Forecasting emerging technologies with deep learning and data augmentation: convergence emerging technologies vs non-convergence emerging technologies

Authors: Yuan Zhou[a], Fang Dong[a], Zhaofu Li[b], JunFei Du[b], Yufei Liu[c*], Li Zhang[b]

[a] School of Public Policy and Management, Tsinghua University, Beijing 100084, China

[b] School of Mechanical Science and Engineering, Huazhong University of Science and Technology, Wuhan 430074, China

[c] The CAE Center for Strategic Studies, Chinese Academy of Engineering, Beijing 100088, China

Abstract：

Forecasting emerging technologies can help governments and enterprises in various countries to grasp the key to success in a new round of technological competition. The emergence of emerging technologies has multiple patterns and the development of emerging technologies with different emerging patterns will have fundamentally different requirements for governments and enterprises. Technological innovation based on multi-field technology convergence which can create numerous opportunities for development has leaded to great attention. The current research to convergence emerging technologies forecasting is mainly network analysis, text mining and evolution path clustering. We can't achieve a full automation forecast process to utilize these unsupervised methods, due to a large amount of domain knowledge is required to analyze the characteristics of the convergence emerging technologies. Supervised learning can automatically obtain forecasting results that use the already trained model, however, which requires the selection of excellent features and a large amount of labeled data to train forecast model. In this paper, we transform the forecasting question of emerging technologies into a multi-classification supervised machine learning question and propose an approach that forecasting convergence emerging technologies, non-convergence emerging technologies with deep learning and data augmentation . Our evaluation reveals that our proposed approach can forecast the convergence emerging technologies and non-convergence emerging technologies 1 year before they emerge with high precision.

Keywords: Forecasting emerging technologies, Data augmentation, Deep learning

* Corresponding author: Yufei Liu, liuyufei0418@qq.com.

1 Introduction

With its characteristics of "creative revolution", emerging technologies have created opportunities for latecomers to achieve leaping development. Forecasting emerging technologies can help governments and enterprises in various countries to grasp the key to success in a new round of technological competition. The emergence of emerging technologies has multiple patterns and the development of emerging technologies with different emerging patterns will have fundamentally different requirements for governments and enterprises.

There are two emerging patterns for emerging technologies according to current research. On the one hand, it may arise from breakthrough innovations and create new technological paths. On the other hand, it may also result from the gradual innovation process and the convergence of two or more unrelated fields (Curran C S et al., 2010, Pennings J M et al., 2000, Stieglitz N et al., 2004). Among them, technological

innovation based on multi-field technology convergence has brought about tremendous changes in society and economy, which prompts changes in the existing market structure and competition rules as well as creates numerous development opportunities.

The current research to convergence emerging technologies forecasting is mainly network analysis, text mining and evolution path clustering. We can't achieve a full automation forecast process to utilize these unsupervised methods due to a large amount of domain knowledge required to analyse the characteristics of the emerging technologies in different patterns. Supervised learning can automatically obtain forecast results that use the already trained model, however, which requires the selection of excellent features and a large amount of labeled data. With the development of artificial intelligence technology, data augmentation can generate a large number of synthetic samples based on the distribution of real samples and deep learning can automatically combine features, which provide possibilities for overcoming above problems.

In this paper, we transform the forecasting question of emerging technologies into a multi-classification supervised machine learning question and propose an approach to forecast convergence emerging technologies (CET), non-convergence emerging technologies (NCET) with deep learning and data augmentation. First, we collect existing CET, NCET, non-emerging technologies (NET) and select multiple patent features for each technology as data samples. Second, the data samples are divided into training samples and testing samples and we use the generative adversarial network(GAN) to generate a large number of synthetic samples to amplify the scale of the training data. Finally, the deep neural network classifier(DNN) is trained through generated synthetic training samples and tested through testing samples.

As the empirical analysis, we used patent data from Thomson Innovation to evaluate the proposed approach. It reveals that our approach can forecast the CET and NCET 1 year before they emerge with high precision. This research provides an efficient and full automatic approach for forecasting emerging technologies. Further, this research is helpful to explore the potential application of artificial intelligence in forecasting emerging technologies.

2 Literature review

Forecasting emerging technologies is full of challenges. While technology forecasting in general has been acknowledged as an effective tool in order to anticipate and understand the potential direction, rate, and effects of technological change (Porter and Roper, 1991), these methods are not easy to forecasting emerging technologies.

We argue that the extent to which these methods are able to forecast emerging technologies, however, it is limited. For example, Qualitative studies were put forward to meet the demand of technology forecasting analysis in the early years. These included the Delphi method, scenario planning, interview analysis, etc. (Rowe and Wright, 2001; Daim et al., 2006; Drew, 2006; Shen et al., 2010; Cho et al., 2016, etc). There are many limitations to these methods, and the data sources for these

researches are subjective and expensive to collect. For the sake of the data reliability, many researchers conducted technology forecasting by combining qualitative studies with objective data. For example, Hsieh (2013) created a model that combined the Delphi method, fuzzy measurement, and a technology portfolio planning (TPP) model to analyze patents in large quantities before they were commercialized. However, due to the fact that these qualitative methods cost too much, they are usage were limited for many years.

