Step-wise Refinement Classification Approach for Enterprise Legal Litigation

Ying Mao , Xian Wang , Jianbo Tang , Changliang Li[*]
Kingsoft
{maoying, wangxian, tangjianbo, lichangliang}@kingsoft.com

Abstract

In the field of finance and lawsuit, data mining technology has absolute broad market prospect but also is a challenging task. Past years have witnessed great successes of data mining in finance and lawsuit related applications. Most existing work usually focus on providing litigation risk assessment and outcome prediction services for the clients. However, the research on the legal litigation type for enterprise is limited. In this paper, we focus on enterprise lawsuit category prediction and propose a novel approach to refine the problem as a classification task. First, We evaluate the possibility distribution of legal documents received by the enterprise, then distinguish the specific legal litigation type. We apply our method on *International Big Data Analysis Competition*[1] launched by *IEEE ISI Conference 2019* and scored the first place in the final leader-board.

1 Introduction

Artificial intelligence is developing very rapidly in today's society, so we hope to introduce it into the legal field. From the rule-based expert legal system (transforming legal experts' legal knowledge and experience into a computer language in the form of rules) to the autonomous system supported by deep learning, machine learning, and big data, the deeper and broader impact of artificial intelligence on the legal industries has only just begun. If we can extract valid information from the company's legal information to predict the type of received legal documents, our work can not only help the company prepare for litigation in advance, but also provide early warnings of its operating status.

Litigation/Lawsuit[2] is a kind of legal action, which is divided into civil and criminal categories. The former plaintiff is the victim's party, and the law is resorted to because there are unresolved disputes. The latter involves criminal offences, and the government authorities sue the suspects (prosecution). The proceedings are divided into first and second instance, and may also be final. A lawsuit may involve dispute resolution of private law issues between individuals, business entities or non-profit organizations. The function of litigation is not limited to the discovery of historical facts that have occurred in the past, but also establishes the link between fault and responsibility, crime and punishment through the process of litigation, thereby conveying to citizens a message of how to behave and accountability.

Enterprises would receive different type of legal litigation documents for various reasons in the business process, such as tax evasion, arbitrary discharge of industrial sewage, arrears of payment, etc. Such information which can be obtained from the legal information report of the enterprise is generally historical time information. The *IEEE ISI Conference 2019* launched the *International Big Data Analysis Competition* and one of the objective aims to predict the type of legal documents that are most likely to be received by companies.

In this paper, we propose a new approach using step-wise refinement strategy to predict the top two most possible legal litigation types for each publicly traded company. In the first step, considering whether the legal documents is received by enterprise, we introduce the binary-classification model. After evaluate the possibility distribution of received legal documents, in the second step multi-class classification processing will then be performed to identify the specific legal litigation type. During these steps, the models refine more detailed information from the output produced by last step. We apply our method on *International Big Data Analysis Competition* and score the first place in the final leader-board. As it turns out, achieving the remarkable result reveals our step-wise refinement approach significantly outperforms state-of-the-art techniques in legal litigation prediction field.

2 Related Work

The traditional solution to the task of legal type prediction is manual judgment, which means experienced professionals draw conclusions based on relevant data. The main analysis process is generally based on the analysis of the number, regional distribution, and subject matter of the litigation cases that the company participates as a litigation participant. Through analyse the corresponding data, the main types of litigation cases of the corresponding enterprises in the earlier period of the data acquisition period can be obtained with a high probability.

[*] the corresponding author

[1] http://www.linkx.ac.cn/#/title

[2] https://en.wikipedia.org/wiki/Lawsuit

Proceedings of the First Workshop on Financial Technology and Natural Language Processing
(FinNLP@IJCAI 2019), pages 27-33, Macao, China, August 12, 2019.

