Journal of Modern Mathematics and Statistics 4 (3): 101-104, 2010
ISSN: 1994-5388

Technology Forecasting Using Frequency Time Series Model: Bio-Technology Patent Analysis

[1]Sunghae Jun and [2]Daiho Uhm
[1]Department of Bioinformatics and Statistics, Cheongju University, Cheongju, Korea
[2]Department of Statistics, Oklahoma State University, USA

Abstract: Most of technology forecasting has been depended on the knowledge and experience of experts. It has caused many problems like inaccuracy as well as waste of time and cost. So, we need more objective and accuracy methods for technology forecasting. Recently, many researches of technology forecasting have been published. They were analyzed methods of patent data. In this study, the researchers propose a technology forecasting model using frequent time series analysis in bio-technology domain. The experimental data are patents of bio-technology. The researchers verify improved performance of frequent time series model in the experimental results.

Key words: Technology forecasting, frequency time series, patent analysis, poisson regression, linear regression, USA

INTRODUCTION

Hall *et al.* (2001) insisted that the patent data were better than others for technology forecasting. Also, Dernis *et al.* (2001) published the approach of time series analysis to forecast technology. In this study, we propose a method for technology forecasting using frequent time series model. Many experts have forecasted technology subjectively based on their experience and knowledge (Lee *et al.*, 2009; Wang *et al.*, 1998; Yoon and Park, 2007; Yoon and Lee, 2008) but we need more objective approach for efficient forecasting.

To settle this problem we consider frequent time series model based on patent data. To verify the method, we use the US patents about bio-technology (Vapnik, 1998).

Related researches: There were many researches in technology forecasting (Fattori *et al.*, 2003; Feinerer *et al.*, 2008; Lee *et al.*, 2009; Wang *et al.*, 1998; Yoon and Park, 2007; Yoon and Lee, 2008). They were depended on keywords from patent document using text mining (Fattori *et al.*, 2003; Feinerer *et al.*, 2008). Also, the methods about citation analysis were used in the study (Hall *et al.*, 2001). These have had many contributions in technology forecasting (Lee *et al.*, 2009). So, we can plan research and development (Rand D) processes by the results of technology forecasting. Generally R and D plan is very important in the government and company (Yoon and Park, 2007; Yoon and Lee, 2008). The plan is deeply involved with project cost. Also, we can avoid overlapping investment by efficient R and D plan. Many company and corporation have suffered from patent violation suits by other company or patent troll (Lee *et al.*, 2009; Yoon and Park, 2007). Before researching and developing a technology, the researchers have to analyze the technology results so far achieved. One of these data is patent document. Patent data are very objective for technology forecasting. More accurate results are needed for objective forecasting of technology. But many forecasting processes have been depended on subjective knowledge of the domain experts. In the research, the reserarchers propose an objective approach to technology forecasting. In the study, a frequency time series method is used to analyze patent data for efficient forecasting of technology.

Frequency time series mode for bio-technology forecasting: Traditional methods of time series are focused on continuous data (Brockwell and Davis, 2002). ARIMA (Auto-Regressive Integrated Moving-Average) is a popular time series model (Tsay, 2005). But this method has a problem for frequent time series data such as patent frequent by year. In the study, researchers forecast a trend of technology using MA (Moving Average). Also, we compare this approach with linear regression, Poisson regression and SVR (Support Vector Regression).

Linear regression and poisson regression: Linear regression finds a dependence of one variable on another (Myers, 1989). Its functional form is defined as the following:

$$Y_i = \alpha + \beta_{Xi} + \epsilon_1 \quad (1)$$

Corresponding Author: Sunghae Jun, Department of Bioinformatics and Statistics, Cheongju University, Cheongju, Korea

Where:

x_i and y_i = Independent and dependent variables, respectively

α and β = The intercept and slope of the regression line

ϵ = Error from Normal distribution with mean 0 and variance σ_ϵ

In the technology forecasting model, x and y are the time (year) and frequent of patent. The distribution of y is normal. This distribution has a continuous data. In the patent analysis, the frequent of patent is not continuous strictly speaking. So, we can another regression models fitted to discrete data type.