Other researchers promote technology roadmapping as an ideal approach for technology forecasting (Walsh, 2004; Jin et al., 2015; Lee et al., 2015; Li et al., 2015; Cho et al., 2016; Phaal et al., 2004, etc.). This method employs various data sources and is used to explore and communicate the evolution of markets, products and technologies, together with the linkages and discontinuities between various perspectives (Phaal et al., 2004). There are two kinds of technology development paths, namely continuous and discontinuous development. While the scope of technology forecasting is limited somewhat to 'posteriori trends', discontinuous technology evolves along totally different trajectories, thus technology road mapping will only provide implications to firms who are focusing on continuous or sustaining technology, and does not provide answers for firms who are faced with discontinuous or disruptive technology (Kim et al., 2016).

After that, bibliometric methods, especially those which are based on patents and academic papers are widely used to conduct quantitative technology forecasting. In bibliometric analysis, it is assumed that the number of patents or papers is related to the validity and quality of R&D activities (Narin, 1994). The researches of relationships between patents or papers are well organized, specifically, various methods are proposed including Bayesian models for patent clustering (Choi and Jun, 2014), keyword-based patent map (Jin et al., 2015; Lee et al., 2009) and topical analysis (Ma and Porter, 2015). The advantages of patent bibliometics mainly lie in the reliability of data sources.

In recent years, the methods of text mining and machine learning began to appear in technology forecasting, and achieved a significant effect. A novel algorithm is proposed to automatically label data and then use the labeled data to train learners to forecast emerging technologies (Kyebambe M N et al., 2017). Dejing Kong (Kong D et al., 2017) proposed a novel method that combines data-mining with experts' knowledge to build patent-training examples, and then used a support vector machine-based classifier to single out all high-quality patents for each innovation attribute.

3 Methodology

3.1 Labeling CET, NCET, NET

In this paper, Gartner Emerging Technologies Hype Cycles (see www.gartner.com) is used to identify CET, NCET, NET. Gartner Emerging Technologies Hype Cycles was proposed by Gartner, the world's first information

technology research and analysis company, established in 1979. Gartner Emerging Technologies Hype Cycles, which began in 1995, aims to describe a specific stage of development of an emerging technology. Each year, Gartner selects a number of technologies that have significant potential or are highly hype-promoted from more than 2,000 technologies, resulting in an Emerging Technologies Hype Cycles. The curve has received extensive attention in various fields of society and is used to forecast and measure the development trend of technology. The lastest released 2017 Gartner Emerging Technologies Hype Cycles is shown in Appendix A .

According to the characteristics of Gartner Emerging Technologies Hype Cycles, we propose a new method of labeling emerging technologies and non-emerging technologies. The technology that has entered Gartner Emerging Technologies Hype Cycles for the first time in a certain year (T year) is a emerging technology. The technology that has left Gartner's emerging technology maturity curve for some time in a given year is a non-emerging technology. Emerging technologies must show strong consistency and persistence over time to meet emerging standards. Before a technology can be labeled, it is important to obtain a relatively stable state, because many new technologies show some emerging technologies features, but they have not gained stability, leading them to not become emerging technologies in the end. Therefore, if a group of patents show features that can produce new technology categories in the near future, a group of patents will be labeled as emerging technologies. The technology that first entered Gartner Emerging Technologies Hype Cycles in a certain year (T year) is labeled as an emerging technology that is associated with the patent data corresponding to that technology in the T-1 year. Simply, this means that T-1 year of patent data tagged as emerging technology may create a new category or technology in the second year.

At present, the pattern of production is usually determined by experts in a focused manner. In this paper, convergence emerging technologies and non-convergence emerging technologies are labeled by reading relevant documents.

3.2 Datasets construction

According to the retrieved Gartner Emerging Technologies Hype Cycles from 2008 to 2017 and the method of 3.1.1, CET, NCET, NET are identified. Then, in the Thomson Innovation database (TI), the patent data of the corresponding technology is retrieved. The technology feature vector of each technology is calculated from the patent data, and is used as the original sample for model training and model testing.

3.2.1 Patent data collection

Thomson Innovation is the only global innovation platform that integrates patents, scientific literature and business and news information, and provides unique tools for analysis, cooperation, and early warning, which is oriented toward corporate R&D, intellectual property and related decision-making departments. It has a global vision, high quality information and deep processing data. In addition to the patent database, there are several other sources of data for forecast technologies, such as social network analysis, scientific literature data, meeting records (Furukawa et al., 2014), etc. Since the TI patent data contains a comprehensive patents and its unique

advantages, the patent data retrieved and downloaded from the TI database is used as a data set.

We retrieve the patent data of convergence emerging technologies, non-convergence emerging technologies, and non-emerging technologies labeled in Section 3.1. For selected technologies that become emerging technologies in a certain year (T year), patent data up to T-1 year is retrieved. Correspondingly, for selected technologies that become non-emerging technologies in a certain year (T year), patent data up to T-1 year is retrieved. The multi-mode emerging technologies are also carrying out corresponding patent data retrieval according to the way push time forward one year.

Taking cloud computing as an example, cloud computing technology first appeared on the Gartner Emerging Technologies Hype Cycles in 2008, which was an emerging technology in that year and was a non-emerging technology since 2015. Using TI as a source of patent data, according to preliminary understanding of cloud computing technology, the search formula was determined as: " ALLD=（cloud computering OR cloud storage OR cloud infrastructure OR private cloud OR public cloud OR Iaas OR PaaS OR SaaS）NOT (IC=A*) NOT (IC=F*) NOT (IP=G01*)". From 2008 to 2014, the search was done on April 15th, 2018, and a total of 8,273 patent documents were retrieved.

