Journal of Pharmaceutical Innovation (2019) 14:28–34
https://doi.org/10.1007/s12247-018-9332-2

ORIGINAL ARTICLE

Intelligent Classifier: a Tool to Impel Drug Technology Transfer from Academia to Industry

Hui-Heng Lin[1] · Defang Ouyang[1] · Yuanjia Hu[1,2]

Published online: 2 June 2018

Abstract

Purpose Pharmaceutical technology transfer is one of the components of pharmaceutical innovation. Currently, a gap exists in pharmaceutical technology transfer from academia to industry. This study aims to develop an objective model to identify valuable pharmaceutical technologies for transferring in order to drive pharmaceutical innovation.

Methods We created a support vector machine classifier model using the data of pharmaceutical patents held by universities to predict the licensing outcomes of those patents. We collected data on 369 United States (US) pharmaceutical patents, using 142 licensed patents as the positive samples and 227 unlicensed patents as the negative samples. We also collected the licensing data of the patents, and the distinguished patent features were selected for model training and generation. Upon optimization, the machine learning model was evaluated using different scoring methods.

Results Our support vector machine-based model achieved a fairly good performance of 82.50% in precision and 88.89% in specificity.

Conclusions To the best of our knowledge, our study is the first to apply the machine learning approach to predict the licensing outcomes for pharmaceutical patent valuation and technology transfer. Our work is a good alternative to the current patent valuation methods available in the market, and it could be further developed for practical use in real business contexts.

Keywords University patents · Pharmaceutical patents · Technology transfer · Patent licensing · Machine learning prediction · Support vector machine

Introduction

In pharmaceutical technology transfer, which is one of the components of pharmaceutical innovation, a gap exists between academia and industry [1, 2]. On the one hand, scientists in academia need financial support to develop or commercialize their early-stage drug discoveries [3], such as candidate drugs in the discovery stage or drugs in the preclinical testing stage. On the other hand, investors believe it is difficult to identify the right projects worth investing in [4, 5]. This raises a crucial question about how to efficiently identify pharmaceutical technologies worth investing or transferring in the early stages. An objective approach for determining potentially valuable pharmaceutical technologies is needed to narrow the gap between technology transfer stakeholders and further facilitate the technology flow from academia to industry.

To address this issue, we considered pharmaceutical patents as indicators of technologies, and we proposed the application of machine learning to analyze pharmaceutical patent data. In this study, the support vector machine (SVM) was chosen from a variety of machine learning techniques [6]. SVM is one of the most popular machine learning algorithms. Although it has several technical limitations, such as a high dependency on the quality of the input data, it is still very popular due to the excellent performance it displays in binary classification tasks. For example, Zhang et al. used SVM to

✉ Yuanjia Hu
yuanjiahu@umac.mo

Hui-Heng Lin
yb57542@umac.mo

Defang Ouyang
defangouyang@umac.mo

[1] State Key Laboratory of Quality Research in Chinese Medicine, Institute of Chinese Medical Sciences, University of Macau, Macau, China

[2] The Research Center of National Drug Policy and Ecosystem, Nanjing, China

predict whether pre-microRNA sequences are the true precursors of microRNA or the pseudo precursors [7]. Similarly, SVM could also be used to predict if patents would be licensed out because the prediction of patent licensing outcome could be viewed as a binary classification task. This present study focused on the pharmaceutical patents held by universities. We analyzed the relevant patents and licensing data, and we selected the possible factors and features that might play an important role in the patent licensing outcomes. Subsequently, feature vector space was constructed and used to train the SVM classifier to generate a predictive model. Afterwards, the model was further optimized and its performance was evaluated.

To the best of our knowledge, our work is the first innovative and pioneering study that has applied the machine learning approach to predict licensing outcomes for pharmaceutical technology transfer. Our work could be further developed and extended for practical use in real business contexts.