The first part is the analysis of the litigation case types in the history of the company. The major categories of litigation cases are generally criminal, administrative and civil. In fact, most enterprises is main body of operation in our country and their main business activities should be based on business behaviors of civil and commercial activities, so the development of disputes and types of litigation are mostly civil and commercial litigation. When civil cases are further subdivided, they can be distinguished by referring to the provisions of the *Supreme People's Court* on civil cases. The civil case generally represents the problems existing in the management methods or business concepts, and may also reflect its operating status and some industry characteristics to a certain extent. If there are a large number of cases of product quality disputes, it indicates that the company may have flaws in the process of product quality control. If there are a large number of labor dispute cases, it indicates that the company may have problems such as daily violation of labor contract law management and large mobility of industry personnel due to poor management. If an enterprise has an administrative case or a criminal case, there are serious illegal activities on behalf of the company's production and management behavior, such as meat product enterprises that have been administratively recalled due to quality problems, and vaccine companies that have been investigated by criminals. Once this type occurs and a case with extensive public opinion influence is formed, it will often have a greater impact on the sustainable operation of the enterprise. It may also result in a series of civil cases, such as quality claims, personal injury claims, shareholder lawsuits, etc. It can be considered that if there is no major change in the management or strategy of the company, in addition to the impact of the phased factors, the above situation may have a certain continuity, thus which also can help to predict the trend of the company's frequent cases.

The second part is the analysis of the area where the case occurred and the location of the main business of the company. The occurrence area may reflect the market share of the business operations in the relevant regions. Generally speaking, the outbreak of the case has a certain positive correlation with the market share of the corresponding region. The larger the business volume, the more the problem will be. Sometimes it can also indicate that there is a problem with the local management and operation of the company, which leads to a concentrated outbreak of controversy.

The third part is the analysis of the subject matter. Generally speaking, in the relevant cases, the greater the type of the subject, the stronger the impact on the business operation. It means that the management problems of the relevant parties may be more prominent.

3 Task Definition

The *International Big Data Analysis Competition* subject aims to predict the top two most possible legal litigation types for each publicly traded company according to companies' full size information.

For this competition, the organizer provides 18 data forms[3], including multi-dimensional information of publicly traded companies in the past years, which can be seen from Figure. 1. These data are from official statistic platform. There is one unique ID for each company.

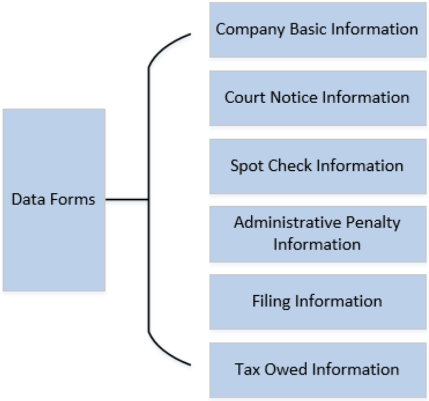


Figure 1: Data forms

The organizer has anonymized, discretized, and normalized some of the information for the protection of data such as corporate secrets, intellectual property rights, and debt information. The court notice is the core form, and the organizer has anonymized the types of legal documents in this form. The types of legal documents processed are expressed as: A-Q and others. As shown in Figure. 2, there is a wide disparity in the amount of different types. We need to extract the top two most possible legal litigation types for each company from this data form.

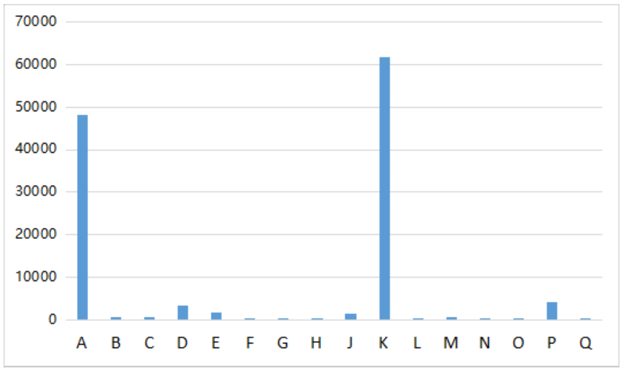


Figure 2: Distribution of the legal litigation type

They provide 3,000 companies as training data sets (actually about 1,500, with about 700 effective data). The company name is anonymized and expressed by enterprise number between 1001 and 4000. The data in some of the tables is incomplete, and the companies provided may not all be used as training sets, and flexible data processing is required. Also the organizer provides about 500 companies as test data sets (actually about 200) with enterprise number between 4001 and 4500. We should calculate the types of legal documents that these companies may receive the most and the second most.

