Advances in Artificial Intelligence, Big Data and Algorithms
G. Grigoras and P. Lorenz (Eds.)

doi:10.3233/FAIA230832

Feature Selection of High Value Patents Based on Random Forest

Zongze LI, Songxian LIU, Yihui QIU[1], Xin YANG
Xiamen University of Technology, Xiamen, China

Abstract. In the process of high value patents evaluation, the important features of patents obtained through feature selection largely affect the performance of the classifier. The traditional feature selection method has the problem of single evaluation factor, and should combine multiple evaluation factors for a more comprehensive screening of features, so this paper proposes a feature selection of high value patents based on random forest, which obtains through the random forest algorithm The method obtains the influence factor, enhancement factor and importance factor of features by random forest algorithm, and combines the three factors for a more comprehensive screening of the original patent feature set to obtain the important features of high value patents. Through comparison experiments with the single factor method, it is proved that the method has better feature selection effect.

Keywords. High value patents, feature selection, random forest, patent valuation

1. Introduction

The existing patent value evaluation research is still at a relatively preliminary stage, because the patent value has the characteristic of being difficult to measure directly, so many evaluation tasks are still done manually by industry experts, and the accuracy and efficiency of the evaluation cannot be guaranteed. However, patent value is often closely related to certain characteristics of patents, so it is of great research significance to realize accurate and efficient patent value evaluation by screening out these characteristics that have an important influence on patent value evaluation through scientific feature selection methods. With the development of computer technology, machine learning has been widely applied to the study of feature selection, and how to filter out the important features that are more effective for patent value evaluation through machine learning has become one of the hot spots for scholars to study [1-3].

2. Related Work

Traditional feature selection methods are mainly based on statistical methods. Harhoff et al [4] considered that the number of citations of patent documents can be used as an important feature for evaluating high value patents; Li et al [5] proposed that citation network has a close relationship with patent value based on the number of citations;

[1]Corresponding Author, Yihui QIU, Xiamen University of Technology, China;
e-mail: qiuyihui@xmut.edu.cn

Bekkers et al [6] proposed that the number and content of claims are the determining factors for measuring high value patents; Ping Xie [7] calculated the high value patents evaluation method by TOPSIS method of feature weights; Yuan Run et al [8] evaluated high value patents by rough set theory on the basis of constructing a high value patents identification framework containing three index features and determining a high value patents index system containing eight index features.

The disadvantage of the traditional feature selection methods are that the process of statistical calculation is relatively complex, the data has a certain degree of complexity, and the data features will grow exponentially due to the increase in dimensionality. Therefore, statistical methods are still difficult and not very practical to use in reality. With the development of machine learning technology, more machine learning technology has been applied to feature selection methods in the field of patent valuation, and considerable research results have been achieved. Heeyong Noh et al. [9] filtered the subject matter features of technologies through machine learning methods, and believed that patents in line with the subject features of technology had higher value. Through research, filtered the subject matter features of technologies through machine learning methods and concluded that patents that conform to the subject matter features of technologies have higher values; Kim C et al. [10] found that high value patents are often characterized by overlapping technical subject features; C. Y. Lee et al. [11] used machine learning technology to construct a patent value evaluation index system and carried out early identification research on high value patents; Yihui Qiu et al. [12] constructed a high value patents index system by screening out important features based on the feature selection method of classification regression tree model; Limin Bai et al. [13] proposed an intelligent evaluation model of potential high value patents based on a neural network algorithm on the basis of constructing three major index systems of basic index, technical index and market index.

In summary, with the development of computer technology and the increase of complexity of patent value evaluation problems, advanced technologies such as machine learning are widely used in the field of patent value evaluation, and more and more scholars apply machine learning technology to the research of feature selection of high value patents, which has an important role in promoting the efficiency and effectiveness of patent value evaluation.

3. Theoretical Background

3.1. High Value Patents

As a form of intellectual property, patent provides legal protection to patent owners and is an important protection barrier for intellectual property. The number of patent applications granted in China has increased rapidly in recent years, with the number of patent applications granted increasing by about five times from 2011 to 2021, as shown in Figure 1. Although the number of patents is increasing, only a few of them are high value patents. The essence of the fierce competition between enterprises in the future is the competition of intellectual property rights, among which high value patents are particularly important, which often represent the core competitiveness and core technology of enterprises.