Material and Methods

Generally, for patent valuation in this study, we considered successfully licensed patents to be valuable patents. The workflow of the in silico analysis began with the retrieval of the patent data from the IMS Lifecycle database. Next, the patent features were selected and retrieved from the database of the United States Patent and Trademark Office (USPTO). After data cleaning, data reorganization, and labelling of the positive/negative training dataset, the structuralized data were converted to the vector space to represent each sample patent. Subsequently, the vector data were stored in plain text format (comma separated values) and transferred into the Linux operating system. In the Linux operating environment, we loaded the SVM-relevant modules into the Python programming interface and carried out the functions of data scaling and normalization, parameter tuning and cross-validation, model training, prediction of testing the dataset, and model evaluation (Fig. 1). Through these processes, we obtained the optimized SVM predictive model. More details on these methods are described below. A series of in silico analytic tasks were executed in the Python programming environment under the Linux system, and the following modules were involved: numpy [8], scipy [9], gnuplot [10], scikit-learn [11], and libsvm [12].

Data Collection Criteria

In general, we selected the US pharmaceutical patents held by universities as the source data for this study. We focused on pharmaceutical patents from universities for several reasons. First and foremost, universities need to transfer or license out technologies, but they face difficulties in doing so, such as the valley of death [2]. Nowadays, many universities produce a significant number of scientific research outputs in addition to their conventional role of teaching. In many cases, such as in the biomedical or pharmaceutical research community, technological innovations can be the fruition of basic research. Hence, universities represent one of the productive sources of scientific and technological innovations, and they need to transfer their technologies or license out their patents [3, 13]. For instance, if university researchers have discovered potential new drugs, due to the universities' limitations in resources and expertise for clinical drug development, universities might patent their discoveries and want to license out their patents for further development of these new drugs. In the present study, we selected pharmaceutical patents because patents carry more weight in the technology-oriented industries, such as in the pharmaceutical and information technology industries. Within these industries, companies' core competencies rely more heavily on state-of-art technological inventions and the protection of intellectual properties [14]. Without patents, their products are very likely to be copied or illegally used by others in the highly competitive business world. Last, but not the least, we further checked and collected the licensing data of patents because we assumed that patents that succeeded in being licensed out would be valuable [5]. In summary, based on these reasons, we selected and collected data to study the licensing of university pharmaceutical patents.

Data Retrieval and Preprocessing

The primary data source for this study is a commercial database IMS Lifecycle. IMS Lifecycle is a powerful tool to analyze the research and development (R&D) pipelines from drug discovery phase to marketed phase, covering more than 23,300 drugs in R&D projects since 1977. Patents and various data relevant to pharmaceutical R&D are recorded in IMS Lifecycle with a strict criteria of data collection. Subsequently, we queried the IMS Lifecycle database (as of February 1, 2017), and we then identified the drugs with patent data. Specifically for this study, we only studied the patents whose patent licensor is either a university or a college. The query results that did not provide the desired patent information were removed. Lastly, we identified a total of 369 US patents. Of the 369 patents, 142 had licensing records; hence, they were used as the positive samples for training the classifier. The other 227 patents without licensing records were treated as the negative samples for training the classifier.

From the IMS Lifecycle database, only limited patent data were provided; however, we were able to obtain the US patent number, the licensing data, and the therapeutic indications. For other patent data and features, as described in the feature representation section, we used the US patent number to retrieve the data from the USPTO database. To prepare the

