ATLANTIS PRESS

Advances in Intelligent Systems Research (AISR), volume 155

2018 International Conference on Advanced Control, Automation and Artificial Intelligence (ACAAI 2018)

Analysis of Criminal Case Judgment Documents Based on Deep Learning

Jinbo Han, Dakui Li, Nanhai Yang*, Zhu Liu and Qiong Nan

School of Software Technology, Dalian University of Technology, Dalian, China

*Corresponding author

***Abstract*—In recent years, along with the improvement of population quality and the advancement of the rule of law society, the market for legal services in the middle and low-end markets has continued to expand, and legal advice has become widespread in daily life. In the process of legal services, the legal provisions play an important role in the lawyer's decision-making. Meanwhile, the historical cases can help the lawyers and the parties to draw lessons from similar cases. However, with the increasing number of judicial documents, it is becoming increasingly difficult to summarize and learn from history. Therefore, this paper proposes a sentencing interval prediction model of criminal cases based on convolutional neural network, and through the method of multi-core convolution, greatly enhances the generalization ability and prediction performance of the model. The experimental analysis of real criminal case verdict verifies that the model is more effective than other classification prediction algorithms.**

Keywords—convolutional neural network; multiple nuclear convolution; criminal case verdict document; prediction of sentencing interval

I. INTRODUCTION

With the rapid progress of the legal system in China, the concept of judicial autonomy in particular is now widely recognized. Chinese nationals are increasingly required the help of lawyers in their daily lives. However, in the teaching of law, the most common sentence is "related cases cannot be exhaustive enumeration, we can only choose typical cases as research objects." As of the end of 2016, a total of 26.8 million articles of arbitration were uploaded by people's courts at all levels across the country in China's arbitration documents. It shows the reference value of historical cases and the number of cases has restricted the analysis for lawyers. With the increase of the number of historical cases, more and more scholars at home and abroad have studied the legal instruments: Liu S analysed the trial results of the criminal cases of the Supreme People's Court and summarized 10 suggestions on judgments of criminal cases in the future [1]. ZH Lin researched of criminal case semantic feature extraction method based on the Convolutional Neural Network [2]. M Lei based on Machine Learning algorithms to implement automatic classification of Chinese Judgment Documents [3]. YL Chen designed a text-mining-based method that allows the general public to use everyday vocabulary to search for and retrieve criminal judgments [4].Similar researches mostly make statistics on the contents of referees' documents through statistical knowledge and find out some superficial rules in the documents. The work of this paper is to predict how many years a criminal suspect can sentence given by continuously studying the features of cases in the arbitration instruments, which can be used as a reference for lawyers and clients.

Although there are fewer direct studies on predicting the judgment range of the cases in the existing literatures, under the actual legal advice scene, the lawyer hopes that the interests of the parties can be best preserved, at the same time of a wide range of case history, find the breakthrough point of the case. The main work of this paper is based on the results of a wide range of historical cases, give to the lawyer a sentencing interval which can be used as a reference for lawyers. This paper mainly analyses the case of homicide cases in the criminal case. In this scenario, we take into account the strong marking language in legal instruments and the extraction of the legal vocabulary, and propose a decision based on the multi-core convolution neural network algorithm Interval prediction model, and through real criminal case data set experiments. Compared with the commonly used prediction models, such as support vector machine (SVM), naive Bayes, the experimental results show that the proposed prediction model based on multi-core convolution neural network has better performance on criminal case analysis indicators.

II. MODEL

In this paper, the sentencing interval prediction model is based on the traditional convolutional neural network model and to transform and parameter tuning. The overall structure of the model shown in Figure I.

The model is divided into four layers, which are input layer, convolution layer, pooling layer and full connection layer. The input layer is a vector matrix that is a word-embedding mapping obtained by word vectorization. Convolutional layer is a multi-core convolution method, composed of convolution kernels of different sizes. Convolution kernels of different sizes extract features of different local perception domains to improve the generalization ability and prediction performance of the model. Then for the extracted local features, the most important feature in each local perception domain is obtained by pooling the layers. At last, the eigenvector extracted by the full connection layer is connected, and the mapping of the output is completed to obtain the possibility of different output.

