time risk model and analyze the effects of differences in internal and external innovation mechanisms, specifically the efficiency of innovation and the spillover effect of trade, on the probability of survival of firms.

As can be seen, the analysis of intellectual property is ongoing and the means of analysis are diverse, but generally involve the use of quantitative tools. The methods proposed in this paper are related to machine learning. A brief state of the art on its applications is presented below.

2.2 Machine Learning

In recent literature there are several machine learning applications that include medicine, energy, education, intellectual property, among others. [19] used machine learning algorithms to predict and diagnose heart disease in India. They compared the accuracy of four machine learning algorithms using 14 attributes obtained from intensive care unit data. A similar study, but applied to fatty liver disease, considering 577 patients was developed by [20].

From other fields, in [21], they apply different algorithms to public transport data from EUBra-BIGSEA (Europe-Brazil Collaboration of Big Data Scientific Research Through Cloud-Centric Applications). [22] applied decode-based learning for machine translation. [23] predict the energy produced in a wind farm from line regression, decision tree, K-neighbor, and cross-validation to reduce variance.

As for the applications of machine learning to intellectual property, in [24], they reviewed 57 papers on artificial intelligence, automatic and in-depth learning associated with intellectual property. In [25], the employed algorithms were Support Vector Machines, Neural Networks and Decision Trees.

3 Method

This section describes the data used, the design of the study, the procedure and the indicators and models used for the analysis.

3.1 Data

This paper uses as its primary source the records of the Superintendence of Industry and Commerce [5], on new creations and distinctive signs requested and granted at the departmental level in 2018 and information from the different types of copyright registrations in 2018 from [6]. Table 1 presents the data, sources and units used for the following sections. The free software used was Orange (Demsar et al. 2013).

3.2 Study Design

This study is quantitative and its scope is descriptive and predictive. The design is non-experimental and cross-sectional. The units of analysis are the departments of Colombia.

Table 1. Data sources and units used

Variable	Unit	Source
Projected population 2018	Number of inhabitants	[27]
Application for patents of invention	Ratio for every 10,000 inhabitants	[5]
Application for utility model patents	Ratio for every 10,000 inhabitants	[5]
Granting of patents of invention	Ratio for every 10,000 inhabitants	[5]
Granting of utility model patents	Ratio for every 10,000 inhabitants	[5]
Chemical Invention Patents Filed	Ratio for every 10,000 inhabitants	[5]
Electricity/Electronics Patents Filed	Ratio for every 10,000 inhabitants	[5]
Industrial design application	Ratio for every 10,000 inhabitants	[5]
Industrial designs awarded	Ratio for every 10,000 inhabitants	[5]
Designations of origin submitted	Ratio for every 10,000 inhabitants	[5]
Commercial School presented	Ratio for every 10,000 inhabitants	[5]
Commercial slogans presented	Ratio for every 10,000 inhabitants	[5]
Collective brand presented	Ratio for every 10,000 inhabitants	[5]
Trademarks presented	Ratio for every 10,000 inhabitants	[5]
Certification brand submitted	Ratio for every 10,000 inhabitants	[5]
Trade names presented	Ratio for every 10,000 inhabitants	[5]
Trade names granted	Ratio for every 10,000 inhabitants	[5]
Certification brand awarded	Ratio for every 10,000 inhabitants	[5]
Trademark granted	Ratio for every 10,000 inhabitants	[5]
Collective brand granted	Ratio for every 10,000 inhabitants	[5]
Commercial slogan granted	Ratio for every 10,000 inhabitants	[5]
Commercial teaching granted	Ratio for every 10,000 inhabitants	[5]
Phonogram registration	Ratio for every 10,000 inhabitants	[6]
Registration of artistic work	Ratio for every 10,000 inhabitants	[6]
Registration of unpublished literary work	Ratio for every 10,000 inhabitants	[6]
Registration of musical work	Ratio for every 10,000 inhabitants	[6]
Registration of contract and other acts	Ratio for every 10,000 inhabitants	[6]
Registration of audiovisual works	Ratio for every 10,000 inhabitants	[6]
Registration of published literary work	Ratio for every 10,000 inhabitants	[6]
Software Registration	Ratio for every 10,000 inhabitants	[6]

Note: filing refers to the application, and granting to the distinctive signs that actually obtained registration before [5] and [6].
Source: Prepared by the Office on the basis of [5, 6, 27]

3.3 Procedure

Information on the statistics of new creations and distinctive signs was searched on the page of [5], and on copyright registrations on [6]. The most recent information was identified, i.e., data available from 2018. The information was organized considering the departments as units of analysis.

