Criminal Punishment Prediction Based on Fuzzy Document Vector Model

Taotao Fu¹, Dezhen Kong¹, Yilin Jin¹, Canghong Jin¹*, Min Zheng², Bing Chen² and Danyan Zhu³

¹Zhejiang University City College, Huzhou Road 51, Hangzhou, China
²Taizhou Public Security Bureau, Kangping Road 2, Taizhou, China
³Xianju Public Security Bureau, 900 South Huancheng Road, Xianju, China
Email: jinch@zucc.edu.cn

Abstract. The drug addict problem is more and more critical to the world during current periods, analysing the criminal process and predicting the final punishment from judgments report is an interesting and important task. Existing studies on text analysis and language model supply methods based on special feature selection and ontology models generation which need external knowledge by human experts. In this paper, however, we creatively leveraged such text data onto prediction in the public judgments without human business. We propose a combined framework to capture the prediction problem by considering both valued based rules and fuzzy document models. This framework contains the complete process as: information extraction, term fuzzy and document vector regression. We setup an experiment on a real-world dataset and compare our model with traditional classification and regression methods. The results show that our model outperforms than others by both RMSE and R squared measures.

1. Introduction
Addiction to drugs among men and women is an acute social problem faced by most of the countries worldwide. The National Narcotics Control Commission reported that the number of drug addicts is still slowly growing in China. There were more than 2.5 million drug users in the country by the end of 2016 with 6.8 percent growth [1], while in 2018 World Drug Report, about 275 million people worldwide, which is roughly 5.6 percent of global population aged 15-64 years, used drugs at least once during 2016 [2]. Drug addict not only deeply affects the individual health but also is thought to be guilty of crime or offense, who would lead to public safety concerns. It is critical to analyse the process of the crime and its final punishment, which could find the key factors of crime and help the judge to estimate the prison term. Therefore, we need to obtain both the context of crime process and predicted punishment value. We use criminal judgements as our study target for three main reasons. First one is the judgement report is well written and follow certain format, which could be relatively easier to extract entities. Second, the judgement report contains rich text information as criminal time, place, motive, purpose, means and circumstances reflecting the facts of the criminal act of the perpetrator of a case and so on. Third, the data of judgment is public and could be downloaded online. Therefore, criminal judgments are good materials for analysing the drug related cases and capturing the prediction problems.

With the rapid advances in machine learning and knowledge graph techniques, the text based prediction problems have been studied by traditional algorithms or ontology based reasoning.
However, most of the existing studies focused on effective features selection, ontology generation and pattern discovery. To our best knowledge, our study is the first to focus on the context of judgement itself and the accuracy of prediction is not depended on the quality of entity and ontologies. In fact, it is possible to infer the result from raw text on judgment with some deep learning models, include Natural Language Program, text vector generation and Neural Network based regression methods.

In summary, to predict the punishment from judgment reports, we are faced with a number of challenges.

- The first challenge is how to extract related context to certain people. Since many judgements usually have several defendants, and their criminal process are across each other, which bring the difficulty for system to divide them correctly.
- Second, for each drug related case, it is hard to distinct which word is more important than others by human business. In particular, most of drug related judgments are looked similar by the description words and the process structures. Identifying such small difference and find the influence of final punishment result is not a trivial work.
- Third, there are many special terms as defendant’s name, birthday, hometown, the address and time of trade process or other mentioned suspects and vehicle, which make the text diversity but useless for prediction model. Thus, we need a mechanism to transfer this information before the learning process.
- Finally, the final punishment not only depend on the description of criminal process but also depends on the number of drug and the number of drug money. We need to a combined model to consider both sides of conditions and improve the prediction accuracy.

To this end, in this paper, a comprehensive approach is taken to cover the above challenges. Specially, we first build an effective extractor to generate the context of each defendant. Furthermore, we establish a combined framework to predict the punishment, which take both number value and text characteristic into account. Finally, we leverage real-world judgment datasets for model training and validation, and then implement a prototype system for end users.

