Model Assisted Statistics and Applications 12 (2017) 239–244
DOI 10.3233/MAS-170398
IOS Press

Penalized regression models for patent keyword analysis

Jong-Min Kim[a], Jea-Bok Ryu[b], Seung-Joo Lee[b] and Sunghae Jun[b,*]
[a]*Statistics Discipline, Division of Sciences and Mathematics, University of Minnesota-Morris, Morris, MN, USA*
[b]*Department of Statistics, Cheongju University, Chungbuk, Korea*

Abstract. Technology analysis is important work in management of technology. Most companies make plans for research and development (R&D) policy, new product development, or technological innovation using the results of technology analysis. In this paper, we propose a methodology of technology analysis using penalized regression models. We analyze the patent keywords extracted from the patent documents using ridge regression, least absolute shrinkage and selection operator, elastic net, and random forest. In addition, to show how our research could be applied to real problem efficiently, we carry out a case study of Apple technology. Our study contributes to perform R&D planning in technology management.

Keywords: Technology analysis, patent data analysis, zero-inflated problem, zero-inflated poisson model, zero-inflated negative binomial model, apple patent

1. Introduction

Technology forecasting is discovering the future trend or aspect of a technology (Hastie et al., 2001). Technology forecasting and management play an important role to improve the technological competition in a company (Hastie et al., 2001). So many studied on technology analysis were performed in management of technology (MOT) (Zou & Hastie, 2005; WIPSON, 2016; USPTO, 2016). The researches were mainly focused on the research and development (R&D) planning, because the technological competition is very important in a company. The methodologies of technology analysis have evolved constantly by many researchers (KIPRIS, 2016; Feinerer et al., 2008; Feinerer & Hornik, 2016; R Development Core Team, 2016). In the most previous researches, patent documents were used as input data for technology analysis, because patent involves so many information on the developed technologies. Furthermore, the exclusive right of technology registered to patent system is protected for a certain period of time (Friedman et al., 2016). So many scientists and engineers apply their developed technology to patent systems in the world. In this paper, we propose an analytical methodology for patent data analysis. This aims to analyze technology for MOT areas such as R&D planning, technological innovation, new product development, technology forecasting, etc. We consider penalized regression models for technology analysis, and use patent keywords as input data to analytical models. The penalized regression models are based on the ridge regression, the least absolute shrinkage and selection operator (LASSO) regression, elastic net, and random forest in this research. The ridge and LASSO regressions are shrinkage models selecting the meaningful variables by minimizing the residuals related to the regression coefficients (Friedman et al., 2010). The elastic net also minimizes the residual equation of regression parameters and select the necessary variables automatically (Han et al., 2012). Using the models based on penalized regression, we perform a case study to show how the methodology could be used to practical problems. Next section shows the penalized regression model for patent keyword analysis. We carry out the case study of Apple technology analysis in Section 3. In the last section, we conclude our research and provide the contribution of this paper.

[*]Corresponding author: Sunghae Jun, Department of Statistics, Cheongju University, Chungbuk 28503, Korea. E-mail: shjun@cju.ac.kr.

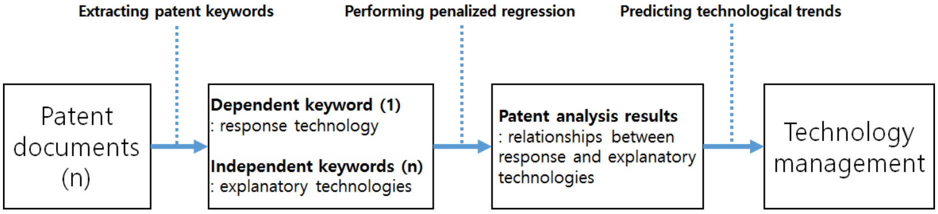
Extracting patent keywords
Performing penalized regression
Predicting technological trends
Patent documents (n)
Dependent keyword (1)
: response technology
Independent keywords (n)
: explanatory technologies
Patent analysis results
: relationships between response and explanatory technologies
Technology management

Fig. 1. Technology analysis using penalized regression models.

2. Penalized regression models for patent analysis

We suppose that the data set involves n rows (observations), p columns (explanatory variables, X) and one response variable (y). In the patent data, n is the number of collected patent documents. In addition, the explanatory and response variables show the technological keywords extracted from the patent document data. Using the penalized regression models, we predict the trend of response technology by the influence of explanatory technologies in this paper. Figure 1 shows the process of proposed methodology.