One of these models is poisson regression model. This consider y as discrete data (frequent). The distribution of y is Poisson in Poisson regression model. The researchers use these regression models to forecast the technology trend in the experiment.

Support vector regression: SVM (Support Vector Machine) is a non-linear model (Haykin, 1999; Vapnik, 1998). Firstly SVM were developed to apply to classification (Vapnik, 1998). Recently, it has solved the regression and clustering. SVM for regression is SVR and SVC (Support Vector Clustering) is the SVM for clustering (Burges, 1998; Haykin, 1999). Using an alternative loss function, SVR can show good performance in regression problem. Given data set, $\{(X_1, y_1), (X_2, y_2),\ldots, (X_n, y_n)\}$ and a linear model, $f(x) = w.X + b$, the researchers can optimal regression function by the following minimum of the functional:

$$\phi(w, \xi) = \frac{1}{2} \|w\|^2 + C \sum_i (\xi_i^- + \xi_i^+) \quad (2)$$

where, C is a constant, also ξ_i^- and ξ_i^+ are slack variables.

Moving average and central limit theorem: MA is a smoothing method in time series analysis (Brockwell and Davis, 2002; Tsay, 2005). Given a time series data set, (y_1 1, y_1 2,...., y_1 T), MA of point t computes F_m, the mean of the previous m points as the following formula:

$$F_m = \frac{y_{t-1} + y_{t-2} + \ldots + y_{t-m}}{m} \quad (3)$$

Traditional MA needs continuous type as a time series data. But the frequent of patent is discrete. To solve this problem, we consider the CLT (Central Limit Theorem) (Casella and Berger, 2002). The patent frequent data are sum according to years (from 1981-1995). Therefore, we use MA to forecast the bio-technology.

RESULTS AND DISCUSSION

The researchers use four IPC (International Patent Classification) codes about typical bio-technology from patents of the US (Vapnik, 1998). Table 1 shows these codes. To verify the proposed mode, the researchers use the patents in biotechnology which are C12M, C12N, C12P and C12Q.

They are popular IPC codes relevant to biotechnology. The researchers got the patent data from USPTO (United States Patent and Trademark Office). The training data has the patent from 1981-1995.

The data from 1996-2000 are used for testing the new model. IPC has a hierarchical structure including section, class, sub-class, maingroup and sub-group. So, the patent data for the experiments are following sub-classes Table 2 and Fig. 1-4 show the trend of frequent of bio-technology.

The researchers found some intervention in 1994, 1995 and 1996 and the researchers can think some events occurred in these periods. In the experiment, linear regression, poisson regression, SVR and MA

Table 1: IPC codes and their subclasses

IPC codes	Technology define
C12M	Apparatus for enzymology or microbiology
C12N	Micro-organisms or enzymes; Compositions thereof; Propagating, preserving or maintaining micro-organisms; Mutation or genetic engineering; Culture media
C12P	Fermentation or enzyme-using processes to synthesise a desired chemical compound or composition or to separate optical isomers from a racemic mixture
C12Q	Measuring or testing processes involving enzymes or micro-organisms; Compositions or test papers therefor; Processes of preparing such compositions; Condition-responsive control in microbiological or enzymological processes

Table 2: IPC codes and their sub-classes

IPC codes	No. of sub-classes	Sub-class items
C12M	59	C12M-001/00, C12M-001/02, …
C12N	282	C12N-000/00, C12N-000/500, …
C12P	209	C12P-001/00, C12P-001/02, …
C12Q	106	C12Q-000/00, C12Q-001/00, …

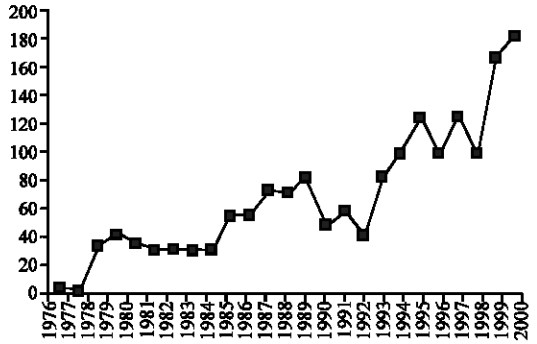

Fig. 1: Trend of patent frequent (C12M)