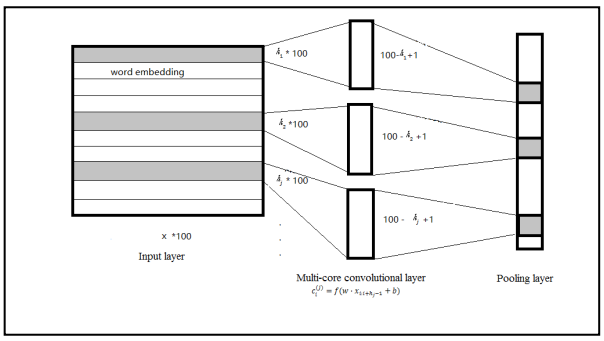

FIGURE I. THE OVERALL STRUCTURE OF THE MODEL.

A. *The Structure of Input Layer*

Suppose our input is a two-dimensional data as n*k, which means that we have a total of n words in the input text, then each line $x_i \in R^k$ represents a k-dimensional vector, which corresponds to the text in the i-th word vector of sentences. So for a length of n text $x_{1:n}$, it is expressed as

$$x_{1:n} = x_1 + x_2 \ldots + x_n \quad (1)$$

Where, + represents the connection operation, $x_{1:n}$ represents to connect the first to n-th words, the connection shown in Figure 2, where n is 100.

B. *Multi-core Convolutional Structure*

Multi-core convolutional layer is to select a number of convolution kernel $w \in R^{h_j,k}$ to extract the input features, where h_j is the j-th convolution kernel's height, that is represents the number of words covered by each convolution, k represents the number of dimensions of each word covered by each convolution. For each convolution window $x_{i:i+hj-1}$ of height h_j , the feature map of the j-th convolution kernel after convolution is

$$c_i^{(j)} = f(w \cdot x_{i:i+h_j-1} + b) \quad (2)$$

Where $b \in R$ is the offset term, w is the weight matrix represented by the convolution kernel, f is the activation function, in this paper, and we choose ReLU function [5]. Apply the j-th convolution kernel to every input window of height h_j , finally return a feature of all the input data:

$$c^{(j)} = [c_1, c_2 \ldots c_{n-h+1}] \quad (3)$$

Where $c^{(j)} \in R^{n-h+1}$.

C. *Pooling Layer Structure*

Because a handful of keywords can get the meaning of a sentence, one or more of the most representative local features in each feature map can be obtained at the pooling level. The addition of a pooling layer also prevents learning features from being overly dimensioned, or over-fitting. There are many ways of pooling in convolutional neural networks, including max-pooling, min-pooling, mean-pooling, and so on. In order to better extract the features of the input layer and reduce the training parameters, and to solve the problem that some information may be lost when the step size is large, max-pooling method is used in this paper to sequentially sample 2 * 2 windows on each feature map's maximum.

D. *Full Connection Layer*

In this part, we map the distributed learning features to the markup space of the sample through the full connection layer. Full connection layer plays a "classifier" role, to be able to take into account all the features extracted, to complete the prediction of sentencing interval.

In order to prevent overfitting, we add Dropout method in the full connection layer to randomly discard a part of neurons each time, which is equivalent to training in different network models. This not only reduces over-fitting but also improves accuracy. That is, for the pooled features $z = [\widehat{c_1}, \ldots \widehat{c_m}]$ (m is the number of convolution kernels), when mapped to the sample space:

$$y = w\ `(z * r) + b \quad (5)$$

Where * is a multiplication operation, r ~ Bernoulli (p) is a Bernoulli random variable with probability p, that is, some neurons are discarded by the probability p. Other neurons are retained by the probability q = 1-p. The output of the rejected neurons is set to zero.

E. *The Training of CNN Model*

In order to minimize the cost function, the gradient descent method is mainly used at this stage. For several variants of the gradient descent method, after many considerations, this paper chooses the mini-batch gradient descent method [6], which avoids a lot of computational costs and convergence time in the case of a large number of datasets.

At the same time, we also need to prevent the occurrence of over-fitting in training process. This paper adopts the method of L2 regularization [7], which reduces the complexity of the model and reduces the risk of over-fitting.

III. EXPERIMENT ANALYSIS

The experiment of this paper is to analyze and verify the validity of the judgment interval prediction model for criminal cases adjudication instruments. In order to compare with the existing classification model, the text of the second part will respectively use the judgment interval prediction model, the support vector machine and the naive Bayesian model to test the civil case judgment instruments, and make statistics on the accuracy of each model.