We retrieve the patent documents from the world patent databases such as the United States Patent and Trademark Office (USPTO), WIPS Corporation (WIPSON), or Korea Intellectual Property Rights Information Service (KIPRIS) (Zhao, 2013; Liaw & Wiener, 2002; Roper et al., 2010). Next we extract patent keywords from the patent documents using R data language and its text mining package (Jun, 2015, 2016; Kim & Jun, 2011). We classify the keywords into dependent and independent keywords by the aim of technology analysis. The dependent keyword (Y) represents the target (response) technology in patent analysis, and the independent keywords (X) are the predictive (explanatory) technologies for target technologies. Using the response and explanatory keywords, we perform the penalized regression models to find the relationships between technologies. The regression provides significant results to predict technological trends of target technology in given domain. Lastly we build the R&D policy for technology management. The detailed explanations of regression models for patent keyword analysis are shown in next sub sections.

2.1. *Ridge regression*

Ridge regression is one of shrinkage methods in statistics. The regression shrinks the parameters of model by adopting penalized approach. The ridge regression minimizes the following squared residuals (Battistella & Toni, 2011).

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2\,(\lambda\geqslant 0) \tag{1}$$

We can control the shrink degree using λ. The larger λ is, the larger the shrink degree is. The problem of high correlated variables in the linear model is settled by the ridge regression. In this paper, we consider the LASSO regression as more advanced model for the penalized regression model in patent data analysis.

2.2. *LASSO regression*

Though the LASSO regression is similar to the ridge regression, this has a clear difference for the ridge regression. The LASSO regression minimizes the following residual expression (Battistella & Toni, 2011).

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2 \quad \textit{subject to} \sum_{j=1}^{p}|\beta_j|\leqslant s \tag{2}$$

Where s is the constraint of regression parameters' sizes. Using the LASSO regression, we can select the significant variables, because this assigns very small values to the parameters of non-significant variables.

Table 1
RMSE of Y = device and X = other variables

Ridge	LASSO	Elastic net
1.785473	1.785563	1.785496

Table 2
RMS of Y = device and X = important variables by random fores

Ridge	LASSO	Elastic net
1.835592	1.835632	1.835603

2.3. *Elastic net*

Elastic net is a regression model penalized with the L1 and L2 norms. This contains the characteristics of both the ridge and LASSO regressions. This shrinks the groups of correlated variables and selects variables automatically (Han et al., 2012). In the elastic net, the response and explanatory variables are assumed centered and standardized respectively (Han et al., 2012). The elastic net minimizes the following equation (Han et al., 2012).

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2+\lambda_1\sum_{j=1}^{p}|\beta_j|+\lambda_2\sum_{j=1}^{p}\beta_j^2 \quad (3)$$

Where λ_1 and λ_1 are non-negative constants.

2.4. *Random forest*

Random forest is an ensemble based on forest consisting of trees (Choi & Jun, 2014). In the random forest, we carry out the process below (Choi et al., 2015; Guo et al., 2013).

Step 1: Sampling n bootstraps.
Step 2: Performing regression tree using n bootstrap samples.
Step 3: Predicting response technology by results of n trees.

In this paper, we analyze patent keyword data using the methods of penalized regression modeling. Next section shows a case study using Apple patent data.

3. Case study of apple patent data

To illustrate how our research could be applied to real problem, we carried out a case study using Apple patent documents. We collected the patent data from the patent databases in the world (Zhao, 2013). We extracted the keywords from the patent documents as follows; access, accessory, address, application, area, assembly, audio, bus, circuit, client, clock, code, color, communication, component, computer, connector, content, control, data, device, digital, disclosed, display, electronic, element, embodiment, frame, generated, graphics, host, housing, image, information, interface, invention, light, list, location, mechanism, media, memory, mobile, multiple, network, node, number, object, operation, pixel, plurality, portable, portion, position, power, present, processor, program, receiving, remote, request, response, screen, sensor, server, set, signal, source, state, storage, surface, system, text, time, touch, unit, user, vector, video, voltage, window, wireless. In this case study, we used the R data language and its package for penalized regression modeling (Kim & Jun, 2011; Hunt et al., 2007; Tibshirani, 1996). Firstly, we determine 'device' keyword as a response keyword, and all the rest were used for explanatory keywords. We compared three methods of the penalized regression models. Table 1 shows the RMSE (root mean square error) of comparative methods.