2. Related Work

Researches on predicting criminal punishment through text of judgement is rare. However, researchers have achieved some related work such as text summary score prediction and legal reasoning. In this section, we provide a brief review of those work.

2.1. Text Summary Score Prediction

In article [3], the author uses a theme model to mine the comment theme which is hidden in user's reviews about dining. He characterizes user preferences and portraits with these review themes and finally, trains the relationship between topic and score by linear regression and logistic regression, to get a more accurate score prediction result. In article [4], the author focuses on English film commentary data, he uses a hidden variable model to describe the dependence between user behaviour preferences and feature attributes, after which he conducts a heuristic structure learning on models through domain knowledge guidance and then infers the movie score by user behaviour preference models. The average prediction accuracy of this model is 72%. At the meantime, in order to improve the accuracy of forecasts, the author of article [5] presents a new score prediction framework based on document vectors and incorporating multiple regression models. The result of this framework is better than the base model.

2.2. Ontology Reasoning

Recently, the convergence of technology in the field of artificial intelligence and law is becoming more and more significant, especially in terms of laws and regulations, formal descriptions of legal cases, and legal reasoning [6]. A. Valente [7] believes that the ontology can establish the missing link between legal theory and artificial intelligence. The legal ontology that has great influence on this research field includes: Deep conceptualization model of legal field (LLD) [8], Legal function
ontology (FO-Law) [9], LRI-core which based on FO-Law [10], The legal core ontology that realizes the interaction of legal knowledge systems (LKIF-core) [11], and so on. The purpose of these ontologies is to store laws, regulations and legal precedents under the ontology framework, so as to preserve the legal concepts, entities and their semantic relationships more completely, and that can be beneficial for facilitating the extraction and application of related information in the later stage.

3. Prediction Framework
As is shown in figure 1, our work consists of data pre-processing part and analysis part. In data pre-processing stage, we have obtained useful information about cases and criminals along with abstract of original text from judgement report by a text extractor, which had been developed by ourselves. And we've also selected some case features with help of human knowledge and judgment experience. In analysis stage, we have used two ways to make prediction of period in prison more accurate and sensible. The first way is predicting by traditional machine learning methods, such as KNN, SVM, Decision Tree on the basis of all criminals' several features. The second way is predicting by document vector on the basis of extracted abstract of origin text. The procedures of this way can be described as follows:

1. Document fuzzy [12-15]. It is obvious that some elements in a document, like stop words, concrete time and place, person names and other numerical values must be removed or replace by some fuzzy words. Because these elements are with no meaning for prediction.

2. Building document vectors. In (1), we have got processed documents, which are fit for further analysis. Thus we built document vectors with the help of algorithm [16-17]. We have also chosen other similar and processed documents for test.

3. Test Documents. For each document to be test, we get similar document vectors and calculate predicted periods in prison.

In addition, importantly, we have validated the performance of these two ways by their RMSE and R-square.

![Figure 1. Process flow diagram.](image)

3.1. Judgement Extraction
Information extraction (IE) is a product between natural language understanding technology and practical applications. Natural language processing has a good prospect to fundamentally solve the problem of man-machine dialogue. However, the current level of natural language processing does not allow for in-depth analysis of arbitrary texts and does not have the ability to understand natural language in depth. Unlike natural language comprehension, information extraction generally does not provide a thorough and comprehensive analysis of the text. Its main function is to extract specific types of information according to pre-defined tasks. To analyse the sentence corresponding to the criminal judgment, we extracted some keywords which have significant influences on the judgment result through the identification of the named entity.
• From the second paragraphs of the judgment, we can easily retrieve the defendant’s basic information and his crimes in the past. For example, we can retrieve the defendant's name, birthday, education, occupation, home address and basic criminal information.

• Generally speaking, from the paragraph beginning with "our court believes", we can just get more detailed crime information such as the type and quantity of drug involved in this case. Besides these, we can also determine whether the defendant played an important role in this case, whether he confessed everything he had done.