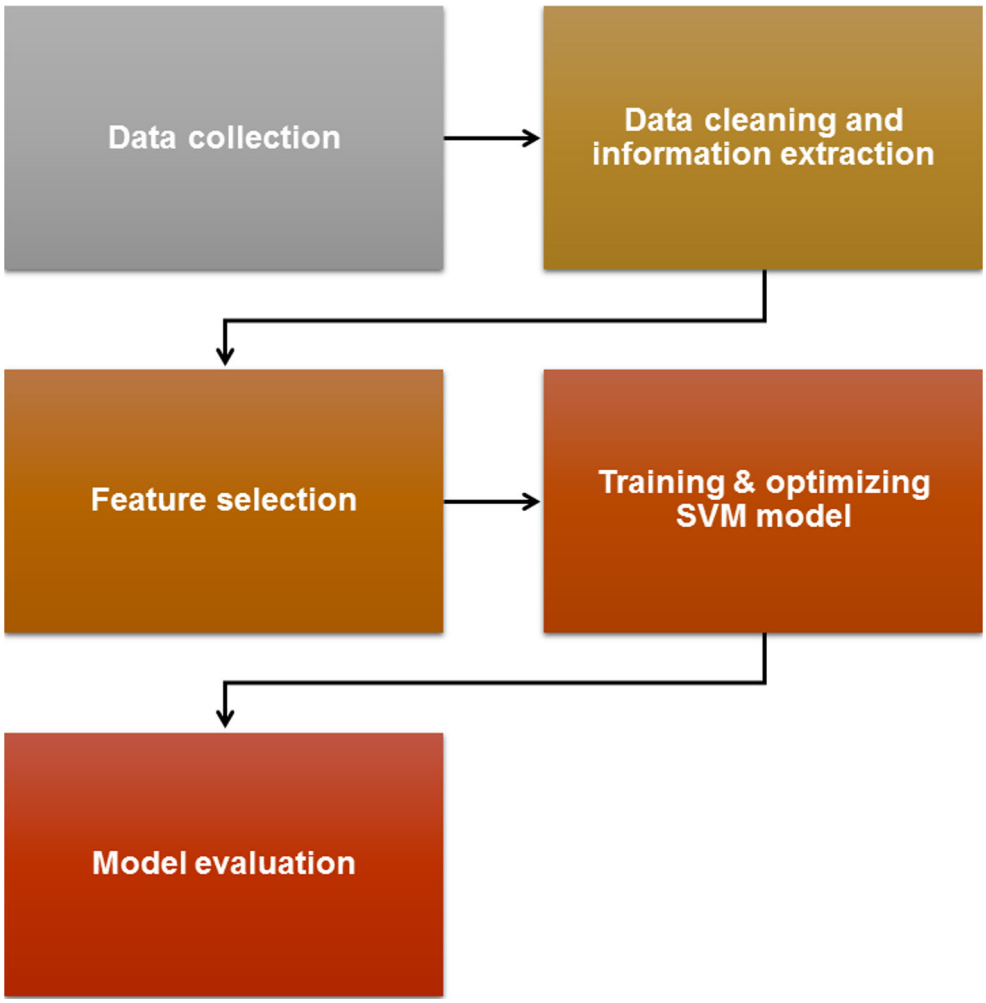

Fig. 1 SVM-based machine learning analytic workflow for this work

datasets for the SVM classifier training, we divided the 369 samples into a primary training set containing 246 samples and a testing set containing 123 samples, thereby creating a 2:1 ratio for the training set to the testing set. To ensure that the distribution of the positive and negative samples was balanced in both the training and testing sets, we confirmed the 1:2 ratio of positive samples to the negative samples in both datasets, i.e., 82 positive and 164 negative samples in the training set and 41 positive samples and 82 negative samples in the testing set. Consequently, there were 123 positive samples in the training set and the testing set and 246 negative samples in the training set and the testing set. Because the primary dataset contained 142 positive (licensed) and 227 negative (unlicensed) samples, the 123 positive samples used in the training and testing sets are a subset of the 142 positive samples in the primary dataset. A total of 246 negative samples were used in the training and testing sets, which is greater than the 227 negative samples in the primary dataset. Therefore, randomized oversampling and undersampling were performed in order to balance the datasets.

Feature Representation

We selected seven features of the patents for training the SVM classifier, given that the patents studied in this work are pharmaceutical patents from universities. Briefly, we considered three relevant aspects of the patents that were studied: the characteristics of the licensor/patentee (universities or colleges), the characteristics of the patent data, and the characteristics of the relevant drug products. Table 1 provides a brief summary of the basic characteristics of the seven selected features. For instance, the first feature is the research prestige of the licensor. We supposed that, usually, a prestigious university is more likely to attract potential licensees due to its well-established fame and reputation, although other universities with strong technological competency could also be attractive to potential licensees. A possible quantitative indicator for this matter is the average number of nonself forward citations of its patent stock until the patent is granted to the licensee. The nonself forward citations of a patent are defined as the forward citations, excluding the forward citations from

Table 1 Summary of the characteristics of the seven features selected for predicting patent licensing outcomes