We knew that the difference between penalized methods was slight. We also made an experiment of comparison of three methods using random forest in Table 2.

Table 3
RMS of Y = device and X = common important variables by random forest

Ridge	LASSO	Elastic net
1.858242	1.858269	1.858249

Table 4
RMSE of Y = data and X = other variables

Ridge	LASSO	Elastic net
3.637373	3.637457	3.637395

Table 5
RMSE of Y = data and X = important variables by random fores

Ridge	LASSO	Elastic net
3.759405	3.759444	3.759415

Table 6
RMSE of Y = data and X = common important variables by random forest

Ridge	LASSO	Elastic net
3.847104	3.847129	3.847110

Table 7
RMSE of Y = system and X = other variables

Ridge	LASSO	Elastic net
1.706308	1.706391	1.706330

Table 8
RMS of Y = system and X = important variables by random forest

Ridge	LASSO	Elastic net
1.782321	1.782358	1.782331

Table 9
RMS of Y = system and X = common important variables by random fores

Ridge	LASSO	Elastic net
1.867590	1.867611	1.867596

Like the result of Table 1, there was little difference between the comparative methods. But we found the RMSE values were less than Table 2. We also used reduced model with common keywords from mean decrease accuracy and mean decrease Gini in Table 3.

The RMSE values among the comparative methods were about the same. The RMSE values of Table 3 were larger than Tables 1 and 2. So we conclude that the performance of Table 1 result is better than others. Next we considered the keyword ‘Data’ as a response technology in Tables 4–6.

The result of standard model (Table 4) is better than other results (Tables 5 and 6). This is similar to the conclusion of response keyword ‘Device’. In Tables 7–9, the results of response keyword ‘System’ are shown.

The results of response keyword ‘System’ are similar to the results of response keywords ‘Device’ and ‘Data’. Based on the results, we forecast Apple technology in Table 10.

We selected five keywords (Device, Data, System, User, and Media) as response keywords, and performed three models (ridge regression, LASSO, and Elastic Net). In addition, we evaluated the performance of technology forecasting by the RMSE. In the results of response keywords ‘Device’, ‘User’, and ‘Media’, the performance of ridge regression is the best. The LASSO provides the best result in the keywords ‘Data’ and ‘System’. Therefore, we conclude that the best model for technology forecasting is based on ridge and LASSO regressions by standard formula.

Table 10
Apple keyword forecasting for year 2007–2010 with year 1980–2006

		Ridge		LASSO		Elastic net	
Year	Device	Predicted	Residual	Predicted	Residual	Predicted	Residual
2007	832	749.38	82.62	747.97	84.03	748.68	83.32
2008	1279	1089.66	189.34	1089.11	189.89	1089.38	189.62
2009	1021	1121.14	−100.14	1116.22	−95.22	1118.69	−97.69
2010	1208	945.05	262.95	942.63	265.37	943.85	264.15
RMSE		174.5317		175.0790		174.7980	
Year	Data	Predicted	Residual	Predicted	Residual	Predicted	Residual
2007	602	233.44	368.56	242.45	359.55	237.97	364.03
2008	717	590.64	126.36	598.36	118.64	594.58	122.42
2009	578	349.51	228.49	355.06	222.94	352.26	225.74
2010	814	726.95	87.05	731.45	82.55	729.27	84.73
RMSE		229.9938		223.5337		226.7390	
Year	System	Predicted	Residual	Predicted	Residual	Predicted	Residual
2007	560	825.66	−265.66	824.60	−264.60	825.13	−265.13
2008	655	905.07	−250.07	904.38	−249.38	904.72	−249.72
2009	413	491.57	−78.57	492.23	−79.23	491.90	−78.90
2010	562	838.31	−276.31	837.63	−275.63	837.97	−275.97
RMSE		232.1814		231.5448		231.8643	
Year	User	Predicted	Residual	Predicted	Residual	Predicted	Residual
2007	605	780.46	−175.46	780.85	−175.85	780.74	−175.74
2008	655	832.14	−177.14	835.49	−180.49	834.06	−179.06
2009	429	306.36	122.64	308.84	120.16	307.71	121.29
2010	533	612.42	−79.42	615.16	−82.16	613.97	−80.97
RMSE		144.4929		145.5062		145.0995	
Year	Media	Predicted	Residual	Predicted	Residual	Predicted	Residual
2007	599	652.80	−53.80	655.25	−56.25	654.03	−55.03
2008	802	1028.58	−226.58	1032.02	−230.02	1030.30	−228.30
2009	483	1004.23	−521.23	1003.34	−520.34	1003.78	−520.78
2010	363	1029.02	−666.02	1030.65	−667.65	1029.84	−666.84
RMSE		438.6037		439.4871		439.0450	

