A feasible method for training classified data with sparsity

Jingsong Zhu, Bing Guo\textsuperscript{1}, Yingting Luo and Jianxin Pan

College of Mathematics, Sichuan University, Chengdu 610063, China

\textsuperscript{1}Corresponding author’s e-mail: guobingm@qq.com

Abstract. Family harmony is an important part of social stability and harmony. Facing the rapid growth of divorce disputes, how to hear the cases quickly and fairly is an urgent matter to be solved. Based on judgment documents from courts located in Southwest of China, this paper studies how to evaluate and predict whether or not a case should be divorced. These documents are characterized by their unbalance and sparsity; besides that, most of variables are bearing with missing value. We propose a feasible method that shows high accuracy on both training and testing datasets. Concisely, oversampling and “clustering” are exploited to data pre-preparation, and recursive feature elimination is applied to deal with variables selection. In this paper we combined Random Forest and XGboost to derive a more feasible and precise model that achieves ninety percent of accuracy.

1. Introduction

With development of China’s economy and improvement of quality of individual life, the divorce rate across China has obviously increased ([1], [2]). Marriage, as a form of family, has a special significant meaning to human society, but divorce, in the contrast, poses a threat to the harmony and stability of family and society [3]. Facing the rapid growth of divorce disputes, how to handle the cases quickly and fairly is an urgent matter to solve. A brunch of works related to divorce have been done. For example, [4], [5], [6] study the factors that affect the divorce rate, and [7], [8], [9] consider how to judge whether a marriage has broken down, and so on.

The main aim of this research is to formalize a method using an extension of machine learning and statistical learning and learn how to collect, process, and analyze the characteristics of pairs to assist court judges. Moreover, the core challenges of the current judgment lie in two aspects: one highlights efficiency of coping with increasing number of cases, and the other is subsumed by prediction accuracy. Through end-to-end machine learning, the approach proposed by this research succeeds to handle a large amount of data and achieve a rapid divorce judgment with ninety percent of accuracy.

Machine learning and artificial intelligence become more and more essential in various fields ([10], [11]). At present, recommendation algorithms can promote advertising to target costumer; intelligent mail classifiers are conducive to blocking spam messages; even in the go chess, deep learning dominates the whole game. The success of these applications is inseparable from two aspects: one is the use of efficient statistical models to extract connections between data, and the other is to make discriminant analysis through large-scale learning systems.

Machine learning algorithms can be generally divided into three classes: supervised learning, unsupervised learning and reinforcement learning [12]. More specific, supervised learning yields a model to derive statistical inference, unsupervised learning analyzes the characteristics of the samples directly, of which the most typical representative is cluster analysis, and reinforcement learning optimizes model along with different circumstances. This paper mainly discusses the application of
supervised learning in divorce judgments by collecting information on past divorce cases, analyzing the identities and behaviors of the pairs, and establishing a feasible model through machine learning methods.

Supervised machine learning is prevalent in the family of machine learning. Its history can be traced back to 1936. Fisher [13] proposed a linear discriminant method which finds a projection to convert high-dimensional problems to one-dimensional problems. Furthermore, based on Bayesian theory, Bayesian classifier was proposed in the 1950s. The main idea is to implement decision-making method under frame of probability, that is, to select the best categorizes using known probability and loss function. Later on, Reed and Pearl [14] discovered the Logistic curve and applied it to simulate population growth. This curve starts to grow slowly, and then grows rapidly at a certain time. After reaching limit, the growth slows down again, showing an elongated "S" shape. Immediately, Berkson [15] proposed the Logit model providing a chance to efficiently predict the probability of a sample belonging to a positive sample. In the next few years, various machine learning methods were proposed such as decision trees, perceptron models, neural network algorithms and so on. Until the 1990s, two well-known methods, Support Vector Machine (SVM) and AdaBoost [16], came out, which not only brought ensemble learning come into people’s attention, but also made a huge leap in efficiency and accuracy. At the same time, Leo Breiman [17] integrated several weak classifiers (decision trees), and proposed Random Forest Model, which, referring to the ideas of bagging and random selection, builds multiple decision trees and fuses their prediction under a benchmark.