• The following paragraphs after the "judgement is as follows" are generally the result of the judgement of the persons involved. For example, we can know that the sentences of the two criminals are eight years' imprisonment and five years' imprisonment, and confiscated 8,000 yuan and 5,000 yuan in property.

3.2. Model 1: Feature Based Prediction

3.2.1. Feature Selection. Based on the rules of sentence, we choose a group of features which measures the punishment value of different suspects. According to our study from judgment description, two main feature categories are individual based features and case based features. Individual based features are like gender, is it principal criminal, if has criminal record, if plead guilty in court, and the case based features are the social influence, the number of drugs, if transport drugs and illegal cultivating original plants. We set the value to one if the condition is true and set value to zero otherwise. Finally, we get 14 features for prediction tasks.

3.2.2. Prediction Model. This section presents the key component of our framework. Specially, in order to predict punishment period more precisely, we do data preprocessing includes data cleaning, instance selection, normalization, feature selection, etc. Since there are many text keywords are treated as features, some redundant or irrelevant or hardly interpret features are removed, and choose subset as our final variable. Then we get the product of preprocessing as final training set. Several traditional classification and regression models are selected to prediction the punishment value as our competing results.

3.3. Model 2: Fuzzy Document Vector Based Prediction

Unlike traditional predictors which depending the effectiveness of selected attributes, document vector based model utilizes the text of judgment directly without human business. However, we still need some fuzzy operations for incomplete and imprecise text information. We first give the definition of document fuzzy mathematics and propose several fuzzy rules for judgments. And then, we use document vector model combined with K-Nearest Neighbor methods to predict the punishment. The main components of our framework are described as figure 2 shows.

![Figure 2. Process flow Diagram.](image)

Document Fuzzy Mathematics. In this section, we define the function of fuzzy operation like
\[ \mu_\tilde{A} : U \to U', u \mapsto \mu_\tilde{A}(u) \in U' \]

where \( \tilde{A} \) is a fuzzy set and \( \mu_\tilde{A}(u) \) is a membership function for mapping \( u \in U \) to a certain value \( u \in U' \).

According to the semantic of judgment context, we propose three different categories of membership functions.

- Placeholder mapping rule: Some special entities as defendant name, trade address, vehicle plate number and so on, which could influence the similarity of documents but useless to sentence. In this case, we use blank placeholder to replace this information. For example, information such as “Born in X City, Y Province”, “Transferring drugs by car with plate number X”, “Make an appointment with X in Y Hotel to deal drugs” will be replace with white spaces.

- Count mapping rule: Under some cases, the judgments describe the defendants’ detail criminal process, which could be summarized as “multiple trades”. Therefore, we should extract key concepts from implicit expressions in document. For example, “Deal drugs X times with amount of Y grams in total” will be replaced with “Trafficking drugs frequently/less frequently” and “Trafficking a lot of drugs / a bit of drugs”.

- Law mapping rule: Traditional membership function defines on a set \( X \) that indicates membership of an element in a subset \( A \) of \( X \). We use membership function not only based on probability theory, but also according to Chinese law. Thus we can replace the concrete period of staying in prison with fuzzy words. For instance, we replace “Under 3 years” with “short”, “3 years to 10 years” with “medium length”. And for amount of drugs, we can also process these numbers similarly. For example, “Below 3 years” can be replaced with “short”, “3 years to 10 years” can be replaced with “median length”, and “above 10 years” can be replaced with “long”. As for amount of drug, “below 7 grams”, “7 to 10 grams”, “above 10 grams” can be replaced with “a small amount”, “median”, “a great amount”, respectively.

After that, we get the fuzzy document without special entities and we could also use TF-IDF model to get the important factor of keywords and repeat these words in documents to modify the similarity measure value.