3.2.2 Patent features selection

We downloaded the required patent data and extracted features from each patent to describe the differences between convergence emerging technologies, non-convergence emerging technologies, and non-emerging technologies. It is the most critical part of our research, because the accuracy of the forecast depends on a large extent on the correlation of these features with the emerging technology's forecast model. Early studies attempted to define the possible features of the emerging technology generation model. The results of these researches were used as the guideline for the following feature selection. We extracted or derived the following characteristics for patent data corresponding to each technology.

a. Emerging technology features

(1) Number of claims

Break through patents have been found to have a higher renewal rate and higher number of claims (Moore, 2004). Inventors of the important invention are more likely to make claims as much as they can to protect their interest.

(2) Number of citations

Inventors and examiners cite other patents to show dependency of the citing technology to the cited technology (Newman, 2010) as well as to pre-empt any later claims that the cited technology invalidates patentability of the citing technology. In both cases, a high number of citations represents stronger relationship between the citing technology and other technologies. So, a patent likely to be a turning point in a given industry is likely to be distinguished from other patents by the number of citations it receives.

(3) Number of citations made to non-patent literature

Non-patent literature is usually result of the investment by government or large firms, cause many new inventions are implementations of scientific research findings. As such, new inventions are more likely to have a strong connection to non-patent literature. This feature is closely related to the Originality and Science indices used by

(Breitzman and Thomas, 2015) to score patent clusters that are likely to contain new inventions.

(4) Technology Cycle Time (TCT)

TCT is the mean age (in years) of patents cited by a patent. This index measures how fast of the technology in a given technological area changes; a small value of TCT indicates fast changing of technological generations within an area. Within a fast growing technological field, we expect patents containing emerging technologies to have relatively smaller TCT values compared to those merely incrementing on existing technologies. TCT is computed as shown in Eq. (1);

$$TCT_i = median_j \left\{ \left| T_i - T_j \right| \right\} \quad (1)$$

where T_i is the application date of patent i and patents i and j are connected.

(5) Patent class

A patent classification comprises of a main class and a subclass(Kyebambe M N et al., 2017). As the features we have discussed and those we are yet to discuss may vary significantly across patent main classes. We include patent class among the features to harmonize differences among other features across patent classes.

(6) Patent cited times

The reference of patents to prior art and scientific papers is a manifestation of the laws of science and technology development. It embodies the accumulation, continuity, and inheritance of science and technology as well as the intersection and penetration of different disciplines and research levels. . Abert et al. (Abert et al., 1991, Harhoff et al., 1999) have considered that the patent cited times can be directly used as an indicator to identify important patents for enterprises. If a patent is cited by many subsequent patents, it indicates that the invention involved in the patent is a relatively important and important technology. Those patents with high citations are often high quality patents.

(7) Family patent size

All patent documents in the same patent family are called patent family members. Patent family size refers to the number that the same invention obtained patents or filed patent applications in different countries, or the number of countries that the applicant sought patent protection for the same invention. With the increase in the number of countries seeking protection, the cost of patents has also increased. Applicants are more willing to do so for high-tech quality inventions with economic value. At the same time, applying for a patent to another country means the applicant judges that the invention may have international competitiveness. If the patent is finally granted by many countries, it means that the invention can stand the test of many parties and has high technical value. Therefore, the size of the patent family also reflects the economic importance and technical importance of the invention.

(8) The type of IPC

The IPC classification was compiled according to the Strasbourg Agreement on International Patent Classifications signed in 1971. It is currently the only internationally-available tool for classifying and searching patent documents and is an essential tool for countries in the world. The number of IPC classification numbers in a field's patents represents the complexity of the field. The more sub-fields, the higher possibility of convergence.

b. Convergence emerging technology features

(1) The entropy of the node

Entropy denotes the degree of "disorder" in a system and also represents a measure of the quality of life activity process. The concept of entropy is introduced into technology fusion to measure the "quality" of a certain technology in technology fusion. The greater the entropy value, the greater the role played by the technology in technology fusion. The measurement of entropy in technology fusion can provide reference for related companies' R&D priorities. Enterprises can pay attention to related technologies with high entropy and high attention(Lou et at., 2017). Entropy is computed as shown in Eq. (2);

$$E_i = -\sum_j P_{j/i} \log_2 \left(P_{j/i} \right) \tag{2}$$

where E_i represents the entropy value of the i-th IPC, and j is an IPC code different from the IPC code i; $P_{j/i}$ is the number of patents associated with both IPC code i and code j divided by the number of patents associated with IPC code i.