Aspects	Factors/features	Potential interpretation	Quantification
Licensor (universities or colleges)	Research prestige [15]	The licensor's well-established fame or reputation for research expertise increases the possibility of licensing out the patent.	Average number of nonself forward citations for a patent stock before sample patent granted to a licensor
Patent data	Technological scope [16]	A broader technological scope of a patent might suggest its higher potential value.	Number of the US patent classes of a patent
	Patent age [5]	A longer period in which the patent is granted results in a longer period of legal protection.	Year the patent was granted minus the year in which the patent application was submitted
	Patent recognition [17]	The extent of international protection for a patent	The number of international patent offices granting the patent
	Patent citation index [18]	The impacts or influences of a patent on other patents	The ratio of the nonself forward citations to the forward citations
	Patent claims [19]	The intellectual property rights reserved by a patent	The number of claims in a patent
Therapy	Therapeutic indications [1]	We assume that patents with more therapeutic indications are of greater interest to licensees, e.g., to license the patent for drug repositioning.	The number of therapeutic indications of a pharmaceutical patent

patents held by the same patent holder. For example, if university *S* holds three patents, *A*, *B*, and *C*, respectively, and if *B* and *C* cited *A*, the nonself forward citations of *A* are the forward citations excluding the forward citations from *B* and *C* because the three patents share the same patent holder, university *S*. In addition to a series of features related to the patent itself and the technological insights, the therapeutic indicator is another interesting feature to mention. Because our study focused on pharmaceutical patents, we naturally considered the therapeutic indication numbers of the relevant drugs, understanding that pharmaceutical patents with multiple therapeutic indications might be sought for further development activities, e.g., drug repositioning.

The SVM Classifier

We choose SVM as the classifier in the present study [6]. SVM is a popular and powerful supervised learning method that is frequently applied in various scientific domains, such as social economics [20], information technologies [21], and computational biology [7, 22, 23]. With a set of vectors and the corresponding known labels, in the binary classification context, the SVM generates a hyperplane to separate the high-dimensional training data into two groups: the positive class and the negative class (Fig. 2). Mathematically, the vectors are defined as *Xi*, where $i = 1, 2, 3, 4 \ldots n$, and the labels are treated as Yi. Yi = 1 or − 1. Then, there exists a hyperplane that not only can correctly separate the data but also has the largest distance (the margin) to the nearest data points. The hyperplane could be expressed in the equation below:

$$g(Xi) = w^{\mathrm{T}} \times Xi + b$$

where w^{T} is for the set of each *Xi*'s weight and *b* is a constant that is known as bias. If *Xi* is given, *w* and *b* determine the hyperplane, together, and there is a specific value for *b* and *w* that can form the best hyperplane, i.e., the correct classifier model with the largest margin.

Model Generation and Evaluation

As mentioned above, 369 labelled sample patents were divided into 246 samples for the training set and 123 samples for the testing set. After primary training trails, in order to improve the performance of the generated model, fivefold cross-validations were carried out to tune the best parameters. Using the optimized model, we classified the 123 labelled testing samples.

Testing the model that was generated by 246 labelled training samples with the 123 labelled testing samples gives the basic performance of the model, i.e., the number of true positive, true negative, false positive, and false negative samples. This enabled us to further evaluate the model using a variety of indices. The following indices, sensitivity (recall), specificity, accuracy, error rate reduction, precision, and F1-measure, were used to evaluate the performance of the generated model [24].

Sensitivity (recall) reflects the ratio of the correctly identified positive samples. The value of sensitivity is defined by the following:

$$\text{Sensitivity (recall)} = \frac{\text{True positive}}{\text{True positive} \ + \ \text{false negative}}$$

The specificity index indicates the ratio of the correctly identified negative samples, and it is defined by the following:

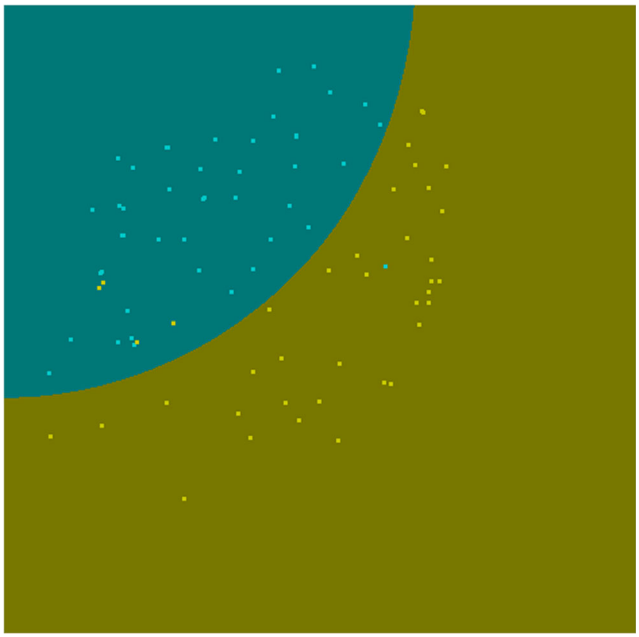

Fig. 2 A simulated result for simple interpretation of the working principle of the SVM classifier

$$\text{Specificity} = \frac{\text{True negative}}{\text{True negatitive} \;+\; \text{false positive}}$$

The accuracy indicator is defined by the ratio of the sum of the number of true positive samples and true negative samples for all instances; hence, it reflects the model's ability to obtain a correct classification. It is defined by the following:

$$\text{Accuracy} = \frac{\text{True positive} + \text{true negative}}{\text{Total number of instance}}$$

The precision index measures the model's performance to correctly classify the positive samples within all of the samples classified as positive by the model. It is defined by the following:

$$\text{Precision} = \frac{\text{True positive}}{\text{True positive} + \text{false positive}}$$

F1-measure is the harmonic average of precision and recall. The F1-measure ranges between 0 (the worst) and the 1 (the best, i.e., perfect precision and recall), and it is defined by the following:

$$\text{F1}-\text{measure} = \frac{2 \times \text{recall} \times \text{precision}}{\text{Recall} + \text{precision}}$$

Data Availability

The data of this study are available for research purposes and one can access the data by sending reasonable request to the corresponding author.

Results

After optimization, the generated model was tested to verify whether it could successfully be used to predict the outcome of patent licensing. The results are described below.

Performance of the Optimized Model

Through optimizations, e.g., cross-validation and grid search, the average accuracy in fivefold cross-validation was improved from 61.25 to 74.79%. Accordingly, we computed the number of true positive, true negative, false positive, and false negative samples classified by the model, as shown in Table 2. The true positive and true negative numbers are the number of positive and negative samples correctly identified by the SVM classifier, respectively. The number of false negative and false positive samples indicates the number of negative and positive samples wrongly classified by the classifier.

Evaluation of the Model Performance

Based on the number of true positive, true negative, false negative, and false positive samples that were obtained, a series of indices were further computed to evaluate the performance of the model. The principles of the indices were described in the "Material and Methods" section, and the values of the indices are shown in Table 3. Sensitivity (recall) represents the ratio of the correctly identified positive instances. In our case, our model only correctly predicted 55.00% of all the licensed out patents. The precision value indicates that 82.50% of all the licensed out patents that the model predicted are correct, i.e., 82.50% are truly licensed out patents. The F1-measure is the harmonic average of precision and recall, and its value of 66.00% provides an overall rating for our model. The specificity measures our model's ability to correctly identify the unlicensed patents. Our model was found to have an advantage in identifying the unlicensed patents by achieving a good specificity value of 88.89%. Finally, the accuracy indicates the ratio of the correctly identified sample number to the total sample number. Our model scored a good accuracy of 72.35%.

Discussion

The present work studied the pharmaceutical patents held by universities. On the one hand, many universities establish technology transfer offices to promote technology transfer; for example, by patenting the inventions created by researchers affiliated with the universities. On the other hand, some argue that patenting the scientific knowledge generated by a university could restrict public access to that knowledge, which contradicts the core missions of universities [25]. We

Table 2 The number of true positive, true negative, false positive, and false negative samples reflect the performance of the SVM classifier

Classification	True	False	Total
Positive	33	7	40
Negative	56	27	83
Total	89	34	

do not debate this issue because it is not a topic that is closely related to this work. However in general, we see the tendency that technology transfer in universities will occur more frequently, especially in state-of-art research fronts (e.g., in the area of artificial intelligence, Google invited Dr. Hinton from the University of Toronto to collaborate on R&D projects, and Google also invited Dr. Martinis from the University of California Santa Barbara to collaborate on building quantum computers).

