International Journal of Multiphysics
Volume 18, No. 3, 2024
ISSN: 1750-9548

Automatic Classification of Intellectual Property Legal Cases based on Support Vector Machine

Geyu Sheng[1], Jun Zhang[2*]
1 School of Liberal Arts and Law,Henan Polytechnic University, Jiaozuo,Henan, 454000, China
2 Center of Information Construction and Management, Henan Polytechnic University, Jiaozuo,Henan,454000, China
*Corresponding author e-mail: zhangjun@hpu.edu.cn

Abstract:
The importance of intellectual property in economic development is growing. Support vector machines have produced cutting-edge outcomes in a variety of applications, including document classification. However, existing study used SVM for the IP classification task but it did not produce as excellent results as alternative learning algorithms like Random Forest and KNN. This is because kernel patent classification differs from traditional classification in many ways. We assess the new methods by classifying the international patent collection of documents using the Gaussian Kernel Support Vector Machine (GKSVM). This study looks at how to recognize specific elements in court decision texts automatically and evaluates how important a role they play. In this paper, we used common classifiers to classify patent documents. The proposed classification method, GKSVM, yields the best results, and the evaluation result shows accuracy for the test set sample.

Keywords: automatic classification, Intellectual property, Support Vector Machine, GKSVM, legal case, patent.

1. Introduction

Intellectual property is a significant element that is held being rightfully protected by a business or individual against unapproved use by third parties. Examples of this type of asset include charters, patents, logos, and trade confidences [1]. Researchers from all around the world claim that intellectual property endorses economics, generates employment, maximises social efficacy, and is important to the modern economy [2]. Related businesses are also growing quickly and have a sizable market. IP valuation is the first step towards realising an insubstantial asset's greatest potential. IP assessment provides a widely comprehensible monetary basis for the involvement of intellectual property to a corporation. The Contemporary Earning Worth Technique, Market Comparative Method, and Cost Approach are the 3 main conventional approaches for appreciating intellectual property. Unfortunately, conventional IP estimate tactics are expensive, take quite a while to value, and are challenging to utilise because of the exclusive nature of intellectual property (IP) and the lack of comprehensive legislation surrounding it. It is significant to note that a quick, precise, and impartial appraisals using machine knowledge techniques that boost the fundamental worth of intellectual property [3].

Support Vector Machines (SVMs) are supervised learning representations in appliance learning that have corresponding learning systems that are capable of statistics analysis [4]. Using sustenance vector machineries, high accuracy patent categorization systems can be created. To reduce the data dimension, you employ self organizing maps (SOMs). Neural networks include maps that organise themselves. The L2-norm distance, among other distance measurements, is used to repeatedly sort the facts based on regular forms and commonalities inside the dataset. This allows us to create distinct data groupings based on their quality. Neural Network is used as the regression model [5]. A neuronal network is an interconnected system or circuitry of biological nerve cells, or, in the modern sense, a neural

network that is artificially made up of synthetic neurons or nodes. Neuronal networks can be utilised in a variety of sectors; however, in this system, they are primarily used for analysis of regression. Because of the growing constantly quantity of patents, the breadth of technological domains covered, and the inherent complexities of patent papers, computerised dispensation and categorization is essential. Machine learning procedures have been effectively applied to text categorization as well as data retrieval [6]. Patent processing of data, which is a subdivision of text processing, can benefit from machine learning, particularly patent categorisation and extraction.

This paper employs a cutting-edge machine knowledge approach, known as the Support Vector Machine. The patent documentation includes numerous items for investigation. These things are categorised into two categories: organised and unorganised. Patent numbers, filing dates, and beneficiaries are examples of organised group elements, while unorganised information is provided in a variety of textual content of varying lengths and material, such as asserts, abstract concepts, titles, and summaries. Some patents papers include patent diagrams which are visual representations of data that is structured as well as unstructured.

This article focused on patented document categorization and efficiency using several classifiers. The study found features by assigning scores to each keyword in a patent application using various weighting algorithms. The feature matrix is then fed into a classifier, and the precision of the classification is observed.

2. Literature Review

In 1997, a researcher stated that patent citation analysis gives information regarding primary referenced patents, influence index, and technological competence [7]. In a study, it was highlighted how patents used mathematical knowledge as a means of innovation. In 2003, another author proposed that the application of knowledge can be utilised to track a firm's technical advancement and diversity [8]. Furthermore, it was found that the used patent records as a measure of a country's technological speciality. The researcher developed a patent grouping system for fundamental analysis of technology and tested two simplistic Bayes algorithms with varying vocabulary lengths [10]. With the progress of computer technological advances, computerised sorting of patent documents can be useful. Technology such as computers can provide automated or partially automatic categorization aid, reducing the ambiguity and inaccuracies associated with traditional categorisation [11]. At the exact same period of time, it can lower the examiner’s burden while increasing classification effectiveness. However, according to the present literature assessment, relevant research remains in the exploratory phase. Some scholars choose to analyse and classify patents using their abstracts or parts [12].