In order to control by the number of inhabitants of each department, the ratio of respective applications or concessions per 10,000 inhabitants was calculated. Different methods (algorithms) of supervised learning were applied for information processing: AdaBoost, Random Forest, SVM (Support Vector Machines), Neural Network, Stochastic Gradient Descent, Linear Regression, KNN and decision tree learning algorithm [28, 29].

4 Results

In this section, the different algorithms used to predict the number of trademarks are analyzed. The novelty of the analysis lies in the use of other information related to the use of industrial property and copyright protection systems, and the handling of artificial intelligence for projection.

In Fig. 1, the results for each of the algorithms used are presented. The results show the root mean square error (RMSE), the root mean absolute error (MAE) and R2, i.e. the proportion of the variance in the dependent variable that is predictable from the independent variables.

The different types of sampling used for prediction are shown: leave one out, test on the training data, boostraping of 100 repetitions with a training sample of 50% and 50% prediction, cross validation of 10 and 20 folds.

Based on the above, the evaluation is carried out considering R2 and the lowest MAE. The algorithms that presented the best behavior were:

– Linear Regression, with an elastic network regulation
– Stochastic Gradient Descent, with Hinge loss function, Ringe regularization (L2) and a constant learning rate
– Neural networks, with 1,000 layers, with Adam's solution algorithm and 2,000 iterations.
– Random forests, with 10 trees

The AdaBoost algorithm has an over-adjustment to the data, so it should not be considered. The algorithms with less predictive capacity were: SVM, KNN and decision trees. The prediction results are presented in Annex 2.

Test & Score

Settings

Sampling type: Leave one out

Scores

Model	MSE	RMSE	MAE	R2
Linear Regression	0.504	0.710	0.527	0.942
SGD	1.226	1.107	0.713	0.858
Neural Network	1.524	1.234	0.780	0.824
Random Forest	2.557	1.599	0.964	0.704
AdaBoost	2.613	1.617	0.916	0.698
Tree	5.174	2.275	1.382	0.401
SVM	6.404	2.531	1.283	0.259
kNN	7.357	2.712	1.858	0.148

Test & Score

Settings

Sampling type: No sampling, test on training data

Scores

Model	MSE	RMSE	MAE	R2
AdaBoost	0.057	0.239	0.133	0.993
Neural Network	0.064	0.252	0.186	0.993
SGD	0.070	0.264	0.195	0.992
Linear Regression	0.106	0.325	0.259	0.988
Tree	0.470	0.686	0.294	0.946
Random Forest	0.668	0.818	0.424	0.923
SVM	4.170	2.042	0.814	0.517
kNN	4.551	2.133	1.398	0.473

Test & Score

Settings

Sampling type: Stratified Shuffle split, 100 random samples with 50% data

Scores

Model	MSE	RMSE	MAE	R2
Linear Regression	1.180	1.086	0.690	0.867
Neural Network	2.602	1.613	1.021	0.707
AdaBoost	4.051	2.013	1.079	0.545
Random Forest	4.427	2.104	1.145	0.502
SGD	4.702	2.168	1.066	0.471
SVM	7.470	2.733	1.487	0.160
kNN	9.951	3.155	2.170	-0.119
Tree	9.557	3.091	1.784	-0.074

Fig. 1. Algorithm results. Source: own elaboration using Orange [26].