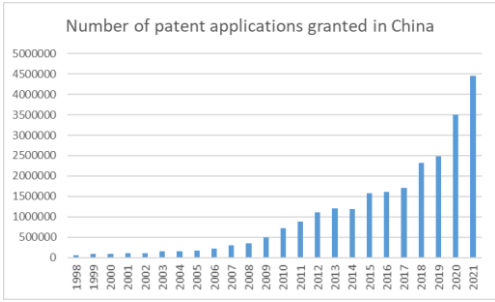

Figure 1. Growth chart of domestic patent application authorization from 1998 to 2021.

To study the feature selection in the process of high value patents evaluation, the connotation of high value patents should be clearly defined. Although there is no clear definition of high value patents in basic theoretical research at present, in practice, it can be found that the value of patent is gradually formed in the process of technology research and development, application confirmation, technology transformation and technology application, which runs through the whole process of patent from creation to application. High value patents often have the characteristics of high technical value, high legal value, high economic value and high market value. Therefore, in the process of feature selection for high value patents, factors to be considered include technical factors, legal factors, economic factors and market factors.

3.2. Random Forest

How to screen features from patent data and measure patent value according to the selected features and index system is an important step in high value patents evaluation. Random forest, as a supervised machine learning integration algorithm, can achieve better generalization performance than single decision tree by constructing and combining multiple decision trees to complete the task. Even though the ability of each decision tree is very weak and the prediction accuracy is very low, the accuracy of random forest algorithm is significantly improved after combination. In the high value patents evaluation task, because of its strong generalization ability and high classification performance, it is outstanding in the patent classification task, and it is also suitable for the screening of patent characteristic index.

As shown in Figure 2, the main steps of building a random forest model are as follows: Suppose the original data set X contains N samples, each with m-dimensional features.

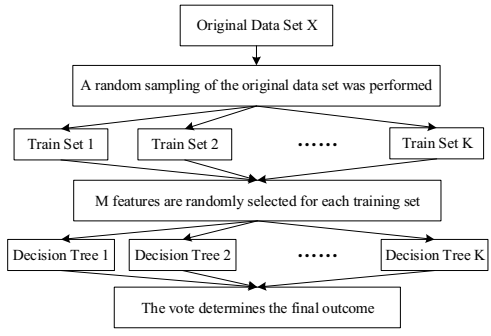

Figure 2. Flow chart of random forest classification algorithm

- Firstly, the Bootstrap idea was used to conduct random sampling with replacement from the original data set, with the sample size of 2N/3 each time;
- m (m<M) features were used as the input of the training decision tree to construct the decision tree;
- Repeat the above two steps K times to produce K decision trees to form a random forest;
- Finally, the classification results of K decision trees are integrated to obtain the final classification results.

Random forests for contain noise and the data with missing values has good prediction accuracy, and can handle a large number of input variables, has faster training speed, in recent years has been widely used in many fields such as classification, feature selection, as high as the research patent value feature selection and screening high value patents important characteristics of the effective methods.

Therefore, this paper proposes a feature selection method for high value patents based on random forest. The influence factor, promotion factor and importance factor of features are obtained through the training of random forest algorithm, and these three factors are used as evaluation indexes to screen out the important features of high value patents. Compared with the single factor method, the feature selection method proposed in this paper has a good feature selection effect, which provides a reference for enriching the feature selection methods of high value patents.

4. Data Source and Method

4.1. Data Source

Wisdom buds patent database with 160 million global patent data, covering 126 countries and regions, with the authority of the richness of patent data and patent evaluation index, the wisdom buds patent database through from the market, economic, legal and technical level four dimensions to measure the value of the patent, and patent value can be divided into five star, The higher the star rating, the greater the value of the patent. The dimension of measuring patent value in Wisdom buds patent database is consistent with the connotation and characteristics of high value patents. Therefore, this paper selects patent data in Wisdom buds patent database as experimental data to verify the effectiveness of feature selection method.