The two basic parts of it are the deployment of automated learning techniques and data preparation. Additionally, this article will conduct research using various machine learning techniques and patent components. According to some academics, the SVM technique performs best when it comes to automatically classifying patents [13].

According to several researchers, in numerous deep learning competitions, the XGBoost approach yields the most sophisticated results [14]. Researchers most frequently use decision trees and random forests as methods for data classification [15]. Consequently, this thesis will employ decision trees, XGBoost, SVM, and random forests as machine learning algorithms. However, other academics pointed out that the assertions could serve as the input information for the categorization of patents. The researchers stated that the claims component meets the requirements for patent classification [16].

Furthermore, several researchers stated that the description section frequently contains detailed information about a single invention that may be utilised for patent categorisation [17]. This is comparable to other studies, which offer a general patent analysis efficiency of operation, with the exception that each analysis performed has a particular purpose. According to the author, this approach is complimentary to the invention cycle, and information about intellectual property assessment has several applications in a variety of industries [18]. The researcher connects the patent lifespan to copyright-related sources of knowledge and different duties along the analysis of patents process [19]. They argue that their patent statistics process is a motivated by purpose procedure that includes pursuit tasks, evaluation duties (micro and macro assessments of business value, technical assessments, and technology recommendations), along with tracking tasks. In a comparable manner the authors suggest that patent examination

constitutes a type of patent information that aids the decision-making process They claim that the word "patent assessment" has a double significance: the process of considering all of the foregoing and the actual study of the patent material [20]. They practice the findings to identify three patent assessment tasks: patent investigating, patent assessment, and patent surveillance, and link the value that data provides from these to the free innovation channel [21].

Undoubtedly, there are still no satisfactory results from accuracy, and there is still no widespread automated cataloging of patent credentials. Therefore, the present research on the use of mechanism knowledge to automatically categorise patent texts is important in terms of its practical position. This research can consequently split a great deal of patent texts conferring to the conceptual characteristics of those patent documents, which may assist more people understand the rich technical knowledge. The goal of this article is to automatically identify patent using machine knowledge as well as text analysis approaches.

3. Methodology

3.1 Support vector machine (SVM)

Support vector machines were created especially for classification into two classes [22]. By utilizing the largest distance between two class vectors, this approach seeks to create an ideal hyper plane as a decision function. Support vector machines require input the feature vector on the high dimensional feature space using non-linear mapping. As the method's initial application, a maximum effectiveness based on decision plane is created to separate the accurate data [23]. "Margin" refers to the separation between the nearest data points on either side of the hyperplane. The effectiveness of classification on every side of the plane increases with increasing margin. This paper discusses automatic classification of intellectual property legal cases and provides an explanation of support vector machines [24].

Gradient Boosting

The boosting algorithm known as gradient boosting operates on the basis of the phase method as described, in which a strong learner algorithm is created as a final model by adding several weak learning algorithms that have all been trained on the same dataset.
The following situations are suitable for the use of gradient boosting:
Regression involves averaging the results produced by the less proficient students.
Classification determining which class prediction appears the most frequently
Because XGBoost and LightGBM are becoming more and more popular, we will examine them both from a theoretical and practical perspective in order to better understand their benefits and drawbacks.

3.2 XGBoosting (XGB)

The full name of XGBoost is eXtreme Gradient Boosting, proposed by Dr. Tianqi Chen who worked in the University of Washington in 2014. XGBoost is a tree integration model, which uses the cumulative sum of the predicted values of a sample in each tree as the prediction of the sample in the XGBoost system

The acronym for Extreme Gradient Boosting is XGBOOST. XGBoost, is integration with tree model that expected a sample in the XGB system using the sum of the simulated data of a sample in each tree[26]. a highly sought-after and well-liked algorithm that is frequently referred to as the platform-specific competition winner. The GB Algorithm has been enhanced by this algorithm. Gradient Boosting Decision Tree Algorithm is the fundamental algorithm. Because of its strong predictive ability and simple implementation method, it is widely used in machine learning notebooks. A few of the algorithm's main points are as

• Figure 1 shows how greedily it builds the tree structure rather than constructing it entirely. In contrast to XG boosting, it divides according to level wise.
• In Gradient Boosting, Taylor's expansion is considered while optimizing the loss function by taking into account

negative gradients.
• The regularisation term discourages the construction of intricate tree models.

Figure 1: Level wise growth in XGB

This paper discusses automatic classification of intellectual property legal cases and provides an explanation of XG boosting [26]

3.3 Light Gradient Boosting (LGB) Machine

LightGBM was introduced as a solution for the issues in time-consuming with the context of a large, high-dimensional sample of data [27].Light Gradient Boosting Machine, or LightGBM, is another boosting algorithm. In the field of machine learning, it is employed. Decision trees in LightGBM are grown leaf-by-leaf, which means that only one leaf at a time will be grown from the entire tree. as below.

Figure 2: leaf wise growth in LGB

This paper discusses automatic classification of intellectual property legal cases and provides an explanation of LGB [27].