A. *Data sets Description*

The data set in this paper is divided into two parts. The first part of the data set includes four types of inheritance case, Intentional killing cases, loan contract cases and intellectual

property cases, mainly for analyzing the main parameters that influence the CNN model in the judgment document data. A total of 2000 data, we randomly divided these samples into a training set and test set, the classification results a total of 1200 training samples and 800 testing samples. The distribution of these seven samples is shown in Table I.

TABLE I. TYPES OF CASES DATA SET SAMPLE DISTRIBUTION

Category	*Labels*	*Training samples*	*Test samples*
0	Inheritance case	300	200
1	Intentional killing	300	200
2	Loan Contract	300	200
3	Intellectual property	300	200

The second part of the data is a data set of murder cases, and mainly to verify the validity of the predictive model of criminal cases judgment interval. Samples and labels are shown in Table II. The labels mean the number of years that the suspect may be sentenced.

TABLE II. THE DISTRIBUTION OF INHERITANCE CASE DATA

Category	*Labels*
0	1-10
1	More than 10 years
2	Life imprisonment

B. Data sets Pre-processing

In order to better apply the model of this article to the judgment documents, the texts need to be pre-processed. First, based on the "plaintiff's claim", "the defendant's argument", "the trial's findings" and "the court's opinion" These highly tagged language will be able to fully describe the case process, the corresponding legal module corresponding. Then the extracted text segmentation and vectorization.

Because the vast majority of the referee instruments are legal norms and common words, this article further expands the lexicon database for word segmentation and adds many common legal terms to achieve the best word segmentation. In the process of word vectorization, we do not use the word vectors trained by Wikipedia. Because the number of words in Wikipedia is too large and the similarity with the text in this article is too low, the resulting word vectors do not show the continuity between words and the words in the data set well .Therefore, in this paper, the Skip-Gram model is trained by combining the word segmentation with the collected legal-specific vocabulary. After training, the resulting word-embedding is derived in a non-binary form.

C. Experimental setup

For the first part of the experiment, in order to analyze the main parameters that affect the performance of the predictive model, we set the pre-training iteration steps to 4,000 steps uniformly and the learning rate of the model to 0.001, and analyzed the height of convolution kernel, the effect of the number of kernels on the model performance. In the experiment, the first part of the data is selected, and the classification task of the text in civil case is taken as the evaluation index of the model.

In the second part of the experiment, the prediction of sentencing interval is mainly based on the optimal parameters obtained in the previous section, and the validity of the prediction is analyzed by taking the case of murder as examples.

D. Experimental Results and Analysis

1) Effect of single convolution kernel height: First of all, we analyze the impact of the height of a single convolution kernel on the performance of model classification. The input dimension is 100 * 100, the number of convolution kernels is 100, the convolution kernel height is5, 9, 11, 15, 20, 25, 30, the results obtained by the experiment shown in Table III.

TABLE III. EFFECT OF SINGLE CONVOLUTION KERNEL HEIGHT

Convolution kernel height	*Experimental results*
5	0.855
9	0.872
11	0.872
15	0.881
20	0.889
25	0.898
30	0.864

Through the above experimental results, it can be found that in the case of a single convolution, the classification performance for the referee documents obtained when the convolution kernel height is 25 is optimal.

The effect of multi-convolution kernel height: Since multiple convolution kernels can extract text features from different perspectives, and the convolution kernel energy with a height near 25 can be better extracted from the above experiments, so the multi-convolution kernels are set as {7,8,9}, {15,20,25}, {7,8,9,10}, {15,20,25,30}, {25,25,25}, {25,25,25,25}, each of them was tested and the results obtained are shown in Table IV.

From Table IV, one can see that using {15, 20, 25}, and {15, 20, 25, 30} that sets near the best single region size can produce the best results. Note that even only using a single good filter region size results in better performance than combining different sizes{7,8,9}. in order to balance training results and training time, we choose {15, 20, 25} in the next experiment.