In China, all the patents selected in this paper are from 2016 because it takes at least 18 months from application to publication and the number of patents cited is divided into the number of citations within 3 years and the number of citations within 5 years. Because the probability of patents being cited is consistent, the influence of the time factor on the experimental results is eliminated. A dataset containing 15 features is obtained on the basis of the Wisdom buds patent database combined with the Patent Value Index Guidebook, and the experimental data are trained supervised with the patent value defined in the Wisdom buds database as the label, as shown in Table 1.

Table 1. Feature set of patent data

No.	Characteristics of the name	No.	Characteristics of the name
1	Number of claims	9	Total number of cited patents
2	Cited the number of patents	10	Number of non-patent citations
3	Refer to the number of patents	11	Number of applicants

4	Number of citations in 3 years	12	Number of inventors
5	Number of citations in 5 years	13	Number of litigation cases
6	Number of members of the same clan	14	Legal status
7	Patent type	15	Priority country
8	License type		

4.2. Method

This paper presents a high value of patent feature selection method based on random forest, based on prediction accuracy improved characteristics influence factor and factor, and through the characteristics of its own characteristics of random forest algorithm is importance factor, influence factor, enhance get patent value evaluation factor and the importance factor of the optimal feature subset, I-I-I Factor for short, as shown in Figure 3.

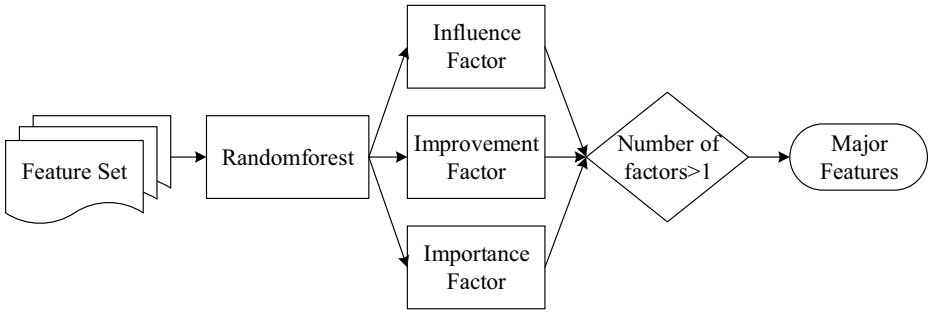

Figure 3. Feature selection of high value patents based on random forest

The basic idea of calculating the Influence Factor of features (Influence Factor) is to define the Influence degree of a feature on the prediction accuracy of random forest model in the case of complete features as the Influence factor of the feature. The higher the Influence degree is, the more important the feature is to the evaluation and prediction; the lower the Influence degree is, the less the feature is related to the evaluation and prediction.

Assuming that the original data set X contains i features, the influence factor d_i of the feature i can be written as follows:

$$d_i = \overline{P} - \overline{A_i} \tag{1}$$

Where $\overline{P}$ represents the prediction accuracy of the random forest model under the complete set of features, $\overline{A_i}$ represents the prediction accuracy of the random forest model after eliminating the i^{th} feature, and the set of feature influence factors is expressed as $D = \{d_1, d_2, \ldots, d_i\}$.

The basic idea of feature improvement factor calculation is to define the improvement degree of prediction accuracy of a certain feature to random forest model in the case of empty set as the improvement factor of this feature, and the higher the improvement degree, the higher the importance of this feature.

The u_i formula of the i^{th} feature is:

$$u_i = \overline{B_i} \tag{2}$$

Where $\overline{B_i}$ represents the prediction accuracy of the random forest model after adding the feature i under the empty set, and the feature improvement factor set is expressed as $U = \{u_1, u_2, \ldots, u_i\}$.

The basic idea Of calculating the importance factor Of features through random forest is to quantify the contribution degree Of each feature to the classification performance Of the constructed decision tree K. The contribution degree is usually expressed by the Important Factor Of features and the Out Of Bag (OOB) error rate is used as the evaluation index.