3.4 Naïve Bayes

A set of probabilities is determined for each class by the probability - based classifier Naive Bayes. The method makes the assumption that every attribute is independent, which is rarely the case in the real world, and applies the Bayes theorem [25]. Naive Bayes is a classifier based on probabilities, which means that given a document d, it gives $c \in C$the class $\hat{c}$ that has the highest posterior probability. We use the symbol ^ to mean "our closest estimate of the correct class" in equation (1).

$$\hat{c} = avg\max_{c \in C} p(c|d) \quad (1)$$

Bayesian reasoning is an idea that has been around since Bayes's work. It was first used to classify text. The idea behind Bayesian classification is used by the Baye formula to change equation 1 into other probably events that are used. The Bayes rule comes in equation (2);. It lets us divide any conditional probability $P(x|y)$ into three other probabilities.

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)} \quad (2)$$

Then, we can put equation (1) into equation (2) to get equation (4):

$$\hat{c} = avg\max_{c \in C} p(c|d) = \operatorname{argmax}_{c \in C} \frac{P(d|c)P(c)}{P(d)} \quad (4)$$

We can make equation (15) easier to understand by taking out the term $P(d)$. We can do this because we will figure out$\frac{P(d|c)P(c)}{P(d)}$) for each possible class. But $P(d)$ stays the same for every class because We're always looking for the best possible class for document d, so it share the same $P(d)$. So, we can pick the class that makes this basic formula work best:

$$\hat{c} = avg \max_{c \in C} p(c|d) = \text{argmax}_{c \in C} P(d|c)P(c) \quad (5)$$

The Naive Bayes model is a model that is generative because equation(5) seems to make a claim about a document is made: first, chosen by a class from $P(c)$, and then chosen by the words from $P(d|c)$. This process could even be used to make fake papers, or at least documents with fake word counts.
In order to find the most likely class $\hat{c}$ for a given document d, we pick the class that has the greatest product of two probably event: the prior likely outcome of the class $P(c)$ and the likely hood of the document $P(d \mid c)$ as illustrated in equation (6)

$$\hat{c} = avg \max_{c \in C} p(d|c)\ P(c) \quad (6)$$

P(d|c) is Likelihood probability: There is a chance that the information given that a theory is true.
P(c) is Prior likely outcome: Chance of a before the hypothesis looking at the facts.
The predict posterior probability based on the prior probability illustrated in equation 7.

$$P(C = c|X = \text{x} = \frac{P(C=c)\Pi_i^I \text{P}(\text{X}_{\text{i=x}_{\text{i}}|\text{C=c}})}{\text{P(X=x)}} \quad (7)$$

3.5 The Proposed method Gaussian Kernel- Support Vector Machine method (GKSVM)

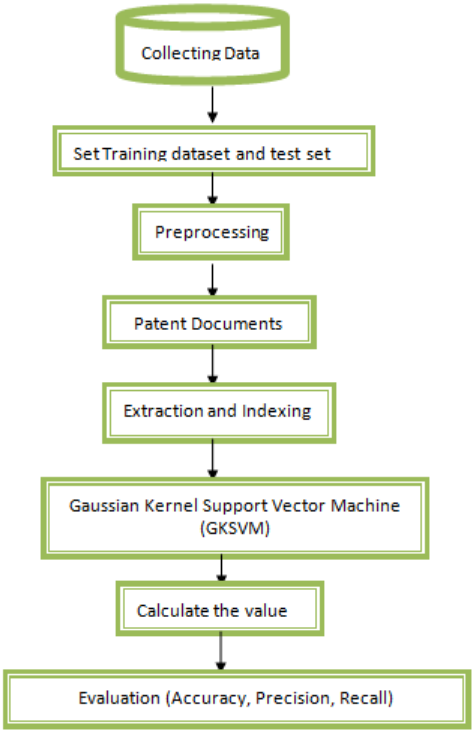

Figure 3: Flow of Proposed method

Data preprocessing using steerable filters (SF)

Since SF relies on the calculation of the patent legal cases derivative of Gaussians, local orientation maps of a property can be created using these filters. In essence, SF is a linear combination of the second derivatives of Gaussian distributions. The following formula (8) computes a 2-dimensional Gaussian at a specific pixel for an image $i\ (a,b)$. Equation (9) describes the SF formulations with a direction of θ. While the variable R, which is the deviation of the Gaussian function, is fixed, the outcome map of an image is created by integrating the outputs of individual SFs with varied θ values. The values of θ in this study vary between $0°\ to\ 360°$ at intervals of $30°$. Equation (10) is also used to compute the final answer map that SFs produce for an image i.

$$g(R,a,b) = \frac{1}{\sqrt{2\pi R}} exp \frac{-(b^2+a^2)}{2R^2} \quad (8)$$

$$f(\theta,R,a,b) = g_{aa}cos^2(\theta) + 2g_{ab}cos(\theta)sin(\theta) + g_{bb}sin^2(\theta) \quad (9)$$

$$R(a,b) = f(\sigma,a,b,\theta) * i\ (a,b) \quad (10)$$

Where the variances of the Gaussian function is represented by its independent parameter$R$. Gaussian 2nd derivatives are indicated by g_{aa}, $g_{ab,}$ andg_{bb}. $*$Indicates the convolutional operators sign.