Test & Score

Settings

Sampling type: Stratified 10-fold Cross validation

Scores

Model	MSE	RMSE	MAE	R2
Linear Regression	0.587	0.766	0.567	0.932
SGD	1.094	1.046	0.687	0.873
Neural Network	1.644	1.282	0.885	0.810
AdaBoost	2.931	1.712	0.964	0.661
Random Forest	3.376	1.837	1.009	0.609
SVM	6.314	2.513	1.312	0.269
kNN	7.308	2.703	1.848	0.154
Tree	8.987	2.998	1.822	-0.040

Test & Score

Settings

Sampling type: Stratified 20-fold Cross validation

Scores

Model	MSE	RMSE	MAE	R2
Linear Regression	0.500	0.707	0.529	0.942
SGD	1.269	1.127	0.697	0.853
Neural Network	1.533	1.238	0.790	0.823
AdaBoost	2.829	1.682	0.933	0.673
Random Forest	3.574	1.890	1.019	0.586
SVM	6.407	2.531	1.286	0.258
kNN	7.366	2.714	1.866	0.147
Tree	8.239	2.870	1.663	0.046

Fig. 1. (*continued*)

5 Discussion and Conclusions

In the case of supervised learning it was possible to identify 4 algorithms with a level of explanation higher than 80%, these are: (i) Linear Regression, with an elastic network regularization; (ii) Stochastic Gradient Descent, with Hinge loss function, Ringe regularization (L2) and a constant learning rate; (iii) Neural Networks, with 1,000 layers, with Adam's solution algorithm and 2,000 iterations; (iv) Random Forest, with 10 trees. The results found in this study are consistent with those of [25], and some algorithms are added.

The originality of this study lies in: (i) the transformation in ratios of the variables, since the other studies on intellectual property make the analysis in absolute values; (ii) the compendium of the information of the systems of protection of the intellectual property in Colombia; (iii) application of the artificial intelligence (machine learning) for the description and projection.

It is suggested that for future research a multivariate spatial analysis be conducted, for example using a geographically weighted regression or a panel data analysis, to determine the behavior of the innovation of the record and granting of distinctive signs in the light of other variables. Similarly, it would be appropriate to extend the study to a greater number of years and to the possibility of predicting other variables.

References

1. Lis-Gutiérrez, J.P., Zerda-Sarmiento, A., Balaguera, M., Gaitán-Angulo, M., Lis-Gutiérrez, M.: Uso del sistema de propiedad industrial para signos distintivos en Colombia: un análisis departamental (2000–2016). En: Campos, G., Castaño, M., Gaitán-Angulo, M. & Sánchez, V. (Comps). Diálogos sobre investigación: avances científicos Konrad Lorenz, pp 193–215. Bogotá: Konrad Lorenz Editores (2019)
2. Lis-Gutiérrez, J.P., Lis-Gutiérrez, M., Gaitán-Angulo, M., Balaguera, M.I., Viloria, A., Santander-Abril, J.E.: Use of the industrial property system for new creations in colombia: a departmental analysis (2000–2016). In: Tan, Y., Shi, Y., Tang, Q. (eds.) DMBD 2018. LNCS, vol. 10943, pp. 786–796. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-93803-5_74
3. WIPO. World intellectual property indicators. Ginebra: OMPI (2018)
4. WIPO. Datos y cifras de la OMPI sobre PI, edición de 2018. Ginebra: OMPI (2019)
5. Superintendencia de Industria y Comercio. Estadísticas PI [base de datos] (2019)
6. Dirección Nacional de Derechos de Autor (DNDA). Estadísticas en línea [Base de datos]. Bogotá: DNDA (2019)
7. Moros Ochoa, A., Lis-Gutiérrez, J.P., Castro Nieto, G.Y., Vargas, C.A., Rincón. J.C.: La percepción de calidad de servicio como determinante de la recomendación: una predicción mediante inteligencia artificial para los hoteles en Cartagena. En: G. Campos, M.A. Castaño, M. Gaitán-Angulo, V. Sánchez (comp). Diálogos sobre investigación. Bogotá: Editorial Konrad Lorenz (2020)
8. Lis-Gutiérrez, J.P., Aguilera-Hernández, D., Escobedo David, L.R.: Análisis de las demandas de los integrantes del Ejército colombiano en calidad de víctimas; una aplicación de machine learning. En: G. Barbosa Castillo, M. Correa, y A. Ciro Gómez (eds.), Análisis de las demandas de los integrantes del Ejército en calidad de víctimas: una aplicación de "machine learning", pp. 437–468. Universidad Externado de Colombia, Bogotá (2020)
9. Alimov, A.: Intellectual property rights reform and the cost of corporate debt. J. Int. Money Finance **91**, 195–211 (2019)
10. Sweet, C., Eterovic, D.: Do patent rights matter? 40 years of innovation, complexity and productivity. World Dev. **115**, 78–93 (2019)
11. Auriol, E., Biancini, S., Paillacar, R.: Universal intellectual property rights: too much of a good thing? Int. J. Ind. Organ. **65**, 51–81 (2019)
12. Campi, M., Dueñas, M.: Intellectual property rights, trade agreements, and international trade. Res. Policy **48**(3), 531–545 (2019)
13. Papageorgiadis, N., McDonald, F.: Defining and measuring the institutional context of national intellectual property systems in a post-trips world. J. Int. Manag. **25**(1), 3–18 (2019)
14. Miric, M., Boudreau, K.J., Jeppesen, L.B.: Protecting their digital assets: the use of formal & informal appropriability strategies by App developers. Res. Policy **48**(8), 103738 (2019)
15. Barroso, A., Giarratana, M.S., Pasquini, M.: Product portfolio performance in new foreign markets: the EU trademark dual system. Res. Policy **48**(1), 11–21 (2019)