(2) Node strength

Node strength is expressed using the node average weight value, which is the ratio between the unweighted average and the weighted average. Therefore, the unweighted average and the weighted average are respectively measured. The unweighted average represents the number of each node, and the weighted average represents the fusion importance of each node (Akade'miai Kiado' Budapest et at., 2015). The formula is as follows:

$$UAD = \frac{1}{2} \frac{\sum_i^n l_{ij}}{n}, \ where \ \ i \neq j \ \ and \ \ l_{ij} = \begin{cases} 1, \ if \ \ w_{ij} > 0 \\ 0, \ else \end{cases} \tag{3}$$

$$WAD = \frac{1}{2} \frac{\sum_i^n w_{ij}}{n}, \ where \ i \neq j \tag{4}$$

$$AWL = \sum_i^n \frac{w_{ij}}{l_{ij}}, \ where \ i \neq j \ and \ l_{ij} = \begin{cases} 1, \ if \ \ w_{ij} > 0 \\ 0, \ else \end{cases} \tag{5}$$

where l_{ij} is the number of technologies that have convergence with a certain technology i; w_{ij} is the total number of fusions of a technology and other technologies, that is, the total number that from one node to another nodes, and n is the total number of nodes.

(3) Jaccard coefficient

The Jaccard coefficient(Lou et at., 2017), also known as the Jaccard similarity coefficient, is used to compare the probability of similarity and dispersion in the sample set, which is used to measure link weights. Jaccard coefficient equals the ratio of sample set intersection to sample set union. Jaccard coefficient is computed as shown in Eq. (6);

$$Jaccard \ \ coefficient = \frac{n(i \cap j)}{n(i \cup j)} \cdot 100\% = w_{ij} \tag{6}$$

where, technology i and technology j are constituted by their technical connotation and technical boundary respectively. The size of the intersection of the two represents the degree of convergence of the technology i and the technology j, that is, the number of links between the technology i and the technology j. The size of the

union represents the total number of patents of the technology i and the technology j, that is, the total number of links from technology i and technology j.

(4) Cross impact

The cross-impact method defines the influence of technology i and technology j $\mathrm{Im}\,pact(i,j)$ as a conditional probability $n(j/i)$. The impact value is the ratio of the number of patents contained in both technology i and technology j to the number of patents contained in technology i. It was proposed by Changwoo et al(CHANGWOO GHOI et at., 2007). Grouping links based on cross- impact value will help further analyze the relationships between technologies. The cross- impact value is computed as shown in Eq. (7);

$$\mathrm{Im}\,pact(i,j)=\frac{n(i\cap j)}{n(i\cup i)} \tag{7}$$

3.2.3 Training & testing datasets construction

The feature vector of each technology is constructed according to the features of 3.2.2. First, for each of the eight features used in the emerging technologies, the patents corresponding to each technology are calculated and an 8-dimensional vector is directly formed, that is, each technology corresponds to an 8-dimensional vector. Then, for the features of convergence emerging technologies, these features are for IPCs of patents, patents for a certain technology involve multiple IPCs, and there are multiple values. Here, the average value is used as a feature of the convergence emerging technology, and a 4-dimensional vector is formed, that is, each technique corresponds to a 4-dimensional vector. Finally, we combine the features of emerging technologies and the features of convergence emerging technologies, so that a technology feature vector is formed. That is, each technology is represented as a 12-dimensional vector.

Use the data obtained in section 3.2.1 to create a corresponding feature vector for each technology, where the dimension of each vector is one of the features described in section 3.2.2. Our ultimate goal is to train models that can forecast the eigenvectors corresponding to the technology domain in a given year as whether they are emerging in a convergence emerging technologies.

Each technology of the convergence emerging technologies, non-convergence emerging technologies, and non-emerging technologies selected in this paper corresponds to a 12-dimensional feature vector. In order to build the independence of the training set and the test set, the original sample is divided into a training set and a test set in a random manner. The technology feature vectors corresponding to convergence emerging technologies, non-convergence emerging technologies, and non-emerging technologies are randomly selected 70% as training set, the remaining 30% as a test set. The WGAN model is trained using the training set, and data augmentation is performed. The DNN classifier is trained using the enhanced data as the training set, and the DNN classifier is tested using the test set.

3.3 Data augmentation based GAN

The number of sample categories is unbalanced, and it is difficult to get a high-quality model based on statistical machine learning classification methods. At the same time, the total number of samples in each category is small, and it is difficult to effectively train a DNN classification model based on deep learning. We use GAN to enhance the original data samples and generate a large number of identically distributed virtual samples. This effectively solves the problem of small and

unbalanced samples. GAN is a powerful type of generative models (Wang K et at., 2017) introduced in 2014 by Goodfellow (Goodfellow I et at., 2014). GAN comprises two deep architecture functions for the generator and discriminator, which can learn from the trained data in an adversarial fashion simultaneously (Radford A et at., 2015).

The process of generating synthetic samples based on GAN contains two stages in the proposed approach. First, the generator begins to generate original synthetic samples when the loss function of generator and discriminator have converged after training of tens of thousands times. Second, according to GAN's idea of adversarial (Mcdaniel P et at., 2016), the generator tries to generate synthetic samples which can fake discriminator out, but the discriminator tries to discriminating real samples and synthetic samples. In other words, Original synthetic samples which can fake discriminator out will be the final synthetic samples.

3.4 Forecasting based on deep learning

In the big data environment, deep learning, as the core of the big-data intelligent method, compared with classical statistical machine learning methodologies, has a more complex model structure. The amount and the quality of a data set can significantly affect the deep learning classifier. It requires large-scaled annotated sample data to make model parameters fully optimized and make the performance superior (Goodfellow I et at., 2016).