First define the indicator function:

$$I(x, y) = \begin{cases} 1, & x = y \\ 0, & x \neq y \end{cases} \tag{3}$$

h_{ki}^{OOB} of the i^{th} feature F_i in the k tree in the original data set X is:

$$h_{ki}^{OOB} = \frac{\sum_{p=1}^{n_o^k} I\left(Y_p, Y_p^k\right)}{n_o^k} - \frac{\sum_{p=1}^{n_o^k} I\left(Y_p, Y_{p,\pi_i}^k\right)}{n_o^k} \tag{4}$$

Where, n_o^k is the number of observed samples of the k decision tree. Y_p is the real classification label corresponding to the p sample, Y_p^k is the predicted classification result of the K^{th} decision tree on the p observation of OOB data before random replacement F_i, and Y_{p,π_i}^k is the classification result of the k decision tree on the p sample after random replacement F_i, where the k decision tree needs to be retrained. When feature F_i does not appear in the k decision tree, $h_{ki}^{OOB} = 0$.

The importance factor of feature F_i in the whole random forest is defined as:

$$h_{ki}^{OOB} = \frac{\sum_{k=1}^{K} H_{ki}^{OOB}}{K\sigma} \tag{5}$$

Where K represents the tree of the decision tree in the random forest, σ represents the standard deviation of h_{ki}^{OOB}. The importance factor of feature F_i, h_{ki}^{OOB} represents the contribution of F_i to classification accuracy. The feature importance factor is jointly determined by the mean and standard deviation of the error rate outside

the bag, and the set of feature importance factor is expressed as $H = \left\{ h_{k1}^{OOB}, h_{k2}^{OOB}, \ldots, h_{ki}^{OOB} \right\}$.

The more times a feature appears in different factors, the higher the importance of the feature is. Therefore, an I-I-I Factor feature selection method based on random forest is proposed in this paper, that is, the influence factor, promotion factor and importance factor of features are firstly obtained through the training of random forest algorithm, and then the repeated features in the influence factor, promotion factor and importance factor are selected as the final feature selection result.

5. Experiment Results

The random forest was used to rank the influence factor, promotion factor and importance factor of 15 patent features in terms of their importance. The training was carried out with 10 fold cross-validation. Each group of experiments was repeated for 100 times, and the average of the results of 100 times was calculated. The sklearn toolkit in Python was used to implement the random forest algorithm and trained to obtain the influence, boost and importance factor of the features.

5.1. Influence Factor

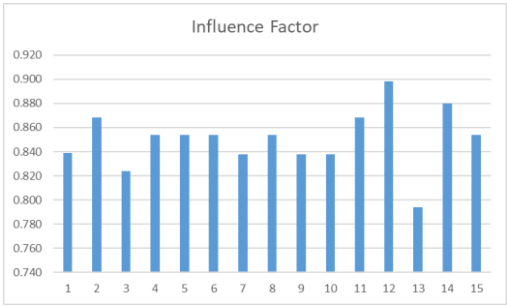

Figure 4. Feature influence factor based on random forest

The feature influence factors based on random forest are shown in Figure 4. It can be found from the figure that " Number of litigation cases " is the feature with the highest impact factor, and " Number of inventors " is the feature with the lowest impact factor. The accuracy rate of the full set features under random forest training is 0.926. The influence factors of features are obtained according to Eq. (1), and the influence factors of features are arranged from high to low in order to obtain Table 2.

Table 2. Feature influence factors based on random forest (from high to low)

No.	Characteristics of the name	Influence factor
1	Number of litigation cases	0.132
2	Refer to the number of patents	0.102
3	Patent type	0.088
4	Total number of cited patents	0.088
5	Cited the number of patents	0.088
6	Number of claims	0.087

7	Number of citations in 3 years	0.072
8	Number of citations in 5 years	0.072
9	Number of members of the same clan	0.072
10	License type	0.072
11	Priority country	0.072
12	Number of patents cited	0.058
13	Number of applicants	0.058
14	Legal status	0.058
15	Number of inventors	0.046

According to the principle of statistics, mode refers to the value with obvious central trend point in statistical distribution, which represents the general level of data, while important features represent features above the general level. Therefore, mode is adopted in this paper as the basis for screening important features through the influence factor. As shown in Table 2, the mode of the feature influence factor is 0.072, and the feature whose influence factor is greater than 0.072 is considered as an important feature. The 6 important features screened by the influence factor in Table 3 are obtained.