Obtaining patent documentation

Approximately 1750 patent documents were gathered from various websites. Unstructured text is extracted from these documents. Given that the data were in HTML format, we were able to extract the contents of patent documents using GKSVM.

Sorting and pulling out terms

In this paper, we took the words out of each document, tokenized them, and got rid of any stop words. For each word or term, we used Porter's stemmer to split the stem part from the affix part. This was done because stemming helps people remember things. Then, an inverted index is made that shows the list of words (vocabulary), how often they appear, and how many times they were posted.

Grouping

We used the Support Vector Machine (SVM) [24], Naïve Bayes (NB) [25], XG Boosting (XGB) [26] and Light GBM (LGB) [27], and classifiers to classify the patent documents. A probabilistic unigram model was employed to categorize the patent documents. According to this model, sample 'x' belongs to class 'y', which has the highest probability and the least amount of risk. 10% of the documents in each class were used as samples were tested, and remaining 90% were used as training samples.

Gaussian Kernel SVM (GKSVM)

The first thing we look at is the use of support vector machines for classification. Equation (11) is utilized on the designated pattern's labeled training data.

$$\{(a_j,b_j)\}_j\ ^m_{\ =\ 1} \quad (11)$$

Using$b_j \epsilon \{-1,+1\}, a_j \epsilon r^M$. Applying a kernel to data points is explained in equation (12).

$$K(a,c): (r^M)^2 \to r \quad (12)$$

Equation (13), when reduced, is used to find the hyper plane that optimally separates the data. It is the internal stresses $\Phi(c) \bullet \Phi(a)$ in an unachieved feature space that may be high dimensional. Equation (14) aids in our optimization in the dual form. Equation (15) provides the sign $(H(a))$ that represents the decision function.

$$\tau(\xi,v) = D\sum_{j=j}^{m} \xi_j + \frac{1}{2}\|v\|^2 \quad (13)$$

$$V(\alpha) = \sum_{j=j}^{m} \alpha_j - 1/2 \sum_{ji} \alpha_j \alpha_i b_j b_i\ K(a_j,a_i) \quad (14)$$

$$H\,(a) = \sum_{L=1}^{N} \alpha_j b_j\, K(a_j, a_i) + y \qquad (15)$$

For the sake of clarification, there is a small infraction of notation: the attributes $a_L: L \in \{1,2,3,4, \ldots., N\}$shall be referred to as SVM. In order to classify a single point using a kernelized SVM, N kernel computations are typically needed, and all N-SVM must be retained. Since $K(a, c) = \langle a, c \rangle$. we can perform better with linear kernals. $H(a) = \langle v, a \rangle + y,,$ where $v = \sum_{L=1}^{N} \propto_j b_j a_j,,$ can thus be expressed as H (). Combining KSVM's strategies with computational evolutionary concepts, GKSVM is a hybrid highway surface recognition method.

The optimization method increases KSVM's classification accuracy for highway surface conditions and is modeled after Gaussian behavior in natural contexts. Gaussian Kernel Support Vector Machine (GKSVM) successfully tackles the problems of fracture detection on patent by employing this hybrid technique. Because of its adaptability and ability to manage complex datasets, it holds great promise as a tool for enhancing highway safety through the timely identification of problem areas and maintenance decision-making. Algorithm 1 presents the suggested approach.

Algorithm 1: Gaussian KSVM

Step 1: Import necessary libraries

$$and import numpy as np$$
$$from sklearn.svm import SVM$$
$$from sklearn.model\_selection import patent\_test\_split$$
$$from sklearn.metrics import accuracy\_score$$

Step 2: Define the GKSVM class

Class GKSVM:

$$def\ \_init\_\ (self, C = 1.0, kernel = 'rbf', gamma = 'scale'):$$
$$self.C\ =\ C$$
$$self.kernel\ =\ kernel$$
$$self.gamma\ =\ gamma$$
$$self.model\ =\ None$$
$$def train\ (self, X\_train, y\_train):$$
$$self.model\ =\ SVC(C = self.C, kernel = self.kernel, gamma = self.gamma)$$
$$self.model.fit(X\_train, y\_train)$$
$$def predict(self, X\_test):$$
$$return self.model.predict(X\_test)$$

Step 3: Load and preprocess the datavalue

Step 4: Split the datavalue into training and testing sets

$$X\_train, X\_test, y\_train, y\_test\ =\ train\_test\_split(X, y, test\_size, random\_state)$$

Step 5: Initialize and train the GKSVM model

$$gfo\_ksvm\ =\ G\_KSVM\ ()$$
$$g\_ksvm.train(X\_train, y\_train)$$

Step 6: Make predictions on the testing set

$$y\_pred\ =\ Gksvm.predict(X\_test)$$

Step 7: Evaluate the model

$$accuracy =\ accuracy\_score(y\_test, y\_pred)$$

Print ("Accuracy:", accuracy)

Data set: This study's data obtained from law offices. This work refines the GKSVM to the actual patent classification problem, which may help lawyers make data-driven decisions in patent cases. 94 original data points

in total were split up into testing and training datasets. Table 1 displays the data distributions.