In recent years, Chen and Guestrin [18] has improved the gradient boosting tree at the level of algorithms and engineering and proposed XGBoost model. XGBoost is a large-scale integrated learning system based on the Boosting Tree Model. It is widely used in different fields including but not limited to text recognition, behavior prediction, software detection and survival risk prediction. It is well-known for efficiency and accuracy in all aspects of applications. For example, in the realistic event-based modelling competitions held by Kaggle in 2015, 17 of the 25 winning teams directly or indirectly take advantages of XGBoost. Why its advantages are so prominent? In fact, compared with other commonly-used machine learning methods, XGBoost is improved in the following aspects: Firstly, it optimizes non-contiguous caches, and combines Column Block Parallel Computing and Out of Core Computation which strengthens its ability to train large-scale data. Secondly, it introduces three methods, column sampling, regularization, and weight-based compression, to prevent model from overfitting. Third, Sparsity-aware Split Finding, which can automatically learn the default classification direction of missing values, helps to process sparse data.

2. XGBoost model
XGBoost is an end-to-end boosting tree algorithm [18]. It builds a regression tree (a weak classifier), and then use another one to model its residual. Repeating this operation to derive a strong classifier:

$$\hat{y}_t = \sum_{k=1}^{K} f_k(x_i),$$

where $f_k(x_i)$ is a regression tree. According to the principle of minimum loss, regularization is imported to determine the direction of regression tree splitting:

$$L_t = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{t-1} + f_t(x_i)) + \gamma T + \mu \frac{1}{2} \| w \|^2$$

$$= \sum_{i=1}^{n} [ l(y_i, \hat{y}_i^{t-1} + f_t(x_i)) + g_t f_t(x_i) + \frac{1}{2} h_t f_t(x_i)^2 ] + \gamma T + \mu \frac{1}{2} \| w \|^2,$$

where $l$ is a differential loss function measuring the differences between prediction $\hat{y}_t$ and true value $y_t$, $\hat{y}_i^{t-1} + f_t(x_i)$ denotes $t$-th strong classifier after adding a regression tree $f_t(x_i)$, $g_t$ and $h_t$ are first and second order derivatives, $T$ is number of leaves in $t$-th regression tree and $w$ accounts for weight of each leaves.
Omitting the constant term and making some mathematical transformation, it is easy to derive that the optimal loss for variable-splitting.

\[ \hat{L} = \frac{1}{2} \sum_{j=1}^{T} \left( \sum_{i \in I_j} g_i \right)^2 + \sum_{i \in I_j} h_i + \mu \gamma T, \]

where \( I_j = \{ i | q(x_i) = j \} \).

Then, transform each partition of each variable, split in turn, find the optimal loss, and compare the optimal loss before and after the split. Define

\[ \text{Gain} = \frac{1}{2} \left[ \sum_{j=1}^{T} \left( \sum_{i \in I_L} g_i \right)^2 + \frac{1}{2} \sum_{j=1}^{T} \left( \sum_{i \in I_R} h_i \right)^2 - \frac{1}{2} \sum_{j=1}^{T} \left( \sum_{i \in I_L} h_i + \mu \right) \right] - \mu. \]

When “Gain” is greater than a particular threshold, then split, otherwise, stop splitting and start to consider the next partition.

3. Data preparation

3.1. Data description

The data of this research comes from the divorce judgments of southwestern courts in China from 2016 to 2018. It provides with basic personal information (“Gender”, “Age”, etc.) of the pairs, facts of the case (“Whether they live together”, “Whether they have pre-marital property”, etc.), Information about child custody and maintenance fee (“Whether their children born in wedlock”, etc.), and dismemberment of property (“Request to divide real estate”, “Request to divide vehicles”, etc.), forming a sample of 5395 divorces with 103 variables, and parts of our data are listed in Table 1.

| Class                          | Name                        | Category                  |
|-------------------------------|-----------------------------|---------------------------|
| Personal Information          | Gender of plaintiff         | Categorical variable      |
|                               | Age of plaintiff            | Continuous variable       |
|                               | Education of plaintiff      | Categorical variable      |
|                               | ...                         |                           |
| Fact of the Cases             | Whether they live together  | Categorical variable      |
|                               | Whether they have pre-marital property | Categorical variable |
|                               | Evaluation of premarital feeling | Categorical variable |
|                               | ...                         |                           |
| Information about child custody and maintenance | Whether their children born in wedlock | Categorical variable |
|                               | Living cost                 | Continuous variable       |
|                               | ...                         |                           |
| Dismemberment of property     | request to divide real estate | Both                      |
|                               | request to divide vehicles  | Both                      |
|                               | ...                         |                           |
3.2. Sparsity
There are three types of deletions: completely missing at random, missing at random and non-random missing. Completely missing at random means that it neither depends on its own value nor the values of other variables, such as the “age of the plaintiff” in the data. Missing at random refers to the fact that the missing value indeed has a certain connection with other variables, but it is independent with its own value. For example, there is a connection between the “separation time” and “whether there is a separation”; the last non-random missing means that the missing value is related to its own value. Generally, completely missing at random can be filled with its mathematical characters like mean value, while non-random deletions are more difficult to handle due to prior knowledge.