In order to further enhance the forecast effect of the multi-pattern emerging technologies, after the GAN performs data augmentation on the original samples, a DNN classifier is selected for forecasting. DNN classifier is trained with a large number of synthetic samples generated by GAN to avoid overfitting and is tested with partially independent real samples in the proposed approach. After DNN classifier is trained, we use three multi-classification metrics based on a confusion matrix: accuracy, F-measure, G-mean. The accuracy is the proportion of predictions that are correct , F-measure is the harmonic mean of precision and recall (Sousa L R E et at., 2017), and the G-mean is the geometric mean of recall (Sun Y et at., 2007).

4 Result and Discussion

4.1 Analysis result of proposed approach

Through retrieving and labeling these technologies except the unrecovered technologies, we identify a total of 57 emerging technologies and 48 non-emerging technologies. On the basis of the identified 57 emerging technologies, the emerging technologies were identified through the support of experts and literature data, and 17 convergence emerging technologies were identified. The data is then divided into three categories: convergence emerging technologies, non-convergence emerging technologies, and non-emerging technologies. According to Gartner Emerging Technologies Hype Cycles from 2008 to 2017, 17 convergence emerging technologies, 40 non-convergence emerging technologies, and 48 non-emerging technologies were extracted, as shown in Appendix A. Retrieve and download patents for all convergence emerging technologies, non-convergence emerging technologies, and non-emerging technologies, and calculate the technology feature vector of each technology as the original sample. Table 1 lists the number of patents retrieved for each technology.

Table 1 The number of patents retrieved for each technology

Technology Category	Technology Name	Number of Patents	Technology Name	Number of Patents
CET	Surface Computers	358	Wireless Power	6
	Internet TV	438	Video Search	6
	Home Health Monitoring	19	Interactive TV	6
	Autonomous Vehicles	409	Broadband over Power Lines	6
	IOT	125	Big Data	6
	Social TV	7	3D Scanners	6
	Hybrid Cloud Computing	7	Connected Home	6
	Machine Learning	3060	IOT Platforms	6
	Nanotube Electronics	66		
NCET	Solid State Drives	96	Digital Dexterity	209
	Green IT	71	Virtual Reality	4598
	ebook readers	11	Biochips	1023
	Human Augmentation	9	Affective Computing	65
	Terahertz Waves	401	Electrovibration	6
	Activity streams	106	Digital Security	6
	Virtual Assistants	50	Data Science	6
	Consumer Generated Media	30	Image/Content Recognition	6
	Media Tablets	27	Micro Data Centers	6
	Tangible/Conversational User Interfaces	17	Volumetric & Holographic Displays	6
	Complex event processing	84	802.11ax	6
	QR/Color Code	3246	4D Printing	6
	Group Buying	111	5G	6
	Application Stores	1755	Deep Learning	6
	Blockchain	44	Edge Computing	6
	HTML5	49	Cognitive Computing	6
	NFC (Payments)	5	Mobile Health Monitoring	6
	Neuromorphic hardware	5	Deep Reinforcement Learning	6
	Cloud Computing	5	Consumer Telematics	6
	Crowdsourcing	17	Digital Twin	6
NET	Surface Computers	118	IP Video/Internet Video	6
	Green IT	22	Terahertz Waves	6
	Interactive TV	130	Microblogging	6
	Video Search	86	Broadband over Power Lines	6

Consumer Generated Media	7	QR/Color Code	6
ebook readers	62	Group Buying	6
Wireless Power	2267	Image/Content Recognition	6
Application Stores	434	Internet TV	6
HTML5	89	Crowdsourcing	6
Private Cloud Computing	26	Media Tablets	6
Home Health Monitoring	9	BYOD	6
Predictive Analytics	82	Cloud Computing	6
Big Data	426	3D Scanners	6
Activity streams	166	Complex event processing	6
NFC (Payments)	73	Gamification	6
Machine to Machine Comm. Services	7	Mobile Health Monitoring	6
In Memory Analytics	7	Data Science	6
IOT	1487	Biochips	6
Digital Security	60	Hybrid Cloud Computing	6
Affective Computing	22	3D Bioprinting	6
Cryptocurrencies	8	802.11ax	6
Social Analytics	4	Prescriptive Analytics	6
Social TV	2	3D Flat Panel Displays	6
Micro Data Centers	15	Solid State Drives	6

According to the approach we proposed, we first use the actual sample training set corresponding to the convergence emerging technologies, non-convergence emerging technologies, and non-emerging technologies respectively selected in Section 4.2, and then use the trained GAN to generate 1000 corresponding synthetic samples. Hyperparameters for the GAN were empirically determined. The generator has 2 hidden layers that respectively contains 4 ReLU units, and 12 softmax units are used as output layer, and the dimension of the noise vector z is set to 5. The discriminator also has 2 hidden layers that respectively contains 4 ReLU units, and 1 activation function unit is used as output layer. The WGAN's development environment is tensorFlow1.1, and it is trained through GPU. In each iteration of WGAN training, discriminator first iterates 100 times, then the generator iterates 1 time.

After synthetic samples have been generated, synthetic samples are used to train DNN classifier. Then the DNN classifier is tested using the divided 30% of the test set samples. The dimension of classifier's input is 12, which is equal to the number of features in liver cancer samples. The classifier has 2 hidden layers, each containing 32 ReLU units while softmax is used as output layer and cross-entropy is used as loss function. TensorFlow1.1 and GPU are used for training DNN classifier as well and number of iteration is set to 3000 times.