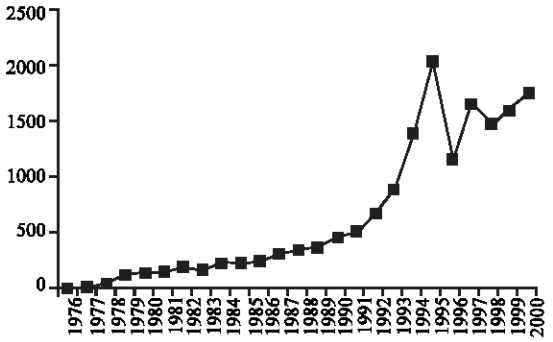

Fig. 2: Trend of patent frequent (C12N)

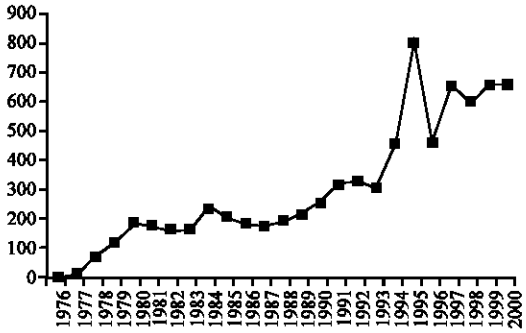

Fig. 3. Trend of patent frequent (C12P)

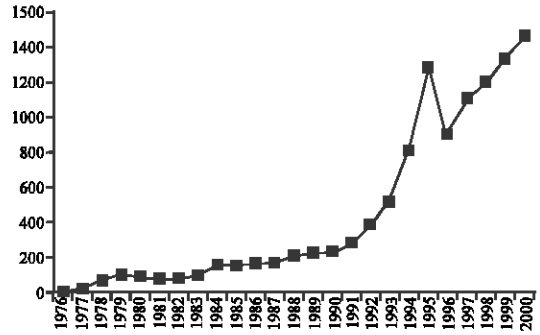

Fig. 4: Trend of patent frequent (C12Q)

are considered as comparative models. Firstly, we got a result for technology forecasting using linear regression. Table 3 shows regression parameters (b_0-intercept, b_1-slope of regression line), R^2 and MSE (mean squared error) (Han and Kamber, 2001; Myers, 1989). The reserachers can find the performance of poisson regression model for bio technology forecasting in Table 4.

Where AIC (Akaike Information Criterion) is a criterion for measuring the performance of poisson regression method (Johnson, 1998; Myers, 1989). The researchers also use SVR with RBF (radial basis function) for bio-technology forecasting model in Table 5. Where

Table 3: Result of linear regression

IPC code	b_0	b_1	R^2	MSE
C12M	5.36	4.31	0.7054	1701.18
C12N	-289.22	68.46	0.6523	82190.48
C12P	-17.74	23.56	0.6656	9708.77
C12Q	-164.03	39.96	0.6016	217366.43

Table 4: Result of Poisson regression

IPC code	b_0	b_1	AIC	MSE
C12M	-174.1322	0.0896	244.8	462.54
C12N	-385.0254	0.1967	702.9	2756068.04
C12P	-214.6075	0.1107	679.8	20067.91
C12Q	-374.5327	0.1912	606.6	512865.70

Table 5: Result of SVR (RBF)

IPC code	No.	C	Gamma	MSE
C12M	19	1	1	4103.13
C12N	7	1	1	402297.08
C12P	13	1	1	65668.39
C12Q	9	1	1	453908.75

Table 6: Result of MA

IPC code	m	MSE
C12M	2	2044.20
	3	3630.32
	4	5039.58
	5	5408.31
C12N	2	29472.60
	3	181772.47
	4	418194.06
	5	624506.17
C12P	2	7343.02
	3	35412.75
	4	56856.87
	5	76446.12
C12Q	2	118890.32
	3	307691.81
	4	480220.86
	5	609314.78

Table 7: MSEs of comparative models

Model	C12M	C12N	C12P	C12Q
Linear regression	1701.18	82190.48	9708.77	217366.43
Poisson regression	462.54	2756068.04	20067.91	512865.70
SVR	4103.13	402297.08	65668.39	453908.75
MA (m = 2)	2044.20	29472.60	7343.02	118890.32

no. of S.V. is the number of support vectors. C is the regularization parameter (constant) and Gamma is parameter of RBF. Finally, we consider the MA method for forecasting model. According to m, the researchers got the results of MA in Table 6.