The differences among cases themselves and human factor in recording judgment documents lead to the mismatch between variables, causing prominent sparsity. Actually, almost all variables contain missing-value and some even lost 60% values. To cope with such circumstance, we propose following procedures:

1. Common steps: Drop the variables whose ratio of missing value is above 60%; fill the data according to the correlation between variables, like whether the “Cohabitation time” being greater than 0 can provide evidence to fill in the missing value of “Whether they cohabit”.
2. Individual steps (Choose one of those method to fill data: 1) Missing-value-sample removal, 2) Direct filling data: filling continues variables with their mean and category variables referring to its proportion randomly, 3) KNN filling method.

3.3. Variable selection
The selection of variables has a significant relationship with the fit on the training set and the accuracy on the prediction set. Specifically, when there are fewer variables, the model is under-fitting which leads to a poor prediction both on training and testing dataset; in the contrast, if too many variables are selected, the model tends to be overfitting causing a mismatch effect on training and testing dataset. Therefore, a benchmark is needed to rank the importance of variables.

Generally, statistical learning or machine learning methods are applied to filter variables, such as Lasso and Recursive Feature Elimination. This study takes Recursive Feature Elimination to recursively screen out unimportant variables through a Stepwise Regression Method. Figure 1 measures the correlation between the number of obtained variables and the accuracy of prediction:

![Figure 1. The correlation between the number of variables and score.](image)

4. A flexible method: application to a divorce data
There are over one hundred variables in the dataset describing 5152 samples, and generally samples with the same values tends to have the same response. However, suppose that we drop too many insignificant variables, the slight differences among original data caused by dropped variables will vanish and form an “overlapping” dataset (samples with the same values might be corresponding to different responses) which yields a poor accuracy. Therefore, we propose a feasible method named Two-Stage Model to cope with this situation.
In the first stage, we applied Random Forest Model to pre-prepared dataset. The model recommends 7 significant variables to conduct the establishment, and the 7 significant variables are listed in Table 2.

However, we find the direct approach built on an unsatisfactory dataset consistent with a poor prediction. In Table 3, samples with the same values correspond to different responses. For example, in Class 1, there are 1377 variables being marked as “Divorce” and 54 variables being charted as “Not Divorce” for the same independent-variables-values which leads to a chaos among the whole dataset.

We suggest applying a “Vote” method to deal with this problem. Considering the ratio of divorce and not divorce, we build a test to see if this ratio is different from 1 significantly. If does, like Class 1 and Class 5, we call the majority dominates the minority, and we should categorize the minority into the majority. If not, like Class 4, we claim that the obtained 7 variables cannot make those samples to be distinguishable, so we are supposed to collect those samples to do further process.

**Table 2. Importance of obtained variables in stage one.**

| Variables | Importance |
|-----------|------------|
| WAD       | 0.363526   |
| NDP       | 0.218969   |
| WNDS      | 0.154530   |
| WRT       | 0.119701   |
| WSE       | 0.070767   |
| WDM       | 0.044132   |
| WQM       | 0.022456   |

Notes: WAD: Whether the defendant agrees to the divorce, NDP: Number of divorces prosecuted, WNDS: Whether exists any cases where the court has not allowed a divorce but pair has been separated for a full year, WRT: Whether the relationship is two years old, WSE: Whether exist separation, WDM: Whether the defendant is missing, WQM: Whether to quarrel over family trivial matters.