The forecast results of the DNN classifier on the test set data are shown in Table 2. In the test set, the 4 of 5 convergence emerging technologies are forecasted to be

correct, and 1 is non-emerging technologies. The 8 of 10 non-convergence emerging technologies are forecasted to be correct, and 1 is non-emerging technologies, and 1 is convergence emerging technologies. The 7 of 10 non-emerging technologies are forecasted to be correct, and 2 are non-convergence emerging technologies, and 1 is convergence emerging technologies. According to the results, there are a total of 15 emerging technologies for the 5 convergence emerging technologies and 10 non-convergence emerging technologies, and only 2 technologies are identified as non-emerging technologies with an accuracy rate of 86.67%, which indicates that the forecast model can forecast emergence of emerging technologies with an accuracy of 86.67% a year before they emerge. Among all emerging technologies, the number of correctly forecast convergence emerging technologies and non-convergence emerging technologies is 12 with an accuracy rate of 80.00%, which indicates that the forecast model can forecast emergence of emerging technologies with an accuracy of 80.00% a year before they emerge.

Table 2 The experiment results of DNN classifier for the test set

	Forecast			Total
	CET	NCET	NET	
CET	4	0	1	5
NCET	1	8	1	10
NET	1	2	7	10
Total	6	10	9	25

4.2 Evaluation of proposed approach

In order to verify the effectiveness of our proposed method, we used the statistical classifiers Random Forest (RF) and Naive Bayes (NB) for comparison experiments. As a classic machine learning classifier, RF and NB have higher classification accuracy and better generalization performance than machine learning classifiers. The results of comparative experiments in the three categories of CET, NCET, and NET are shown in Table 3.

Table 3 The results of classifier comparison experiment

Model	Accuracy	F-measure	G-mean
NativeBayes	0.4000	0.4150	0.4160
RandomForest	0.5200	0.4981	0.4932
GAN-DNN	0.7600	0.7573	0.7652

From Table 3, we can see that on the same data set, the Accuracy, F-measure, and G-mean of RF classifiers and NB classifiers are lower than the GAN-DNN proposed by us. The comparison of evaluation indicators shows that the classification quality of classical classifiers RF and NB based on statistics is lower than the combination of the GAN-based data augmentation algorithm and the DNN-based deep learning algorithm proposed by us. The comparison test results show that GAN combined with DNN deep learning method can effectively solve the forecast problems of CET, NCET, and NET.

4.3 Discussion

According to the above results, the proposed method of data augmentation combined with deep learning has achieved good results in the forecast of CET and NCET. This result has important significance for forecasting emerging technologies.

First, there are multiple patterns for the emergence of emerging technologies. This paper has effectively explored the forecast of convergence emerging technologies and non-convergence emerging technologies, and has also provided an effective research method for forecasting other emerging patterns of emerging technology. Second, through the method proposed by us, we can effectively transform the emerging technology forecast problem into a supervised machine learning classification problem, providing an automated solution for forecasting emerging technologies. Finally, the deep learning and data augmentation methods effectively solve the data problems in the forecast of CET and NCET using supervised machine learning methods. This also provides ideas for the application of artificial intelligence methods such as data agumentation and deep learning in forecasting emerging technologies.

5 Conclusions

In this paper, we proposed an approach based on deep learning and data augmentation that can forecast the CET and NCET 1 year before they emerge with high quality. Compared with the traditional methods, we not only forcast the emerging technology from the generation pattern, but also effectively transform the forecasting question of emerging technologies into a multi-classification supervised machine learning question, which can automatically forecast convergence emerging technologies and non-convergence emerging technologies. Furthermore, the effective application of artificial intelligence methods such as deep learning and data augmentation in this research also provides an effective way for artificial intelligence technology to be applied in forecasting emerging technologies.

Acknowledgement

This paper is one of the research achievements of the National Natural Science Foundation of China "Supporting Technology Foresight for Multi-source Heterogeneous Big Data Fusion and Time Series Text Prediction" (No. 91646102), Project leader: Zhou Yuan. One of the research results of the National Natural Science Foundation "Research on the 2035-oriented China Engineering Science and Technology Development Road Map Drawing Theory and Method" (No. L1624045), Project Leader: Zhou Yuan. One of the research results of the National Natural Science Foundation of China "National Engineering Science and Technology Development Road Map Application Case and Software Research for the 2035 (No. L1624041), Project Leader: Liu Hualan. One of the research results of National Natural Science Foundation of China “Bibliometrics and Bibliometric Analysis of Development Strategies” (No. L1524015), Project Leader: Zhou Yuan. One of the research results of National Natural Science Foundation of China "Study on the Mechanism of the Demonstration Project to the Industrial Technology Orbit - Taking the New Energy Vehicle and the New Energy Industry for Example" (No. 71203117), Project Leader: Zhou Yuan.

References

Abert M B, Avery D, Narin F, et al. Diretc validation of citation counts as indicators of industrially important patents[J]. Research Policy, 1991, 20(3): 251-259.

Akade'miai Kiado' Budapest, Hungary. Technology Convergence: What Developmental Stage are we in? [J]. Scientometrics, 2015, 104(3): 841-871.

Breitzman, A., Thomas, P., 2015. The emerging clusters model: a tool for identifying emerging technologies across multiple patent systems. Res. Policy 44, 195–205.