Variables	Legal cases	Expression	Max	Min	Avg
M	Month of the case	Jan to dec	10	1	6.1
IPC-M	Patent documents in different country	8 Categories	8	0	4.5
CC	Three types intellective property	Invention,Utility, design patents	3	1	1.3
CC-M	Rights in patents	One rights in patents	25	1	2.8
CL	Identity of clients	Public,medium,person,court	4	1	3.42
AC	Scale of resources	Authorized capital	2,45,000	0	4569000
A	Law office	Small, large, medium, law office	4	2	4.55

Table 1: Intellectual legal property based on dataset

Gaussian Kernel σ and Intellectual property *C*

Since Gaussian kernel SVM has outstanding learning performance, it is widely used. The feature space will be mapped with samples to determined by the kernel width σ, which also has a significant impact on classification accuracy. When $\sigma \to 0$, all training samples can be classified correctly; however, the learning machine's generalization performance is poor, making SVM incapable of classifying new samples. When $\sigma \to \infty$, the entire sample set is trained with classified as a individual class. There is a mathematical explanation for this property. The feature space are mapped with samples using function ϕ(x). When $\sigma \to \infty$, Equation (16)

$$k(z_{i,}z_{i,}) = k(z_{j,}z_{j,}) = 1$$
$$k(z_{i,}z_{j,}) = 1$$
$$||\phi(zi) - \phi(zj)||2 = k(zi, zi) - 2k(zi, zj) + k(zj, zj)|| \quad (16)$$

When $\sigma \to 0$, it is simple to find

$$k(z_{i,}z_{i,}) = k(z_{j,}z_{j,}) = 1$$
$$k(z_{i,}z_{j,}) = 0 \quad (17)$$

Equation (16) becomes

$$\phi(zi) - \phi(zj)||2 = 2 \quad (18)$$

Equation (18) shows that any two samples in feature space are separated by $\sqrt{2}$ when $\sigma \to 0$. To ensure accurate classification of all the training data, samples belonging to the same class will not aggregate and will be classifier as

a single class. But because of over-fitting, the system is unable to categorize fresh samples and as equation (19) can be illustrated as

$$||\phi(zi) - \phi(zj)||2 = 0 \qquad (19)$$

According to equation (19), once samples are mapped to feature space, they become identical points with no separation between them, when $\sigma \to \infty$. As a result, all samples will be categorized into a single class, making it impossible for the computer to differentiate between the training set.

4. Result Analysis and Discussion

To compare the experiment's results, a number of indicators are required. The likelihood that the classifier will yield accurate predictions is referred to as the accuracy rate. The recall rate is the proportion of a given document classification's accuracy to all documents in that category within the document. The related documents are obtain divided by the total documents are obtained is the precision value. It predicts the obtained system's accuracy. The performance is better when the value is higher and closer to 1. Typically, we evaluate the impact of classification using accuracy.

Here the expression:

• Recall = (correct classification with patents count) / (count of patents that should fall into this category).

•Accuracy = (correct classification with patents count) / (all the patents documents in this experiment)

• Precision = (count of patents to be classified in this category) / (correct classification with patents count).

Accuracy

We assessed each classifier using the balanced F-measure, recall, and precision of existing techniques like SVM, NB, XGB, and LGB. As indicated in Table 1, a classifier model was derived by importing 1750 patents in order to compare the efficiency of the suggested methodology. We discovered that 10% of the total samples are test samples and 90% of the samples are training samples. The likelihood that the text classification will yield accurate predictions is referred to as the accuracy rate.

Accuracy examines the percentage of events that are reliably and effectively classified. Table 2 and Figure 5 show the accuracy's outcome. Our suggested method was superior to the current SVM (90.7%), NB (75.4%), XGB (85.4%), and LGB (88.7%) methods were GKSVM (96.5%). Comparing the different scenarios, GKSVM, new techniques, the effectiveness patent has been greatly enhanced.

Training Methods	**Test set Accuracy (%)**
SVM	90.7
NB	75.4
XGB	85.4
LGB	88.7
GKSVM [Proposed]	96.5

Table 2: Result value of Accuracy

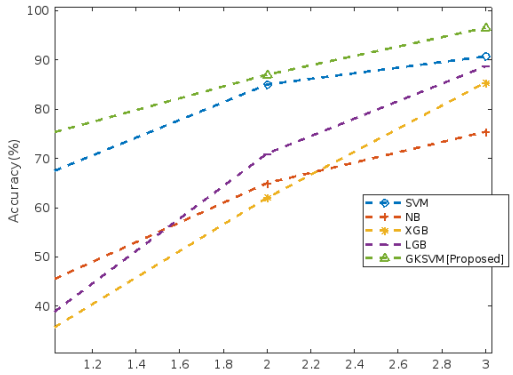

Figure 5: Accuracy

Precision

The precision measure indicates the proportion of precise patent legal documents; Figure 6 and Table 3 present the findings. In contrast to the current approach, which uses SVM (90.2%), NB (73.4%), XGB (84.8%), and LGB (85.6%), We suggested using a higher GKSVM of 95.7%. When compared to existing techniques, the new proposed approach, GKSVM, has greatly enhanced intellectual property legal case test set prediction.