[3] https://pan.baidu.com/s/1YaJcn-lpWVN2TSFDUFLwYw

4 Approach

In this section, we will present end-to-end procedures of our strategy. This work consists of four different parts which are preprocessing, feature engineering, model establishment and ensemble. The overall procedure is illustrated in Figure. 3.

We found that splitting the original problem into three sub-problems can help to better fulfil the forecasting task. The three sub-problems aim to evaluate whether the enterprise has received the legal documents, the type of legal document that the enterprise receives the most, and the type of legal document that the enterprise receives the second most. The training data required for the three sub-problems are not the same, moreover the three will affect each other. Therefore, the first strategy aims to train the model separately, and then combing the three classification algorithms, according to the results.

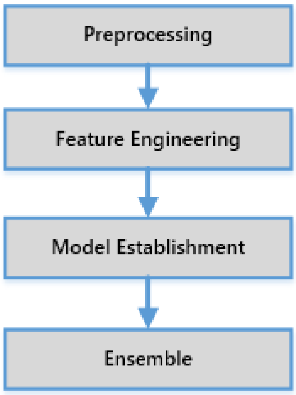


Figure 3: Overall procedure

4.1 Preprocessing

The basic principles of preprocessing for our raw data consist of five parts. Firstly, all data have to be combed on multiple time granularity, for instance we not only count the number of civil cases received by the company in the last year but also in the last two years etc. Secondly, if there are several different time stamp for same data, they should be sorted out specifically. Thirdly, the original data need to be eliminated redundancy before it is counted. Fourthly, there are plenty of data related to capital provided by organizer which need to be reprocessed carefully. For example unified unit of money, count stocks and cash respectively. Fifthly, in order to handle the missing values problem we should attempt to use different strategies as much as possible such as mean, mode, median and so on.

Simplifying prediction labels is one of our special treatment. As shown in Figure. 4, we find there are unbalanced distribution of legal litigation type. The sum of proportion of type A, D and K is over 90%, on the contrary some labels like type C, F, O and Q are less than 1%. To deal with this situation, some labels have been considered as exception value if their proportion are less than 1% and some labels have been combined into one if their proportion are larger than 1% and less than 5%. The same strategy is applied to organize the second most legal litigation type data.

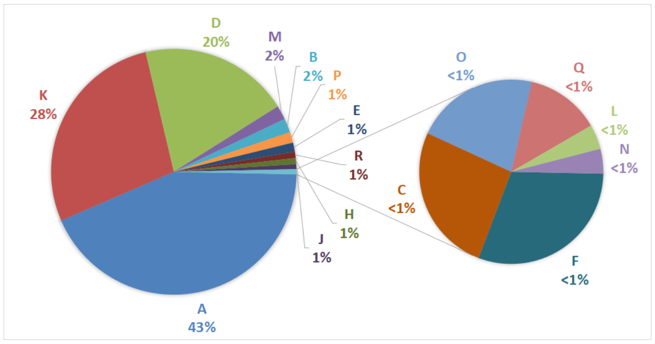


Figure 4: Distribution of the most legal litigation type

Another special treatment we have applied on the raw data is using external knowledge. When we simplify the information about administrative penalty, using the external knowledge make the entire preprocessing process more rational. In the raw data, the descriptions about context of administrative penalty for company every year are different and complicated. After reviewing the relevant regulations on administrative punishment, we make a summary of the descriptions about administrative punishment. Extracting important information from the complicated description, and based on this we roughly divided them into new categories, such as commodity trademark related, computer infringement related, illegal occupation related and so on.

As shown in Figure. 5, the presence of outliers in the data sets can be easily confirmed. Dealing with outliers is also a very important part of data preprocessing. The key to this procedure is eliminate outliers while keeping adequated data volume. Because the amount of relevant data provided by the official is limited, it is very important to be careful when choosing the right method to ensure that do not eliminate too much data. By comparing several different common methods, we choose to detect outliers by artificial experience and PauTa Criterion principle[Li *et al.*, 2016], and at the end the problem of outliers has been solved partly.