CHANGWOO GHOI, SEUNGKYUM KIM, YONGTAE PARK. A patent-based Cross Impact Analysis for Quantitative Estimation of Technological Impact: the Case of Information and Communication Technology[J]. Technological Forecasting & Social Change, 2007（74）: 1296-1314.

Cho, C., Yoon, B., Coh, B.Y., Lee, S., 2016. An empirical analysis on purposes, drivers and activities of technology opportunity discovery: the case of Korean SMEs in the manufacturing sector. R&D Manag. 46 (1), 13–35.

Choi, S., Jun, S., 2014. Vacant technology forecasting using new Bayesian patent clustering. Tech. Anal. Strat. Manag. 26 (3), 241–251.

Curran C S, Bröring S, Leker J. Anticipating converging industries using publicly available data[J]. Technological Forecasting & Social Change, 2010, 77(3):385-395.

Daim, T.U., Rueda, G., Martin, H., Gerdsri, P., 2006. Forecasting emerging technologies: use of bibliometrics and patent analysis. Technol. Forecast. Soc. Chang. 73(8), 981–1012.

Drew, S.A., 2006. Building technology foresight: using scenarios to embrace innovation. Eur. J. Innov. Manag. 9 (3), 241–257.

Furukawa, T., Mori, K., Arino, K., Hayashi, K., Shirakawa, N., 2014. Identifying the evolutionary process of emerging technologies: a chronological network analysis of World Wide Web conference sessions. Technol. Forecast. Soc. Chang. 91, 280–294.

Goodfellow I, Bengio Y, Courville A. Deep Learning[M]. The MIT Press, 2016.

Goodfellow I, Pougetabadie J, Mirza M, et al. Generative Adversarial Nets[J]. Advances in Neural Information Processing Systems, 2014:2672-2680.

Harhoff D, Narin F, Schere F M, et al. Citation frequency and the value of patented inventions[J]. The Review of Economics and Statistics, 1999(8): 511-515.

Hsieh, C.H., 2013. Patent value assessment and commercialization strategy. Technol. Forecast. Soc. Chang. 80 (2), 307–319.

Jin, G., Jeong, Y., Yoon, B., 2015. Technology-driven roadmaps for identifying new product/ market opportunities: use of text mining and quality function deployment. Adv. Eng. Inform. 29 (1), 126–138.

Kim, J., Park, Y., Lee, Y., 2016. A visual scanning of potential disruptive signals for technology roadmapping: investigating keyword cluster, intensity, and relationship in futuristic data. Tech. Anal. Strat. Manag. 1–22.

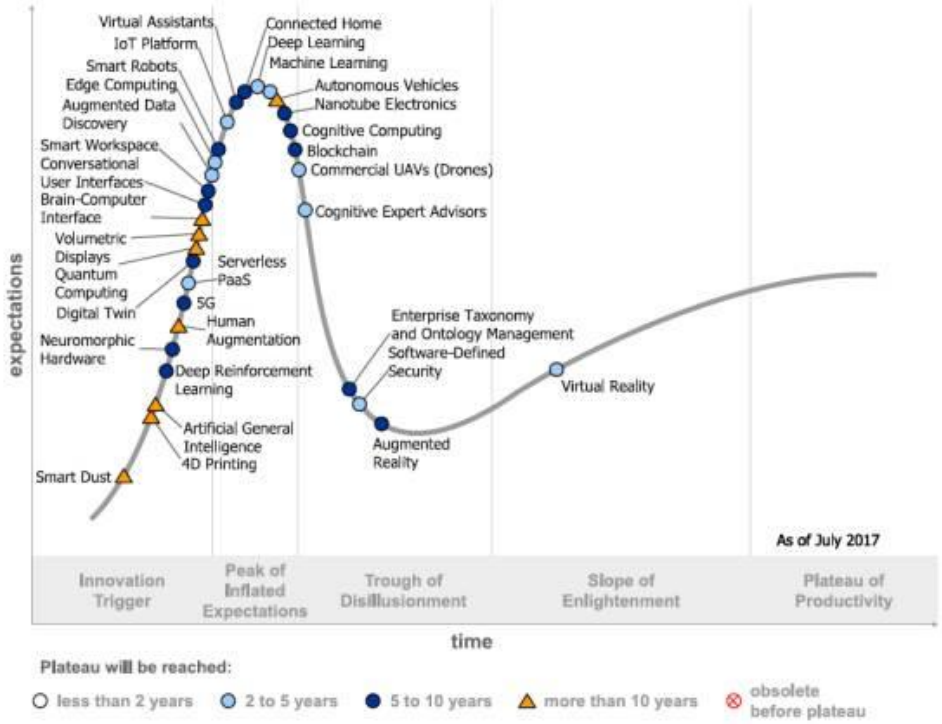
Kong D, Zhou Y, Liu Y, et al. Using the data mining method to assess the innovation gap: A case of industrial robotics in a catching-up country[J]. Technological Forecasting and Social Change, 2017, 119: 80-97.

Kyebambe M N, Cheng G, Huang Y, et al. Forecasting emerging technologies: A supervised learning approach through patent analysis[J]. Technological Forecasting and Social Change, 2017, 125: 236-244.

Lee, C., Song, B., Park, Y., 2015. An instrument for scenario-based technology roadmapping: How to assess the impacts of future changes on organisational plans. Technol. Forecast. Soc. Change. 90, 285–301.