3.4. Prediction Framework with Document Vector

In this section, we develop a framework based on document vector, which is an unsupervised manner that learns continuous distributed vector representations for pieces of texts [13]. We treat fuzzed document as input model and then get the vectors of each document. For each word in document, we use CBOW model to maximize the average log probability

\[ \frac{1}{T} \sum_{t=1}^{T} \sum_{c=cs/jc} \log p(w_{t+j}|w_t) \]

where \( c \) is the size of the training context, the larger \( c \) value the higher accuracy. The \( p(w_{t+j}|w_t) \) could be defined by softmax function:

\[ p(w_{t+j}|w_t) = \frac{\exp ((v'_{t+j})^T V w_{t+j})}{\sum_{j} \exp ((v'_j)^T V w_{t+j})} \]

where \( V w_t \) is input vector representation of word \( w_t \), and \( (v'_{t+j})^T, (v'_j)^T \) are output vector representations of words \( w_{t+j}, w_t \).

Based on the word vector, the document vector framework map paragraph to a unique vector by concatenating or averaging the vectors of words in this paragraph. Moreover, we could use Convolution methods to reshape the output of vector.

We propose a document vector based prediction model as Algorithm 1 shows, in which for each given input Query, we first find several related document by their vectors. And then we use average value as our output predicted value. Since the special entities have been masked or transferred to other types of words, we could focus the relatively significant terms in judgments.
Algorithm 1: Document Vector based Prediction

Input: trained document vector set $M$, a document to be test $D$, the number of most similar documents to be compared $n$

Output: predicted period in prison $P$

1. $A = \text{empty list}$ //A stores $<\text{Document, Similarity}>$ pair
2. $V = \text{transformToVector}(D)$
3. $\text{foreach} m \text{ in } M$ do
4. add similarity($m, V$) to $A$
5. $A = \text{selectTop}(\text{sortBySimilarity}(A), n)$
6. $\text{sum} = 0$
7. $\text{totalWeight} = 0$
8. for $i = 0$ to $A$.length do
9. $\text{sum} = \text{sum} + A[i].\text{similarity} \ast A[i].\text{doc.period}$ //to get weighted average
10. $\text{totalWeight} = \text{totalWeight} + A[i].\text{similarity}$
11. $P = \frac{\text{sum}}{\text{totalWeight}}$
12. return

4. Result Evaluation

4.1. Experiment Platform
All the experiments are executed on a Dell server 64-bit system (16 core CPU, each with 2.6GHz, GPU GTX 1080ti, 32G main memory). The algorithms and models in our paper were implemented by Python 3. We use Scikit-learn library for competing methods in our experiments.

4.2. Dataset Description
We have collected 2478 criminal judgments involving drug-related cases. According to the Sentencing Standard for Drug Trafficking in the People's Republic of China, we have cleaned up the data. The specific operations are as follows: Firstly, we remove those judgments that do not meet the sentencing standards, for example, a suspect sells more than 10 grams of drugs, but his sentence is less than seven years. Abnormal data like that must be deleted to ensure the accuracy of our model. After that, we have 2578 defendants left, they are from 2214 judgments; Secondly, since the maximum sentence of data we have received is 15 years, all remaining defendants are divided into four groups according to the amount of drug trafficking: 0 grams (no drug trafficking), 0-7 grams (sentencing provisions should not exceed 3 years). 7-10 grams (sentencing provisions should be sentenced to three to seven years), 10-50 grams (sentencing provisions should be sentenced to seven to fifteen years). The number of words in judgments of different groups is shown as figure 3. We can see the judgment with more grams of drug trafficking, the more words it has.

Figure 3. Number of words in judgments of different groups.

In each group, however, we extracted 19 attributes. Since different defendants have different types of drugs, we must unify the drug information. According to the Criminal Law of the People’s Republic
of China, we can convert the grams of different drug types into grams of heroin. Now we have 17 attributes left, including name, sex, age, social influence, surrender to the police, whether the principal criminal, multiple drug trafficking, criminal record, drug-related recidivist, shelter others from taking drugs, confession of guilt, plead guilty in court, make contributions to solving the case, transport drugs, a lot of drugs, illegal cultivation of drugs, and the tag is the period of sentence.