4.2 Feature Engineering

After preprocessing, some important features which are not mentioned in the official material need to be constructed. There are three methods have been used to construct the features, basically the construction of features is designed by data characteristics.

First, considering that the growth rate of important indicators in the short term can predict the company's changes in a certain area, trend features have been made by assessing the growth of data in the current two years relative to the company's previous data. Such as for the data sets of the judicial documents, the work to calculate the average number of civil cases of each company in 2016-2017 and the number of civil cases in 2015 needed to be done simultaneously. Soon afterwards, using the previously obtained data, the growth rate of civil cases can be calculated from the period of 2016 to the end of 2017 relative to the previous year.

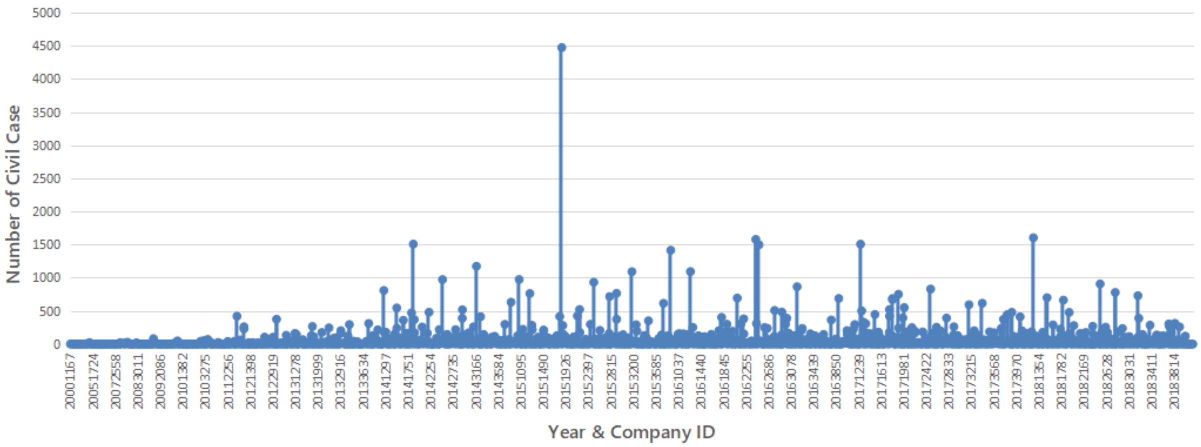


Figure 5: Distribution of number of civil case

Second, considering the differences between regional policies and the problems and challenges faced by various industries, statistical features need to be created to improve system classification capabilities. The main method for constructing statistical features is count the ranking information between different regions, cities, industries, and sub-sectors as well as the proportion of prediction labels, according to the average registered capital and other information. The details are as shown in Figure. 6 , there is significant difference between industries for whether the legal documents will be received. Based on empirical analysis, our team believes that the risks of receiving legal documents in diverse industries are different. The traditional industries with mature product quality control and standardized personnel management procedures are less threatened by legal documents, on the other hand new industries are still at the stage of exploration, management is not standardized, and legal instruments are more likely to be received. Similarly, the difference of scale of enterprise may also cause different preferences of legal litigation type. This is also why we construct the feature for describing enterprise scale by using registered capital and number of employees.

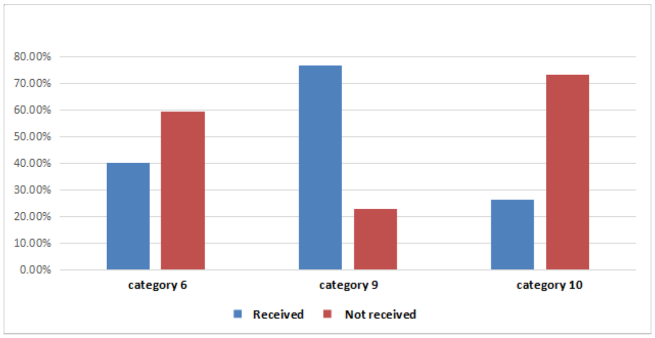


Figure 6: Distribution of whether to receive legal litigation in different industy category

Third, due to constructing polynomial features is one of the common methods for constructing new features, besides it is widely used in statistical models to explore the effect of compound variables on y. Based on experience, we use second-order polynomial features to enrich the feature sets.