Lee, S., Yoon, B., Lee, C., Park, J., 2009. Business planning based on technological capabilities: patent analysis for technology-driven roadmapping. Technol. Forecast. Soc. Chang. 76 (6), 769–786.
Li, X., Zhou, Y., Xue, L., Huang, L., 2015. Integrating bibliometrics and roadmapping methods: a case of dye-sensitized solar cell technology-based industry in China. Technol. Forecast. Soc. Chang. 97, 205–222.
Lou Y, Yang, P P, Huang L C, Miao H, et al. Patent-based technology fusion measurement method——using the fusion of information technology and electric vehicle technology as an example[J].Journal of Modern Information,2017,37(08):142-153.
Ma, J., Porter, A.L., 2015. Analyzing patent topical information to identify technology pathways and potential opportunities. Scientometrics 102 (1), 811–827.
Mcdaniel P, Papernot N, Celik Z B. Machine Learning in Adversarial Settings[J]. IEEE Security & Privacy, 2016, 14(3):68-72.
Moore, K.A., 2004. Worthless patents. http://dx.doi.org/10.2139/ssrn.566941.
Narin, F., 1994. Patent bibliometrics. Scientometrics 30 (1), 147–155.
Newman, M., 2010. Networks: An Introduction. Oxford University Press.
Pennings J M, Puranam P. Market convergence & firm strategies: towards a systematic analysis[J]. Retrieved August, 2000.
Phaal, R., Farrukh, C.J., Probert, D.R., 2004. Technology roadmapping—a planning framework for evolution and revolution. Technol. Forecast. Soc. Chang. 71 (1), 5–26.
Porter, A.L., Roper, A.T., 1991. Forecasting and Management of Technology vol. 18. John Wiley & Sons.
Radford A, Metz L, Chintala S. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks[J]. Computer Science, 2015.
Rowe, G., Wright, G., 2001. Expert opinions in forecasting: the role of the Delphi technique. Principles of forecasting. Springer, US, pp. 125–144.
Shen, Y.C., Chang, S.H., Lin, G.T., Yu, H.C., 2010. A hybrid selection model for emerging technology. Technol. Forecast. Soc. Chang. 77 (1), 151–166.
Sousa L R E, Miranda T, Sousa R L E, et al. The Use of Data Mining Techniques in Rockburst Risk Assessment[J]. Engineering, 2017, 3(4):552-558.
Stieglitz N. Strategie und Wettbewerb in konvergierenden Märkten[J]. 2004.
Sun Y, Kamel M S, Wong A K C, et al. Cost-sensitive boosting for classification of imbalanced data[J]. Pattern Recognition, 2007, 40(12):3358-3378.
Walsh, S.T., 2004. Roadmapping a disruptive technology: a case study: the emerging microsystems and top-down nanosystems industry. Technol. Forecast. Soc. Chang. 71 (1), 161–185.
Wang K, Gou C, Duan Y, et al. Generative adversarial networks: introduction and outlook[J]. IEEE/CAA Journal of Automatica Sinica, 2017, 4(4):588-598.

Appendix A



2017 Gartner Emerging Technologies Hype Cycles

Appendix B

2008 - 2017 CET, NCET, NET List

Year	CET	NCET	NET
2008	Surface Computers	Solid State Drives, Green IT, Cloud Computing	
2009	Wireless Power, Internet TV, Video Search, Home Health Monitoring	ebook readers, Human Augmentation	Solid State Drives
2010	Interactive TV, Autonomous Vehicles, Broadband over Power Lines	Terahertz Waves, Activity streams, Virtual Assistants, Consumer Generated Media, Media Tablets, Tangible/Conversational User Interfaces	Surface Computers, IP Video/Internet Video, Green IT
2011	IOT, Big Data, Social TV	Image/Content Recognition, QR/Color Code, Group Buying, NFC (Payments)	Terahertz Waves, Interactive TV, Microblogging, Video Search, Broadband over Power Lines, Consumer Generated Media, 3D Flat

			Panel Displays
2012	3D Scanners, Hybrid Cloud Computing	Application Stores, Complex event processing, HTML5, Crowdsourcing, Volumetric & Holographic Displays, Consumer Telematics	QR/Color Code, ebook readers, Group Buying, Social TV
2013		Virtual Reality, Biochips, Affective Computing, Electrovibration, Mobile Health Monitoring	Wireless Power, Image/Content Recognition, Application Stores, Internet TV, HTML5, Crowdsourcing, Private Cloud Computing, Media Tablets, Home Health Monitoring, BYOD, Social Analytics
2014	Connected Home	Digital Security, Data Science	Predictive Analytics
2015	Machine Learning, IOT Platforms	Digital Dexterity, Micro Data Centers	Cloud Computing, Big Data, 3D Scanners, Activity streams, Complex event processing, NFC (Payments), Gamification, Machine to Machine Comm. Services, In Memory Analytics, Mobile Health Monitoring, Data Science, Prescriptive Analytics
2016	Nanotube Electronics	Blockchain, 802.11ax, 4D Printing, Neuromorphic hardware	IOT, Biochips, Digital Security, Hybrid Cloud Computing, Affective Computing, 3D Bioprinting, Cryptocurrencies
2017		5G, Deep Learning, Edge Computing, Cognitive Computing, Digital Twin, Deep Reinforcement Learning	802.11ax, Micro Data Centers