Method	Precision (%)
SVM	90.2
NB	73.4
XGB	84.8
LGB	85.6
GKSVM [Proposed]	95.7

Table 3: Result value of Precision

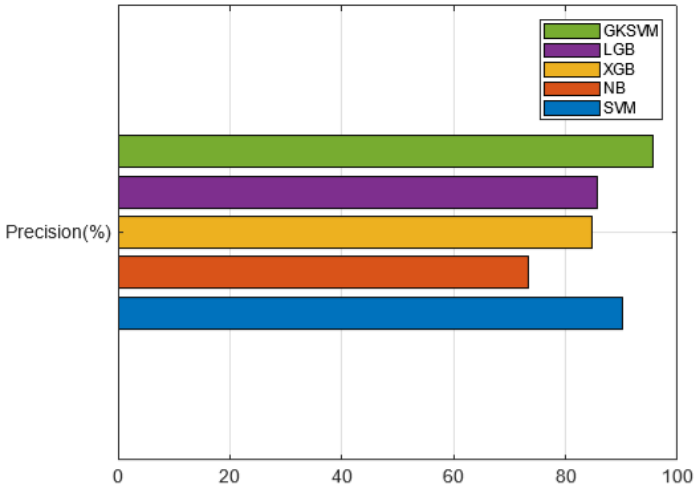

Figure 6: Precision

Recall

Out of all real positives, recalls indicate the proportion of true positives that are successfully recognized. Table 4 and Figure 7 show the recall's outcome. Our proposed approach, GKSVM (95.1%), outperformed the current methods, SVM (90.1%), NB (72.1%), XGB (84.2%), and LGB (88.2%). Consequently, it is advised to employ the proposed

methods, The description will be used by GKSVM as the classifier's input data and in the patent's automatic classification.

Method	Recall (%)
SVM	90.1
NB	72.1
XGB	84.2
LGB	88.2
GKSVM [Proposed]	95.1

Table 4: Result value of Recall

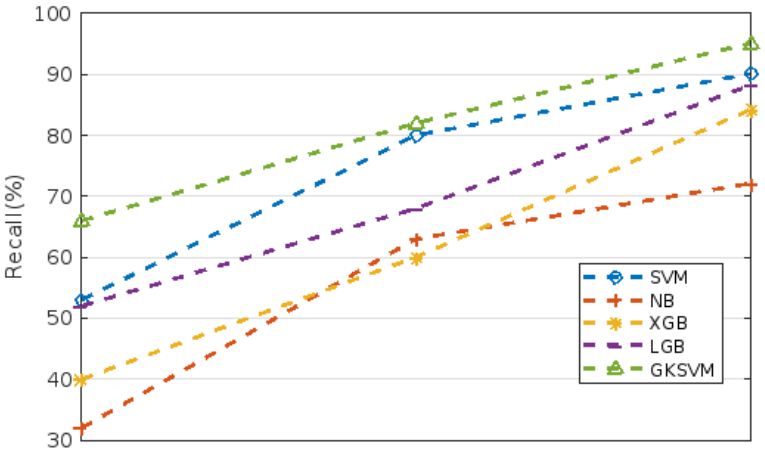

Figure 7: Recall

F-Measures

Compares the uneven margins of SVM's F-measure results with the standard SVM's. As we can see, the Precision obtained by the standard SVM ($\tau = 1$) was significantly higher than the Recall. On the other hand, the SVM with uneven margins produced a higher F1 and balanced Precision and Recall. Using the F-measure on the 110 data points of F1, we were able to determine that the difference's mean was 0.0482 and that its 95% confidence interval was [0.0421, 0.0542]. With uneven margins, the GKSVM produced a statistically significantly better F1 than the SVM ($t = 15.75$ and $P < 0.0001$). Table 5 and Figure 8 show the F-Measure's outcome. In way of comparison to SVM (90%), NB (73.2%), XGB (83.8%), and LGB (87.5%), which are the current methods We proposed using a higher GKSVM of 95.5%. When compared to existing methods, the proposed approach, GKSVM, has significantly improved intellectual property legal cases effectiveness prediction.