4.3 Model Establishment

Through consulting literature and investigation, four models enter the candidate program which are LightGBM[Ke *et al.*, 2017], Xgboost[Chen and Guestrin, 2016], RandomForest[Svetnik *et al.*, 2003] and Neural Network[Hansen and Salamon, 1990]. After experimental verification, the results show that LightGBM, XGBoost work best for this problem.

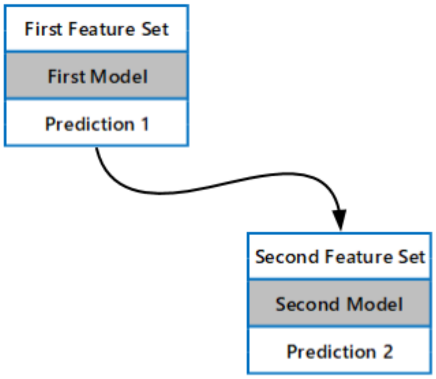


Figure 7: Single classification model for multi-class

To better solve the classification problem, a novel stepwise refinement strategy is designed. First, the problem is composed of one binary classification problem and two multi-class classification problems. The binary classification model is used to predict whether a company will receive legal documents over a period of time by LightGBM model. The multi-class classification models are used to predict the specific legal litigation type by LightGBM and Xgboost models. The detailed procedure for multi-class classification is to predict the type of legal litigation that the company will receive the most firstly, then the predicted result need to be added as a feature to the input feature sets of the second model. At the end, the second model is used to predict the type of legal litigation that the company may receive the second most. The final results are derived from the binary classification results and the multi-class classification results. The details are shown in Figure. 7.

For the LightGBM and XGBoost models of the multi-class classification problem, 10 models are constructed by 10 fold cross-validation, and the final prediction value is obtained by the voting mechanism as the single model final prediction result. This action improves model accuracy and reduces overfitting. The details are shown in Figure. 8.

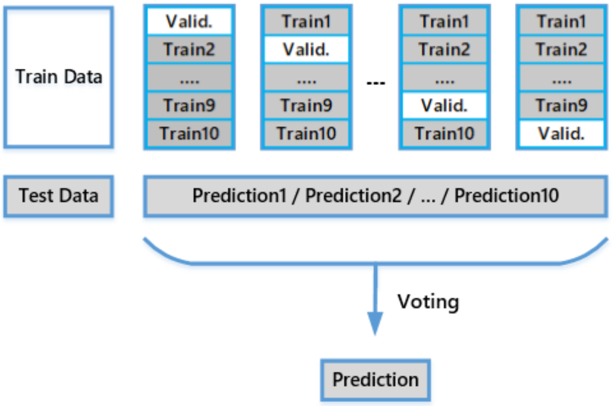


Figure 8: 10-fold cross validation

One challenge is that LightGBM and XGBoost models contain some super-parameters, and the selection of parameters have a certain impact on the classification results. Commonly used parameter search methods include: Random Search[Bergstra and Bengio, 2012], Grid Search and Bayesian Search[Feurer *et al.*, 2015]. After many attempts, Bayesian search is finally selected due to its superior performance.

4.4 Ensemble

In real-world cases, training models to generalize on datasets can be a very challenging problem as it can contain many underlying distributions. Certain models will do well in modelling one aspect of this data while others will do well in modelling the others. Ensemble provides a solution where we can train these models and make a joint prediction where the final accuracy is better than each of the individual models.[4]

There are several commonly used ensemble methods, such as voting[Rokach, 2010], averaging, and stacked generalization[Williams and Gong, 2014] and blending. After many experiments and verification applied in the multi-class classification model task, voting is finally adopted. As shown in Figure. 9, in order to ensure the effect of model ensemble, there are differences in the input feature sets of the four multi-class classification models. All of the feature sets which are used are selected based on their performance in single model.