Table 3. Important features screened out by influence factor

No.	Characteristics of the name	Influence factor
1	Number of litigation cases	0.132
2	Refer to the number of patents	0.102
3	Patent type	0.088
4	Total number of cited patents	0.088
5	Number of non-patent citations	0.088
6	Number of claims	0.087

5.2. Improvement Factor

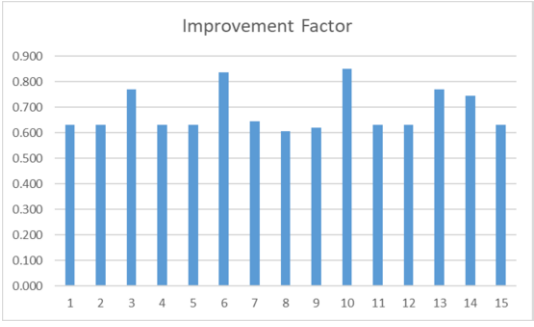

Figure 5. Feature improvement factor based on random forest

The feature improvement factors diagram based on random forest are shown in Figure 5. It can be found that " Number of non-patent citations " is the feature with the highest promotion factor, and " Number of inventors " is the feature with the lowest promotion factor. The mode of accuracy of the full set features under random forest training is 0.926. According to Eq. (2), the feature improvement factor table based on random forest is obtained in Table 4.

Table 4. Feature improvement factors based on random forest (from high to low)

No.	Characteristics of the name	Influence factor
1	Number of non-patent citations	0.850
2	Number of members of the same clan	0.835
3	Refer to the number of patents	0.770
4	Number of litigation cases	0.770
5	Legal status	0.745
6	Patent type	0.745
7	Number of claims	0.655
8	Cited the number of patents	0.645
9	Number of citations in 3 years	0.630
10	Number of citations in 5 years	0.630
11	Number of applicants	0.630
12	Number of inventors	0.630
13	Priority country	0.630
14	Total number of cited patents	0.620
15	License type	0.605

As shown in Table 4, according to the principle of statistics, mode refers to the value with obvious central trend point in statistical distribution, representing the general level of data, while important features represent features above the general level. Therefore, mode is adopted in this paper as the basis for screening important features through promotion factor. As shown in Table 4, the mode of the feature promotion factor is 0.630, and the feature whose promotion factor is greater than 0.630 is considered as an important feature. The important features screened by the promotion factor are obtained as shown in Table 5.

Table 5. Important features screened out by improvement factor

No.	Characteristics of the name	Influence factor
1	Number of non-patent citations	0.850
2	Number of members of the same clan	0.835
3	Refer to the number of patents	0.770
4	Number of litigation cases	0.770
5	Legal status	0.745
6	Patent type	0.745
7	Number of claims	0.655
8	Cited the number of patents	0.645

5.3. Importance Factor

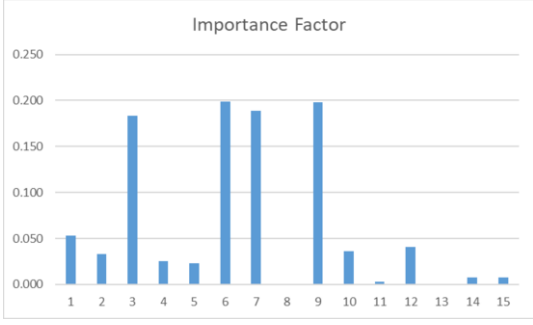

Figure 6. Feature importance factor based on random forest

The feature importance factors diagram based on random forest are shown in Figure 6. Feature importance factors were obtained through random forest algorithm training, and the obtained feature importance factors were arranged from high to low, and the results were shown in Table 6.