In the results of MA model, m = 2 showed best performance in all IPC codes. All MES values of comparative models are showed in Table 7. In the result of IPC code C12M, the researchers found the MSE value of poisson regression was smallest in the comparative methods.

On the other hand, the researchers could get the best MES of C12N, C12P and C12Q using MA (m = 2). The patterns of time series plots Fig. 1-4 can be classified into two groups. C12M is belonged to one group. Another

group has C12N, C12P and C12Q. Generally, the researchers can expect the trend of bio-technology like the group including C12N, C12P and C12Q. So the researchers can know the MA model is effective to forecast technology.

CONCLUSION

Using MA and CLT, a technology forecasting model by frequent time series analysis was showed in this study. The researchers applied the model to bio-technology forecasting analysis. The experimental data were constructed using frequency by year in bio-technology patents. The researchers found the performance of MA was better than comparative methods which were linear regression, poisson regression and SVR. To get more advanced results, the researchers will new learning methods like hybrid neural networks and diverse time series models.

REFERENCES

Brockwell, P.J. and R.A. Davis, 2002. Introduction to Time Series and Forecasting. Springer-Verlag, New York, ISBN: 0387953515.

Burges, C.J.C., 1998. A tutorial on support vector machine for pattern recognition. Data Mining Knowledge Discovery, 2: 121-167.

Casella, G. and R.L. Berger, 2002. Statistical Inference. 2nd Edn. Duxbury Press, California, USA., ISBN: 0-534-24312-6.

Dernis, H., D. Guellec and B.V. Pottelsberghe, 2001. Using patent counts for cross-country comparisons of technology output. STI Rev. 27, OECD, 2001.

Fattori, M., G. Pedrazzi and R. Turra, 2003. Text mining applied to patent mapping: A practical business case. World Patent Inform., 25: 335-342.

Feinerer, I., K. Hornik and D. Meyer, 2008. Text mining infrastructure in R. J. Statistical Software, 25: 1-54.

Hall, B.H., A.B. Jaffe and M. Trajtenberg, 2001. The NBER patent citations data file: Lessons, insights and methodological tools. NBER Working Paper, 8498, http://www.nber.org/patents/

Han, J. and M. Kambr, 2001. Data Mining Concepts and Techniques. Higher Education Press, Beijing.

Haykin, S., 1999. Neural Networks a Comprehensive Foundation. Prentice-Hall, New Jersey.

Johnson, D.E., 1998. Applied Multivariate Methods for Data Analysts. Higher Education Press, USA., pp: 319-396.

Lee, S., B. Yoon and Y. Park, 2009. An approach to discovering new technology opportunities: Keyword-based patent map approach. Technovation, 29: 481-497.

Myers, R.H., 1989. Classical and Modern Regression with Applications. Duxbury Press, USA.

Tsay, R.S., 2005. Analysis of Financial Time Series. 2nd Edn., Wiley-Interscience, New York.

Vapnik, V.N., 1998. Statistical Learning Theory. 1st Edn., John Wiley and Sons, New York.

Wang, P., I.M. Cockburn and M.L. Putterman, 1998. Analysis of patent data-A mixed poisson regression model approach. J. Bus. Econ. Stat., 16: 27-41.

Yoon, B. and S. Lee, 2008. Patent analysis for technology forecasting: Sector-specific applications Proceedings of the IEEE International Conference on Engineering Management, June 28-30, IEEE International, pp: 1-5.

Yoon, B. and Y. Park, 2007. Development of new technology forecasting algorithm: Hybrid approach for morphology analysis and conjoint analysis of patent information. IEEE Trans. Eng. Manage., 54: 588-599.