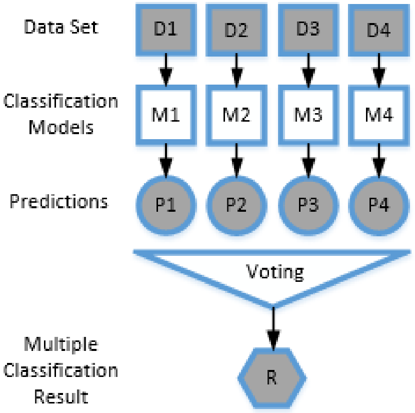


Figure 9: Model ensemble

As shown in Figure. 10, our final result is coupled by the prediction results of the binary classification model and all the multi-class classification models.

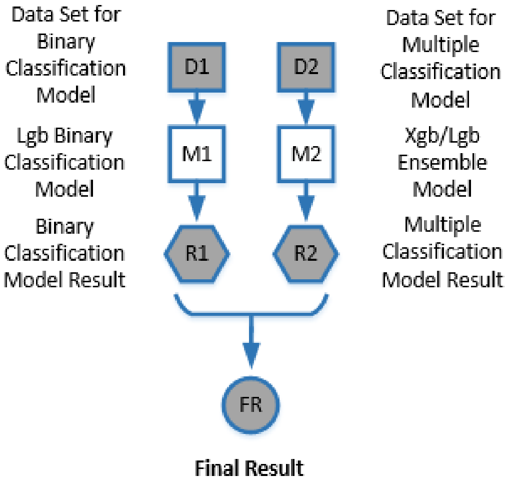

Data Set for Binary Classification Model
D1
D2
Data Set for Multiple Classification Model
Lgb Binary Classification Model
M1
M2
Xgb/Lgb Ensemble Model
Binary Classification Model Result
R1
R2
Multiple Classification Model Result
FR
Final Result

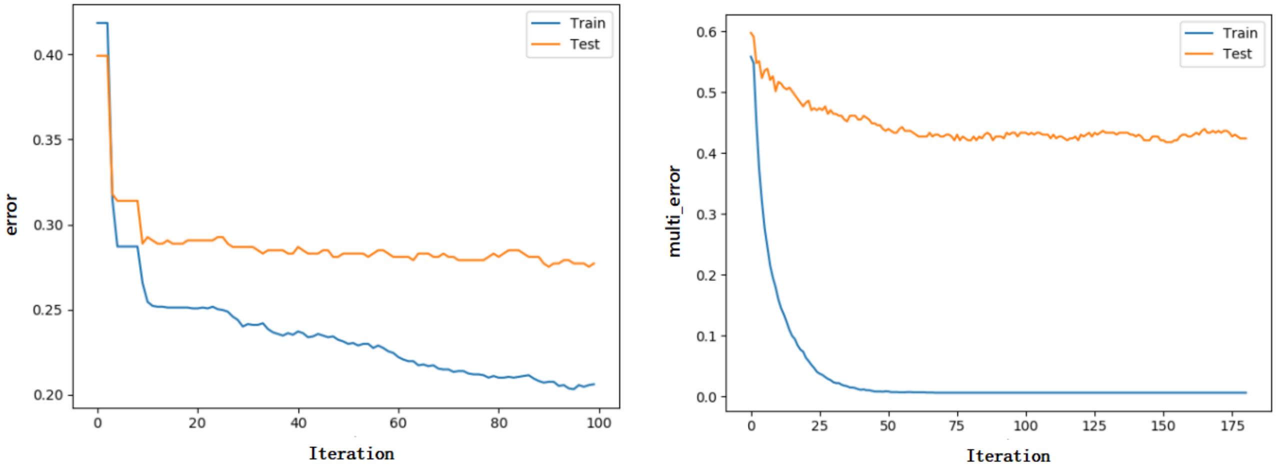
Figure 10: Combination of binary model and multiple model

The final strategy we have adopted when merging the results of the binary classification result and multi-class classification result can be summarized as: firstly, if the prediction result of binary classifications is positive, then focus on the multi-class classification result, otherwise it is classified as negative. Secondly, if the multi-class classification result is displayed as others which is created when we have merged some labels, further prediction is performed according to the distribution of the original data sets, otherwise it is recorded as the original multi-class classifier result.