Table 6. Feature importance factors based on random forest (from high to low)

No.	Characteristics of the name	Influence factor
1	Number of members of the same clan	0.199
2	Total number of cited patents	0.198
3	Patent type	0.189
4	Refer to the number of patents	0.183
5	Number of claims	0.053
6	Number of inventors	0.041
7	Number of non-patent citations	0.036
8	Cited the number of patents	0.033
9	Number of citations in 3 years	0.025
10	Number of citations in 5 years	0.023
11	Priority country	0.008
12	Legal status	0.008
13	Number of applicants	0.003
14	License type	0.000
15	Number of litigation cases	0.000

Generally, in the feature selection process of random forest, features with an importance factor greater than 0.15 are considered as important features. It can be seen from Table 6 that the importance factors of "Number of members of the same clan", "Total number of cited patents", "Patent type" and "Refer to the number of patents" are greater than 0.15. Therefore, the important features screened by the importance factors are obtained, as shown in Table 7.

Table 7. Important features screened by importance factor

No.	Characteristics of the name	Influence factor
1	Number of members of the same clan	0.199
2	Total number of cited patents	0.198
3	Patent type	0.189
4	Refer to the number of patents	0.183

5.4. I-I-I Factor

Table 8. Statistical frequency of occurrence of important features in different factors

Characteristics of the name	Influence factor	Improvement factor	Importance factor	Occurrences
Refer to the number of patents	√	√	√	3
Number of litigation cases	√	√		2
Patent type	√	√	√	3
Total number of cited patents	√		√	2
Number of non-patent citations	√	√		2
Number of members of the same clan		√	√	2
Number of claims	√			1
Legal status		√		1

Table 8 shows the number of times features appear in different methods. The more times a feature appears in different factors, the higher importance of the feature is. Therefore, the frequency of the occurrence of important features in different factors is counted, and the repeated features in the influence factor, promotion factor and importance factor are selected as the final feature selection result. The important features selected by I-I-I Factor feature selection method are shown in Table 9.

Table 9. Important features after the final feature selection

No.	Characteristics of the name
1	Refer to the number of patents
2	Number of litigation cases
3	Patent type
4	Total number of cited patents
5	Number of non-patent citations
6	Number of members of the same clan

In order to verify the effectiveness of the random forest-based I-I-I Factor feature selection method proposed in this paper, The feature sets before and after I-I-I Factor feature selection were put into four commonly used machine learning classifiers, SVM, LR, CART and GBDT, to test the effectiveness of the I-I-I Factor feature selection method proposed in this paper, and the results are shown in Table 10. It can be seen from the experimental results that the important feature sets filtered by I-I-I Factor have higher classification accuracy in different classifier models, indicating that the random forest-based I-I-I Factor feature selection method proposed in this paper is effective.

Table 10. Model accuracy changes before and after feature selection

Classifier Name	Before Feature Selection	After Feature Selection	Δ Accuracy
SVM	63.8 ± 1.1	84.3 ± 1.4	+20.5 ± 0.3
LR	90.0 ± 0.4	91.2 ± 0.5	+1.2 ± 0.1
CART	91.0 ± 0.6	91.1 ± 0.8	+0.1 ± 0.2
GBDT	88.1 ± 0.4	89.6 ± 0.6	+1.5 ± 0.2

Meanwhile, in order to further verify the effectiveness of the random forest-based I-I-I Factor feature selection method proposed in the text, the sets of features filtered by Influence Factor, Improvement Factor, Importance Factor and I-I-I Factor were brought into SVM, LR, CART and GBDT four commonly used machine learning classifiers to test the effectiveness of the I-I-I Factor feature selection method proposed in this paper, and the results are shown in Table 11. From the experimental results, it can be seen that the set of important features screened by I-I-I Factor performs well under different classifier models, further indicating that the I-I-I Factor feature selection method based on random forest proposed in this paper is effective.