Method	F-Measures (%)
SVM	90
NB	73.2
XGB	83.8
LGB	87.5
GKSVM [Proposed]	95.5

Table 5: F-Measure's outcome value

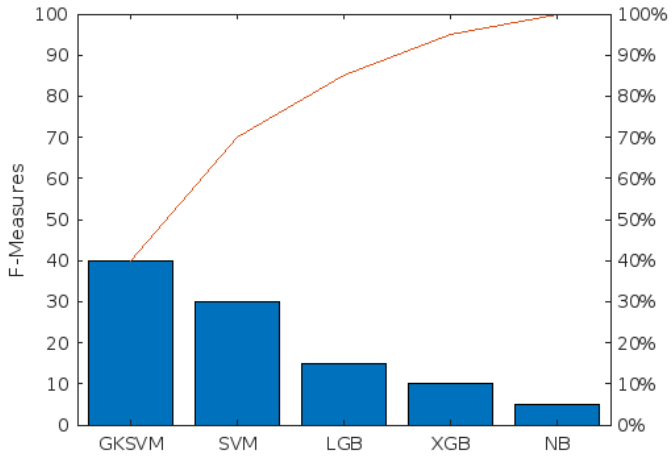

Figure 8: F-Measures with other methods

Conclusion

The enormous volume of data has drawn the serious attention of many researchers to the automatic classification of intellectual property legal cases based on SVM. A database containing documents from the international patent collection is called international patent documentation. About 1750 patent documents were collected from various subclasses with different websites; 90% of the documents were used as training samples and 10% as test samples. We used the F-measure, accuracy, precision, and recall metrics to evaluate the classifier model's performance. Furthermore, we noticed that the performance of certain classifiers improved with an increase in features, and the suggested method GKSVM was evaluated in terms of time and accuracy, yielding efficient results such as accuracy (96.5), precision (95.7), recall (95.1), and F-Measures (95.5).All things considered, the GKSVM patent classification presents an intriguing chance for the ML community to advance the techniques and systems because of its more difficult nature.

In the future, it would be interesting to study other types of machine learning datasets with GKSVM. We could use encryption to incorporate categorical variables into our model. Future work could also incorporate additional information preprocessing methods to enhance the SVM.

Reference

[1] Venugopalan, Subhashini, and Varun Rai. “Intellectual Property.” *Technological Forecasting and Social Change*, vol. 94, May 2015, pp. 236–250, https://doi.org/10.1016/j.techfore.2014.10.006. Accessed 14 May 2020.

[2] Aristodemou, Leonidas, and Frank Tietze. “The State-of-The-Art on Intellectual Property Analytics (IPA): A Literature Review on Artificial Intelligence, Machine Learning and Deep Learning Methods for Analysing Intellectual Property (IP) Data.” *World Patent Information*, vol. 55, Dec. 2018, pp. 37–51, www.sciencedirect.com/science/article/pii/S0172219018300103, https://doi.org/10.1016/j.wpi.2018.07.002.

[3] Miric, Milan, et al. “Using Supervised Machine Learning to Create Categorical Variables for Use in Management Research: The Case for Identifying Artificial Intelligence Patents.” *Strategic Management Journal*, 29 June 2022, https://doi.org/10.1002/smj.3441. Accessed 9 July 2022. [4]M. Kim and Y. Geum, “Predicting Patent Transactions Using Patent-Based Machine Learning Techniques,” *IEEE Access*, vol. 8, no. 5, pp. 188833–188843, 2020, doi: https://doi.org/10.1109/access.2020.3030960.

[5]C.-J. Lin, “Formulations of Support Vector Machines: A Note from an Optimization Point of View,” *Neural Computation*, vol. 13, no. 2, pp. 307–317, Feb. 2001, doi: https://doi.org/10.1162/089976601300014547.

[6]C. Lee, O. Kwon, M. Kim, and D. Kwon, “Early identification of emerging technologies: A machine learning approach using multiple patent indicators,” *Technological Forecasting and Social Change*, vol. 127, no. 3, pp. 291–303, Feb. 2018, doi: https://doi.org/10.1016/j.techfore.2017.10.002.

[7]R. Hall, "The strategic analysis of intangible resources," *Strategic Management Journal*, vol. 13, no. 2, pp. 135–144, Feb. 2020, doi: https://doi.org/10.1002/smj.4250130205.

[8]M. Zhao, "Conducting R&D in Countries with Weak Intellectual Property Rights Protection," *Management Science*, vol. 52, no. 8, pp. 1185–1199, Aug. 2006, doi: https://doi.org/10.1287/mnsc.1060.0516.

[9]J. Yun and Y. Geum, "Automated classification of patents: A topic modeling approach," *Computers & Industrial Engineering*, vol. 147, no. 7, p. 106636, Sep. 2020, doi: https://doi.org/10.1016/j.cie.2020.106636.

[10] Chen, Yen-Liang, and Yuan-Che Chang. "A Three-Phase Method for Patent Classification." *Information Processing & Management*, vol. 48, no. 6, Nov. 2012, pp. 1017–1030, https://doi.org/10.1016/j.ipm.2011.11.001. Accessed 14 Aug. 2020.

[11] Chen, Hongshu, et al. "Topic-Based Technological Forecasting Based on Patent Data: A Case Study of Australian Patents from 2000 to 2014." *Technological Forecasting and Social Change*, vol. 119, June 2017, pp. 39–52, https://doi.org/10.1016/j.techfore.2017.03.009. Accessed 30 Apr. 2020.