5 Experiments

In this section, the final result will be presented. The experiment environment we used is Python 3.6 and calls tools such

[4] https://medium.com/weightsandbiases/an-introduction-to-model-ensembling-63effc2ca4b3



(a) binary

(b) multiple

Figure 11: Training and testing curve for classification

as pandas 0.23.0, sklearn 0.19.1, lightgbm 2.2.3, etc.

After data preprocessing, the data volume of the multi-class classification model is about 3200. On the contrary, the input volume for binary model is more than 4,700 and the validation set data volume is 133. The number of features for binary model and four multi-class model are all around 20 to 25. Negative samples of relatively long time (such as 2011-2012, 2012-2013) can not fully indicate that the company did not receive legal documents during the year, but because of the loss of information and the incomplete statistics. By counting the proportion of positive and negative samples from 2011 to 2018, our team found that the proportions fluctuated significantly, which clearly confirmed our conjecture. Therefore, in order to eliminate the influence of observation errors of these historical data, we only used the negative samples of 2017-2018 to join the training set.

The training/testing curve for binary and multiple classification tasks shows in Figure. 11. The two figures show that the prediction accuracy of the binary classification model as well as multiple classification models all converge to a high level.

The organizer provides an evaluation index algorithm for the model effect. For every enterprise, if the model predicts the correct results for both the most legal litigation type and the second most legal litigation type, then the score for this enterprise should be the max score, which is 100. If the model predicts the correct result for the most legal litigation type but the wrong result for the second most legal litigation type, then the score for this enterprise should be a small amount of max score, for example 35. The final score of the test datasets should be the average score of all the enterprises.

Finally we scored 44 points and achieved the first place in the competition.[5] The excellent score of the competition also shows the superiority of the model in this legal litigation type prediction field.

6 Conclusion

In this paper, we presented a step-wise refinement classification approach for legal litigation prediction which combines binary and multi-class classification model based on parameter search and ensemble methods. The innovation for this design is to evaluate the possibility distribution of legal documents received by enterprises before further distinguish the specific type of legal litigation. When evaluated on officially provided datasets, our model significantly outperforms all the couterparts.

While our approach provides substantial performance gain, there is still room for improvement. In the future, we would like to discover more distinguishing features and illustrate the interpretation of the proposed model.

References

[Bergstra and Bengio, 2012] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(Feb):281–305, 2012.

[Chen and Guestrin, 2016] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794. ACM, 2016.

[Feurer *et al.*, 2015] Matthias Feurer, Jost Tobias Springenberg, and Frank Hutter. Initializing bayesian hyperparameter optimization via meta-learning. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[Hansen and Salamon, 1990] Lars Kai Hansen and Peter Salamon. Neural network ensembles. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (10):993–1001, 1990.

[Ke *et al.*, 2017] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and

[5] http://www.linkx.ac.cn/#/ranking

Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, pages 3146–3154, 2017.

[Li *et al.*, 2016] Limin Li, Zongzhou Wen, and Zhongsheng Wang. Outlier detection and correction during the process of groundwater lever monitoring base on pauta criterion with self-learning and smooth processing. In *Theory, Methodology, Tools and Applications for Modeling and Simulation of Complex Systems*, pages 497–503. Springer, 2016.

[Rokach, 2010] Lior Rokach. Ensemble-based classifiers. *Artificial Intelligence Review*, 33(1-2):1–39, 2010.

[Svetnik *et al.*, 2003] Vladimir Svetnik, Andy Liaw, Christopher Tong, J Christopher Culberson, Robert P Sheridan, and Bradley P Feuston. Random forest: a classification and regression tool for compound classification and qsar modeling. *Journal of chemical information and computer sciences*, 43(6):1947–1958, 2003.

[Williams and Gong, 2014] Trefor P Williams and Jie Gong. Predicting construction cost overruns using text mining, numerical data and ensemble classifiers. *Automation in Construction*, 43:23–29, 2014.