TABLE IV. THE EFFECT OF MULTI-CONVOLUTION KERNEL HEIGHT

Convolution kernel height	*Experimental results*
{7,8,9}	0.854
{15,20,25}	0.901
{7,8,9,10}	0.863
{15,20,25,30}	0.881
{25,25,25}	0.899
{25,25,25,25}	0.872

2) The number of convolution kernels: In Ye Zhang's paper[8], he said that the increasing number of maps beyond 600 yields at best very marginal returns, and often leads to over fit. Another salient practical point is that it takes a longer time to train the model when the number of feature maps is increased. Therefore, in order to balance training results and training time, we choose 150 convolution kernels.

3) Criminal case sentencing interval prediction model validity analysis: Based on the parameters obtains in the first part of the experiment, and the experimental data of murder cases, this part will analyze the experimental results of the predictive model for the sentencing interval of criminal cases. At the same time, Some classification models are compared. In this paper, support vector machine and naive Bayesian model are used to predict the compensation cases in civil cases. The experimental results are shown in Table V.

TABLE V. THE AMOUNT OF SENTENCING INTERVAL FORECAST RESULTS

Model	*Experimental results*
Criminal Case sentencing interval prediction model	0.913
Support Vector Machines	0.791
Naive Bayes	0.772

After the above experiment, we use the cross-validation method to evaluate the accuracy of each model. We can find that the Criminal Case sentencing interval prediction model has the highest accuracy of training data, which shows that convolutional neural network model is superior to ordinary machine learning methods in extracting features. At the same time, it shows that applying deep learning to the legal profession has good application value.

CONCLUSIONS

The purpose of this paper is to give the forecasted range of lawyer's criminal case judgment results in specific scenarios. The actual meaning is that a lawyer for a new case, by providing the background of the case, the story of the case, the information about the parties concerned, can give to the layer a sentencing interval. Through this sentencing interval, the lawyer understands the court's punishment for such cases is generally given, which will be an important reference factor. Based on the results of this article, we will help lawyers and the parties concerned to make better decisions which is not only according to the law but also combined with the analysis of history cases.

The main work of this paper includes data acquisition and model selection. In the process of model selection, in addition to the choice of the model itself, the paper also analyzes the influence of the parameter setting of the sentencing interval prediction model and obtains the optimal parameter setting for the classification of the referee instrument data by multiple groups of experimental analysis. Based on the obtained parameters, a series of experimental results are obtained by comparing different classification prediction algorithms. Finally, the validity of the interval prediction model of criminal case decision results based on the multi-core convolutional neural network is also verified.

For the future work, firstly, in the actual use, select a larger number of data sets to establish a predictive model and test the rate of different regions of the court in different cases of the degree of propensity; secondly, in practice, after entering the information, the system can not only provides a range of sentencing, but also extracts similar typical cases from the database for the lawyer to further analyse the case process.

ACKNOWLEGEMENT

This work is supported by National Natural Science Foundation of China (Grant No. 61672131) and the Fundamental Research Funds for the Central Universities (DUT16QY27).

REFERENCES

[1] Liu S. Analysis of Summary of Judgment in Criminal Cases by Supreme People's Court (Part II)[J]. China Law, 2012.

[2] Lin Z H, Chi H, Bao-Guang X U. Research of Criminal Case Semantic Feature Extraction Method Based on the Convolutional Neural Network[J]. Mathematics in Practice & Theory, 2017.

[3] Lei M, Ge J, Li Z, et al. Automatically Classify Chinese Judgment Documents Utilizing Machine Learning Algorithms[J]. 2017.

[4] Chen Y L, Liu Y H, Ho W L. A text mining approach to assist the general public in the retrieval of legal documents[J]. Journal of the American Society for Information Science & Technology, 2013, 64(2):280–290.

[5] Liu J, Fan D, Tian R. Neural network prediction model of rolling force based on ReLU activation function[J]. Forging & Stamping Technology, 2016.

[6] Konečný J, Liu J, Richtárik P, et al. Mini-Batch Semi-Stochastic Gradient Descent in the Proximal Setting[J]. IEEE Journal of Selected Topics in Signal Processing, 2016, 10(2):242-255.

[7] Demirkavuk O, Kamada M, Akutsu T, et al. Prediction using step-wise L1, L2 regularization and feature selection for small data sets with large number of features[J]. BMC Bioinformatics,12,1(2011-10-25), 2011, 12(1):1-10.

[8] Zhang Y, Wallace B. A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification[J]. Computer Science, 2015.