16. Denicolai, S., Hagen, B., Zucchella, A., Dudinskaya, E.C.: When less family is more: trademark acquisition, family ownership, and internationalization. Int. Bus. Rev. **28**(2), 238–251 (2019)
17. Teixeira, A.A., Ferreira, C.: Intellectual property rights and the competitiveness of academic spin-offs. J. Innov. Knowl. **4**(3), 154–161 (2019)
18. Zhang, D., Zheng, W., Ning, L.: Does innovation facilitate firm survival? Evidence from chinese high-tech firms. Econ. Model. **75**, 458–468 (2018)
19. Kannan, R., Vasanthi, V.: Machine learning algorithms with roc curve for predicting and diagnosing the heart disease. Soft Computing and Medical Bioinformatics. SAST, pp. 63–72. Springer, Singapore (2019). https://doi.org/10.1007/978-981-13-0059-2_8
20. Wu, C.C., et al.: Prediction of fatty liver disease using machine learning algorithms. Comput. Methods Programs Biomed. **170**, 23–29 (2019)
21. Alic, A.S., et al.: BIGSEA: a big data analytics platform for public transportation information. Future Gen. Comput. Syst. **96**, 243–269 (2019)
22. Banik, D., Ekbal, A., Bhattacharyya, P.: Machine learning based optimized pruning approach for decoding in statistical machine translation. IEEE Access **7**, 1736–1751 (2019)
23. Aguilar, R., Torres, J., Martín, C.: Aprendizaje Automático en la Identificación de Sistemas. Un caso de estudio en la generación de un parque eólico. Revista iberoamericana de automática e informática industrial **16**(1), 114–127 (2018)
24. Aristodemou, L., Tietze, F.: The state-of-the-art on Intellectual Property Analytics (IPA): a literature review on artificial intelligence, machine learning and deep learning methods for analysing intellectual property (IP) data. World Patent Inf. **55**, 37–51 (2018)
25. Havermans, Q.A., Gabaly, S., Hidalgo, A.: Forecasting European trademark and design filings: An innovative approach including exogenous variables and IP offices' events. World Patent Inf. **48**, 96–108 (2017)
26. Demsar, J., et al.: Orange: data mining toolbox in Python. J. Mach. Learn. Res. **14**(Aug), 2349–2353 (2013)
27. Departamento Administrativo Nacional de Estadística (DANE). Proyecciones de Población Departamental [Base de datos]. Bogotá: Dane (2020)
28. Quitian, O.I.T., Lis-Gutiérrez, J.P., Viloria, A.: Supervised and unsupervised learning applied to crowdfunding. Adv. Intell. Syst. Comput. **1108**, 90–97 (2020)
29. Viloria, A., Lis-Gutiérrez, J.P., Gaitán-Angulo, M., Stanescu, C.L.V., Crissien, T.: Machine learning applied to the H index of colombian authors with publications in scopus. Smart Innov. Syst. Technol. **167**, 388–397 (2020)