[12] Venugopalan, Subhashini, and Varun Rai. "Topic Based Classification and Pattern Identification in Patents." *Technological Forecasting and Social Change*, vol. 94, May 2015, pp. 236–250, https://doi.org/10.1016/j.techfore.2014.10.006. Accessed 14 May 2020.

[13] Kim, Mujin, et al. "Generating Patent Development Maps for Technology Monitoring Using Semantic Patent-Topic Analysis." *Computers & Industrial Engineering*, vol. 98, Aug. 2016, pp. 289–299, https://doi.org/10.1016/j.cie.2016.06.006. Accessed 1 Dec. 2019.

[14] Kyebambe, Moses Ntanda, et al. "Forecasting Emerging Technologies: A Supervised Learning Approach through Patent Analysis." *Technological Forecasting and Social Change*, vol. 125, Dec. 2017, pp. 236–244, https://doi.org/10.1016/j.techfore.2017.08.002.

[15] Hu, Jie, et al. "Patent Keyword Extraction Algorithm Based on Distributed Representation for Patent Classification." *Entropy*, vol. 20, no. 2, 2 Feb. 2018, p. 104, https://doi.org/10.3390/e20020104. Accessed 29 May 2022.

[16] Li, Chuanxiao, et al. "A Patent Retrieval Method and System Based on Double Classification." *Information Sciences*, 1 Apr. 2024, pp. 120659–120659, https://doi.org/10.1016/j.ins.2024.120659. accessed 22 June 2024.

[17]M. Wang, Hiroki Sakaji, Hiroaki Higashitani, Mitsuhiro Iwadare, and K. Izumi, "Discovering new applications: Cross-domain exploration of patent documents using causal extraction and similarity analysis," *World patent information/World patent information (Online)*, vol. 75, no. 3, pp. 102238–102238, Dec. 2023, doi: https://doi.org/10.1016/j.wpi.2023.102238.

[18]C. Jiang, Y. Zhou, and B. Chen, "Mining semantic features in patent text for financial distress prediction," *Technological Forecasting and Social Change*, vol. 190, no. 3, p. 122450, May 2023, doi: https://doi.org/10.1016/j.techfore.2023.122450.

[19] Zaini, Wan Mohammad Faris, et al. "Identifying Patent Classification Codes Associated with Specific Search Keywords Using Machine Learning." *World Patent Information*, vol. 71, Dec. 2022, p. 102153, https://doi.org/10.1016/j.wpi.2022.102153. Accessed 27 Jan. 2023.

[20]Y. Lin, X. Wang, J. Yang, and S. Wang, "Core Technology Topic Identification and Evolution Analysis Based on Patent Text Mining—A Case Study of Unmanned Ship," *Applied sciences*, vol. 14, no. 11, pp. 4661–4661, May 2024, doi: https://doi.org/10.3390/app14114661.

[21]J. Wang *et al.*, "Enhancing patent text classification with Bi-LSTM technique and alpine skiing optimization for improved diagnostic accuracy," *Multimedia tools and applications*, vol. 1, no. 5, May 2024, doi: https://doi.org/10.1007/s11042-024-18806-8.

[22]J. Wright, V. Weber, and G. M. Walton, "Identifying potential emerging human rights implications in Chinese smart cities via machine-learning aided patent analysis," *Internet policy review*, vol. 12, no. 3, Jul. 2023, doi: https://doi.org/10.14763/2023.3.1718.

[23]T. Ha and J.-M. Lee, "Examine the Effectiveness of Patent Embedding-Based Company Comparison Method," *IEEE access*, vol. 11, no. 3, pp. 23455–23461, Jan. 2023, doi: https://doi.org/10.1109/access.2023.3251664.

[24]C.-H. Wu, Y. Ken, and T. Huang, "Patent classification system using a new hybrid genetic algorithm support vector machine," *Applied Soft Computing*, vol. 10, no. 4, pp. 1164–1177, Sep. 2010, doi: https://doi.org/10.1016/j.asoc.2009.11.033.

[25] McCallum, Andrew, and Kamal Nigam. "A Comparison of Event Models for Naive Bayes Text Classification." *National Conference on Artificial Intelligence*, 1 Jan. 1998, pp. 41–48. Accessed 26 June 2024.

[26]T. Han, "Research on Chinese Patent Text Classification in the Field of New Energy Vehicles Based on the XGBoost Model," *EITCE '23: Proceedings of the 2023 7th International Conference on Electronic Information Technology and Computer Engineering*, vol. 5, no. 1, pp. 481–485, Oct. 2023, doi: https://doi.org/10.1145/3650400.3650479.

[27]J. Liu, P. Li, and X. Liu, "Patent lifetime prediction using LightGBM with a customized loss," *PeerJ. Computer science*, vol. 10, no. 2, pp. e2044–e2044, May 2024, doi: https://doi.org/10.7717/peerj-cs.2044.