4.3. Result Analysis

Baseline setup. Our method is compared with a variety of competing methods with ten traditional classification models: decision Tree (DT), support vector machines (SVM), linear regression (LR), k-Nearest Neighbor (KNN), Random Forest (RF), Adaboost, Gradient Boosting Decision Tree (GBDT), Bagging, Extra Tree (ET) and Xgboost. For all the methods with parameters, we optimize the parameters with 10-fold cross-validation by further dividing the training set into 70% for model fitting and 30% for parameter validation.

Evaluation metrics. We use root mean square error (RMSE) and R Squared computed with test set to evaluate the performances of different methods. The RMSE is a frequently used measure of the predicted value and observed value. The RMSE is defined as:

$$RMSE = \sqrt{\frac{1}{m} \sum_{i=1}^{m} (y_i - \hat{y}_i)^2}$$  \hspace{1cm} (3)

where $m$ is number of instances, $y$ is the observed value and $\hat{y}$ is the predicted value.

R-squared ($R^2$) is a statistical measure that represents the proportion of the variance for a dependent variable that’s explained by an independent variable or variables in a regression model. The R-squared is defined as:

$$R^2 = 1 - \frac{MSE(y, \hat{y})}{\text{var}(y)}$$ \hspace{1cm} (4)

$$MSE = \frac{1}{m} \sum_{i=1}^{m} (y_i - \hat{y}_i)^2$$ \hspace{1cm} (5)

$$\text{var}(y) = \frac{1}{m} \sum_{i=1}^{m} (y_i - \bar{y})^2$$ \hspace{1cm} (6)

The large value of R-squared presents the two data points are high correlated.

We first analyse the result of RMSE in table 1, which is the mean of the square root of the error between the predicted and true values. The results show that the linear regression has the lowest RMSE among all kinds of traditional machine learning models. Maybe a lack of attributes can explain this phenomenon. On the contrary, our model FDV have lower RMSE values in each group, especially in the second and the fourth groups, which are 15% lower than the lowest RMSE of traditional machine learning.

| Algorithm | Group-1 | Group-2 | Group-3 | Group-4 |
|-----------|---------|---------|---------|---------|
| DT        | 3.3921  | 9.4066  | 11.5117 | 28.8151 |
| SVM       | 2.4852  | 6.9463  | 8.7788  | 21.3231 |
| LR        | 2.1685  | 6.1714  | 7.8960  | 19.5557 |
| KNN       | 2.4606  | 7.1802  | 8.7388  | 21.5606 |
| RF        | 2.9671  | 8.0555  | 9.0046  | 22.0731 |
| Adaboost  | 2.7294  | 7.3016  | 8.6990  | 21.3280 |
| GBDT      | 2.7785  | 7.4723  | 8.5394  | 21.7171 |
| Bagging   | 2.9669  | 8.2592  | 9.4277  | 22.4691 |
| ET        | 3.3298  | 9.4401  | 11.5121 | 28.5724 |
| Xgboost   | 3.2923  | 10.8801 | 9.3471  | 39.2583 |
| FDV (our model) | **1.96** | **5.25** | **6.25** | **16.61** |

Table 1. A performance comparison by RMSE.
As for R-square (table 2), it describes how well the model predictions are better than the average. We can see that the linear regression method is still the best in the traditional machine learning method, but our model FDV has an obvious advantage. It is also worth noting that many R square indicators in the traditional machine learning model are negative, which means that the model has no effect at all. Therefore, in terms of text prediction, if you only use the named entity to extract the relevant attributes for prediction, there will be a great limitation. And the word vector representation based on deep learning can effectively avoid this problem.