This study has several limitations. First, in the real world, the values of patents might not be viewed as simply positive or negative, as in how we tagged or labelled the sample dataset. Due to the difficulty in quantifying the values of patents, we chose to simplify this issue by using a binary classification approach. Second, our predictions could be better supported by a real-world, data-driven profitability analysis, such as the discounted cash flow analysis, if such an analysis is available. However, it is difficult to conduct a discounted cash flow analysis because it is complex and requires a lot of extra information, such as the data on patent holders/licensees, the financial status of companies, their business plans, and their intellectual property strategies. Because these data are not readily accessible, we were unable to conduct this type of analysis.

Third, this study had several technical limitations. One is the selection and number of the patent features. We selected the patent features by reading the current literature. Under ideal conditions, a significant number of patent features should be available, and features highly related to the licensing outcome should be selected through quantitative analysis, such as automated feature selection technologies. The result generated under such ideal conditions should be less biased than the outcomes we obtained from our manual selection of the patent features. However, due to the limitations of the data sources, we were unable to ensure that our analytic workflow was ideal. Next, when counting the number of patent claims, using the number of independent claims would be better than the total number of claims since the dependent claims depend on the validity of the independent claims.

Table 3 A series of indices to evaluate the model performance

Index	Value (%)
Sensitivity (recall)	55.00
Precision	82.50
F1-measure	66.00
Specificity	88.89
Accuracy	72.35

Furthermore, in this study, the generation of the predictive model relied on the SVM. One of the advantages of the SVM approach is that it is possible to use binary classification and moderate-sized datasets; however, its performance is limited for multi-class classification and processing large-scale datasets. The binary classification function of SVM fits the purpose of this study because predicting if patents could be licensed out is a type of binary classification task. However, it is difficult to extend SVM to other research purposes, such as classifying patents into different types or classes of technologies, which is a multi-class classification task. In that context, the possible solution is to either construct multiple SVM classifiers or utilize other kinds of machine learning algorithms, e.g., the random forest algorithm. Alternatively, to overcome the SVM issue of high computation costs and low computation efficiency when loading large amounts of datasets, deep learning or other techniques might be used. Additionally, in terms of the model's performance, e.g., the sensitivity (recall) index and the F1-measure in the evaluation process, the values for our model were not as high as we expected. One of the potential solutions for this issue might be to conduct further testing on a greater number of feature sets and identify the more suitable feature subsets to improve the performance of the machine learning model.

In this study, we collected the pharmaceutical patents of universities. After selecting the relevant factors, we generated an SVM-based machine learning model to predict the licensing outcomes of the relevant patents. After optimizing the model and evaluating its performance, our model achieved a high performance for specificity (88.89%), precision (82.50%), and accuracy (72.35%). Using actual patent data, we innovatively conducted the first study to predict patent licensing outcomes using the powerful and solid SVM-based machine learning approach. Our work presents a relatively objective approach, and it could be further developed for practical use in a real business context, e.g., both the potential patent buy side and sell side (or patent licensors and licensees). Moreover, even third-party agents can use our tool if they need to predict the potential value of patented pharmaceutical technology. For example, the universities and scientists that desire to license out patents might want to use our tool to predict the licensing outcome of the pharmaceutical patents they own. Meanwhile, the potential licensees or investors might also want to use our tool to estimate the value of pharmaceutical patents by predicting the licensing outcome. The predictive results could assist decision-makers in their licensing and pricing activities. Furthermore, our work provides insights about the patent valuation framework or criteria. For instance, the

features employed in the SVM training could be potential factors for valuating pharmaceutical patents.

Funding This study was funded by the grant MYRG2015-00145-ICMS-QRCM from the University of Macau.

Compliance with Ethical Standards

Conflict of Interest The authors declare that they have no conflict of interest.