4. Conclusions

We proposed a penalized regression models for patent keyword analysis. Each patent keyword is assigned to corresponding technology. So we forecast future technology by technology analysis based on patent keyword analysis. In this paper, we used ridge regression, LASSO, and Elastic net for penalized regression modeling. Also we compared three formula of model (standard, important variables by random forest, and common important variables by random forest) by RMSE measure. From the results of Apple case study, we found that ridge and LASSO regression models were better than elastic net model. In addition, the standard formula is better than others. In the future works, we will consider more advanced shrinkage model for patent data analysis.

References

Hastie, T., Tibshirani, R., & Friedman, J. (2001). The elements of statistical learning, data mining, inference, and prediction. New York: Springer.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301-320.

WIPSON. (2016). WIPS Corporation, http://www.wipson.com/, http://global.wipscorp.com/.

USPTO. (2016). The United States Patent and Trademark Office, http://www.uspto.gov/.

KIPRIS. (2016). Korea Intellectual Property Rights Information Service, www.kipris.or.kr/.

Feinerer, I., Hornik, K., & Meyer, D. (2008). Text mining infrastructure in R. *Journal of Statistical Software*, *25*(5), 1-54.

Feinerer, I., & Hornik, K. (2016). Package 'tm'. *Text Mining Package*, CRAN of R project.

R Development Core Team. (2016). R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org/.

Friedman, J., Hastie, T., Simon, N., & Tibshirani, R. (2016). Package 'glmnet'. *Lasso and Elastic-Net Regularized Generalized Linear Models*, CRAN of R project.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1.

Han, J., Kamber, M., & Pei, J. (2012). Data mining: Concepts and techniques, Third Edition, Waltham, MA: Morgan Kaufmann.

Zhao, Y. (2013). *R and Data Mining – Examples and Case Studies*, Academic Press.

Liaw, A., & Wiener, M. (2002). Classification and regression by random Forest. *R news*, *2*(3), 18-22.

Roper, A. T., Cunningham, S. W., Porter, A. L., Mason, T. W., Rossini, F. A., & Banks, J. (2011). *Forecasting and Management of Technology*, Hoboken, NJ, John Wiley & Sons.

Jun, S. (2015). Patent statistics for technology analysis. *International Journal of Software Engineering and Its Applications*, *9*(5), 155-164.

Jun, S. (2016). Patent big data analysis by R data language for technology management. *International Journal of Software Engineering and Its Applications*, *10*(1), 69-78.

Kim, J.-M., & Jun, S. (2011). Zero-inflated poisson and negative binomial regressions for technology analysis. *International Journal of Software Engineering and Its Applications*, *10*(12), 431-448.

Battistella, C., & Toni, A. F. D. (2011). A methodology of technological foresight: A proposal and field study. *Technological Forecasting & Social Change*, *78*, 1029-1048.

Choi, S., & Jun, S. (2014). Vacant technology forecasting using new Bayesian patent clustering. *Technology Analysis & Strategic Management*, *26*(3), 241-251.

Choi, J., Jang, D., Jun, S., & Park, S. (2015). A predictive model of technology transfer using patent analysis. *Sustainability*, *7*(12), 16175-16195.

Guo, B., Gao, J., & Chen, X. (2013). Technology strategy, technological context and technological catch-up in emerging economies: Industry-level findings from Chinese manufacturing. *Technology Analysis & Strategic Management*, *25*(2), 219-234.

Hunt, D., Nguyen, L., & Rodgers, M. (2007). *Patent Searching Tools & Techniques*, Hoboken, NJ: Wiley.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B (Methodological)*, 267-288.