| Algorithm | Group-1 | Group-2 | Group-3 | Group-4 |
|-----------|---------|---------|---------|---------|
| DT        | -0.1640 | -0.3705 | -0.6514 | -0.5520 |
| SVM       | 0.1338  | 0.0671  | -0.0217 | -0.0074 |
| LR        | 0.3239  | 0.2166  | 0.0814  | 0.0689  |
| KNN       | 0.1361  | -0.0197 | -0.0440 | -0.0411 |
| RF        | 0.1257  | 0.0502  | -0.1009 | -0.0214 |
| Adaboost  | 0.1604  | 0.1262  | 0.0145  | 0.0552  |
| GBDT      | 0.2083  | 0.1731  | 0.0354  | -0.0306 |
| Bagging   | 0.1600  | -0.0462 | -0.1050 | -0.0990 |
| ET        | -0.0755 | -0.3746 | -0.5605 | -0.6557 |
| Xgboost   | 0.0798  | 0.1183  | 0.0068  | 0.0372  |
| FDV (our model) | 0.3756 | **0.2615** | **0.2727** | **0.1782** |

5. Conclusion
Prediction on text is not a trivial task for its ambiguous and rich semantic meanings. Traditional methods should extract features based on external knowledge. In this paper, however, we creatively leveraged such text data for prediction in the public judgments without human business. We propose a combined framework to capture the prediction problem by considering both valued based rules and fuzzy document models. This framework contains the complete process as information extraction, term fuzzy and document vector regression. We setup experiments on a real-world dataset and compare our model with traditional classification and regression methods. Our further work is to estimate the effectiveness of each term in document and learn how to modify the similarity value by weighted terms in document for higher accuracy.

References
[1] XINHUANET 2018 http://www.xinhuanet.com/english/2017-03/27/c_136161743.
[2] Supplementary Document 2018 http://www.unodc.org/doc/wdr2018/WDR_2018_Press_ReleaseENG.PDF.
[3] Gao Y F, Yu W Z, Chao P F, et al. 2015 Analyzing reviews for rating prediction and item recommendation Journal of East China Normal University (Natural Science) (3) 80-90.
[4] Hu M Y 2016 Analysis of Movie Rating Data and User Preference Model (Yunnan University).
[5] Mu Y L, Zhou C H, Yu D J, et al. 2016 A rating prediction framework based on distributed representation of document and regression model Computer Era (5) 24-29.
[6] Athan T, Boley H, Governatori G, et al. 2013 OASIS legal RuleML Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law pp 3-12.
[7] Valente A, Breuker J 1994 Ontologies: The missing link between legal theory and AI & law Legal Knowledge Based Systems 94 (2) 138 -150.
[8] McCarty L T 1989 A language for legal discourse Proceedings of the 2nd International Conference on Artificial Intelligence and Law (Washington: ACM) pp 180-189.
[9] Valente A 1995 *Legal Knowledge Engineering: A Modelling Approach* (Netherlands: University of Amsterdam).

[10] Varzi A C and Vieu L 2004 Core concepts of law: Taking common sense seriously *Proceedings of the Third International Conference* (Amsterdam: IOS Press) pp 210-223.

[11] Hoekstra R 2009 Ontology representation design patterns and ontologies that make sense *Proceedings of the 2009 Conference on Ontology Representation* (Amsterdam: IOS Press) pp 1-236.

[12] Quan T, Hui S and Cao T 2004 A fuzzy based approach for citation-based document retrieval *IEEE Conference on Cybernetics and Intelligent Systems* pp 578-583.

[13] Gao Y, Zheng B, Chen G, et al. 2010 Efficient mutual nearest neighbor query processing for moving object trajectories *Information Sciences* 180 2170-2195.

[14] Liu Y, Wang N, Li B H and Luo Q 2017 Application of fuzzy grammar method in crime text classification *Computer Engineering and Design* (7) 1965-1971.

[15] Hu J Z, Shu J B and Huang Y Y 2008 Text feature extraction based on the extension of topic words and fuzzy set *Proc. of International Conference on Computer Science and Software Engineering* (IEEE Press).

[16] Hinton G E 1986 Learning distributed representations of concepts *Proceedings of the Eighth Annual Conference of the Cognitive Science Society* (New Jersey: Lawrence Erlbaum Associates) pp 1-12.

[17] Xi X F and Zhou G D 2016 A survey on deep learning for natural language processing *Acta Automatica Sinica* 42 (10) 1445-1465.