Research Involving Human Participants or Animals This article does not contain any studies with human participants or animals performed by any of the authors.

Informed Consent Not applicable.

References

1. Ni J, Shao R, Ung COL, Wang Y, Hu Y, Cai Y. Valuation of pharmaceutical patents: a comprehensive analytical framework based on technological, commercial, and legal factors. J Pharm Innov. 2015;10(3):281–5.
2. Butler D. Translational research: crossing the valley of death. Nature News. 2008;453(7197):840–2.
3. Henderson R, Jaffe AB, Trajtenberg M. Universities as a source of commercial technology: a detailed analysis of university patenting, 1965–1988. Rev Econ Stat. 1988;80(1):119–27.
4. Fernandez JM, Stein RM, Lo AW. Commercializing biomedical research through securitization techniques. Nat Biotechnol. 2012;30:964–75.
5. Ruckman K, McCarthy I. Why do some patents get licensed while others do not? Ind Corp Change. 2017;26(4):667–88.
6. Vapnik VN. An overview of statistical learning theory. IEEE Trans Neural Netw. 1999;10(5):988–99.
7. Zhang Y, Yang Y, Zhang H, Jiang X, Xu B, Xue Y, et al. Prediction of novel pre-microRNAs with high accuracy through boosting and SVM. Bioinformatics. 2011;27(10):1436–7.
8. Walt SVD, Colbert SC, Varoquaux G. The NumPy array: a structure for efficient numerical computation. Comput Sci Eng. 2011;13(2):22–30.
9. Oliphant T, SciPy TE. Open source scientific tools for Python. Comput Sci Eng. 2007;9:10–20.
10. Racine J. Gnuplot 4.0: a portable interactive plotting utility. J Appl Econ. 2006;21(1):133–41.
11. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
12. Chang CC, Lin CJ. LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol. 2011;2(3):27.
13. Cervantes M. Academic patenting: how universities and public research organizations are using their intellectual property to boost research and spur innovative start-ups. WIPO Small And Medium-sized Enterprises Documents 2003. http://www.wipo.int/sme/en/documents/academic_patenting.html. Accessed 1 April 2018.
14. Salazar A, Hackney R, Howells J. The strategic impact of internet technology in biotechnology and pharmaceutical firms: insights from a knowledge management perspective. Info Technol Manag. 2003;4(2):289–301.
15. Laursen K, Leone MI, Torrisi S. Technological exploration through licensing: new insights from the licensee's point of view. Ind Corp Change. 2010;19(3):871–97.
16. Leone MI, Reichstein T. Licensing-in fosters rapid invention! The effect of the grant-back clause and technological unfamiliarity. Strat Mgmt J. 2012;33(8):965–85.
17. Neuhäusler P, Frietsch R. Patent families as macro level patent value indicators: applying weights to account for market differences. Scientometrics. 2013;96(1):27–49.
18. Smith DKW. A new methodology for citation dependent patent evaluations. Carleton University. 2014. (Electronic, M.Sc. thesis) http://curve.carleton.ca/system/files/theses/31557.pdf. Accessed 1 Oct 2017.
19. Sakakibara M. An empirical analysis of pricing in patent licensing contracts. Ind Corp Chang. 2010;19(3):927–45.
20. Pai PF, Hong WC, Change PT, Chen CT. The application of support vector machines to forecast tourist arrivals in Barbados: an empirical study. Int J Manag. 2006;23(2):375.
21. Tseng CY, Chen MS. Incremental SVM model for spam detection on dynamic email social networks. IEEE CSE Int Conf. 2009;4:128–35.
22. Hua S, Sun Z. Support vector machine approach for protein subcellular localization prediction. Bioinformatics. 2001;17(8):721–8.
23. Yu CS, Lin CJ, Hwang JK. Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions. Protein Sci. 2004;13(5):1402–6.
24. Powers DM. Evaluation: from precision, recall and F-measure to ROC, 2011 informedness, markedness and correlation. J Mach Learn Tech. 2011;2(1):37–63.
25. Campos TC. The idea of patents vs. the idea of university. New Bioethnol. 2015;21(2):164–76.