Table 11. Compares with the single factor method

Classifier Name	Influence Factor	Improvement Factor	Importance Factor	I-I-I Factor
SVM	66.3 ± 0.1	74.9 ± 0.2	82.5 ± 0.9	84.0 ± 0.7
LR	89.8 ± 0.3	86.7 ± 0.3	89.7 ± 0.3	91.2 ± 0.5
CART	91.1 ± 0.3	85.2 ± 0.3	92.6 ± 0.4	91.1 ± 0.8
GBDT	86.6 ± 0.8	83.5 ± 1.3	92.6 ± 0.2	89.6 ± 0.6

6. Conclusion

In this paper, a high value patents feature selection method based on random forest is proposed. The idea of the method is mainly reflected in the evaluation of patent features by using a variety of factors, while the traditional feature selection methods mostly use a single factor to evaluate patent features. In this method, the influence factor, promotion factor and importance factor of features are obtained by training the original feature set through random forest, and the I-I-I Factor feature selection method is proposed by integrating the three factors. The connotation of feature selection by I-I-I Factor is that the more times a feature repeats in different factors, it indicates that the higher the importance of this feature is, the important characteristics of high value patents can be screened out. Experimental results show that the feature subset screened out by the random forest-based I-I-I Factor feature selection method in this paper has good performance under various classifiers.

Acknowledgment

This work was sponsored by the National Natural Science Foundation of China (Grant No. 71804157).

References

[1] N. Kozodoi, S. Lessmann, K. Papakonstantinou, Y. Gatsoulis, and B. Baesens, "A multi-objective approach for profit-driven feature selection in credit scoring," Decision Support Systems, vol. 120, pp. 106-117, 2019.
[2] B. Mla, C. Mm, C. Rc, "Mixed integer linear programming for feature selection in support vector machine," Discrete Applied Mathematics, vol. 261, pp. 276-304, 2019.
[3] A. Jayaprakash, C. Keziselvavijila, "Feature selection using Ant Colony Optimization (ACO) and Road Sign Detection and Recognition (RSDR) system," Cognitive Systems Research, vol. 58, pp. 123-133, 2019.
[4] D. Harhoff, F. Narin, F. M. Scherer, K. Vopel, "Citation Frequency and the Value of Patented Inventions," Review of Economics and Statistics, vol. 81, pp. 511-515, 1999.
[5] L. Jiang, P. Willett, "ArticleRank : A PageRank-based Alternative to Numbers of Citations for Analyzing Citation Networks," Aslib Proceedings, vol. 61, pp. 605-618, 2009.
[6] R. Bekkers, R. Bongard, A. Nuvolari, "An Empirical Study on the Determinants of Essential Patent Claims in Compatibility Standards," Social Science Electronic Publishing, vol. 40, pp. 1001-1015, 2011.
[7] Ping Xie, Run Yuan, Guo Qian, "Research on core patent identification based on TOPSIS method," Information Studies : Theory & Application, vol. 38, pp. 88-92, 2015.
[8] Run Yuan, Guo Qian, "A Rough Set Theory Model for Identifying Core Patents," Library and Information Service, vol. 59, pp. 123-130, 2015.
[9] H. Noh, Y. K. Song, S. Lee, "Identifying emerging core technologies for the future: Case study of patents published by leading telecommunication organizations," Telecommunications Policy, vol. 40, pp. 956-970, 2016.
[10] C. Kim, H. Lee, H. Seol, C. Lee, "Identifying core technologies based on technological cross-impacts: An association rule mining (ARM) and analytic network process (ANP) approach," Expert Systems with Applications, vol. 38, pp. 12559-12564, 2011.
[11] Lee, Changyong, Kwon, Ohjin, Kim, Myeongjung, et al, "Early identification of emerging technologies: A machine learning approach using multiple patent indicators," Technological forecasting and social change, vol. 50, pp. 291-303, 2018.
[12] Y. Qiu, C. Zhang, S. Chen, "Research on Patent Value Evaluation Index System Based on Classification Regression Tree Algorithm," Journal of Xiamen University(Natural Science), vol. 56, pp. 244-251, 2017.
[13] L. Bai, Z. Zhu, L. Liu, P. Xia, "Research on the intelligent identification method of potential high value patents in the Internet field," China Invention & Patent, vol. 15, pp. 28-32, 2018.