**Table 3. Distinguishable and indistinguishable class after first fitting.**

| WQM | WAD | NDP | WSE | WRT | WNDS | WDM | Divorce | Not Divorce |
|-----|-----|-----|-----|-----|------|-----|---------|-------------|
| Class 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1377 | 54 |
| Class 2 | 1 | 0 | 0 | 0 | 0 | 0 | 245 | 180 |
| Class 3 | 1 | 0 | 0 | 0 | 0 | 0 | 389 | 44 |
| Class 4 | 0 | 0 | 0 | 1 | 0 | 0 | 94 | 97 |
| Class 5 | 0 | 0 | 0 | 0 | 1 | 1 | 5 | 65 |

In stage two, we apply XGboost combined with Recursive Feature Elimination to those indistinguishable samples. And the results show that, after stage two modelling, there are only very limited samples still being indistinguishable, so it is acceptable to only process two stages and the new obtained variables are listed in Table 4.

**Table 4. Importance of obtained variables in stage two.**

| Variables | Importance |
|-----------|------------|
| WDS       | 0.213606   |
| LTN       | 0.164118   |
| WDAM      | 0.141410   |
| PME       | 0.113910   |
| WDV       | 0.104457   |
| WGD       | 0.081487   |
| WQM       | 0.046685   |
| WCBW      | 0.044116   |
Notes: WDS: Whether the defendant has a spouse but cohabits with another person, LTN Living together or not, WDAM: Whether the defendant has abused and abandoned family members, PME: Pre-marital relationship evaluation, WDV: Whether exists domestic violence, WGD: Whether the defendant has gambling and drug abuse, WQM: Whether to quarrel over family trivial matters, WCBW: Whether their children born in wedlock.

The main concept of this Two Stages Model is establishing two different models corresponding to different obtained variables. We summarize the above procedure in Algorithm 1.

Algorithm 1: Two Stages Model

1. Data preparation
   - Filling in data with kinds of methods under a benchmark
   - Apply SMOTE function to cope with sparsity

2. Stage one
   - Apply Random Forest Model combined with Recursive Feature Elimination to divide dataset into distinguishable dataset and indistinguishable dataset
   - In distinguishable dataset, categorize the minority into the majority
   - Re-apply Random Forest Model to first-refined dataset

3. Stage two
   - Apply XGboost combined with Recursive Feature Elimination to indistinguishable dataset
   - Drop out very limited samples which are still indistinguishable
   - Categorize the minority into the majority in left dataset
   - Re-apply XGboost Model to second-refined dataset

5. Comparison and discussion

In order to improve the efficiency and quality of the trial of divorce dispute cases, in this paper we propose a new feasible method for training classified data with sparsity. Then the new method is applied to a divorce data collected from somewhere of southwest in China. In the first stage, seven binary variables are selected by random forest model. And in the second stage, eight variables are selected for classifying samples that cannot be distinguished by variables in the first stage. Compared with the traditional methods, there are two innovations in the proposed method: the first is dealing with sparsity and the second is coping with indistinguishability caused by variables-dropping. Meanwhile, the merits showed by Two-Stage Model are also prominent. This kind of method, on the one hand, are capable of extracting compact and comprehensive information in dataset as much as possible without ignoring too many samples whose values are not complete. On the other hand, recursive modelling on indistinguishable dataset provides with a solution to “overlapping” dataset and enhances the stability of prediction consistent with a fair accuracy. In fact, these two merits lead to a huge leap in accuracy. More concisely, apply Random Forest and XGboost directly shows around 82% accuracy on testing dataset while the Two-Stage Model promotes the accuracy to 89%.

Acknowledgments

This work was supported by the National Key R & D Program of China (2018YFC0831900), Sichuan University Post-Doctor Research Project, the National Natural Science Foundation of China (11871357), and SCU LAIW.

References

[1] Stevenson B and Justin 2007 Marriage and divorce: changes and their driving forces. J. Econ. Perspect. 21 27–52
[2] Wang Q and Zhou Q 2010 China's divorce and remarriage rates: trends and regional disparities J. Div. Rem. 51 257–268
[3] Zhang C, Wang X and Zhang D 2014 Urbanization, unemployment rate and China's rising divorce rate. Chin. J. Popul. Res. Environ. 12 157–64 (in Chinese)
[4] Gao M 2011 Rural divorce rate and outdoor employment: research on village panel data from
2003 to 2009 in China. *J. World Economy* **10** 55–69 (in Chinese)

[5] Liu H and Yang L 2014 Spatial heterogeneity of imbalanced sex ratio at birth in rural China and its influence factors: based on geographically weighted regression model *J. Population* **36** 5–15 (in Chinese)

[6] Su L, Liang C, Yang X and Liu Y 2018 Influence factors analysis of provincial divorce rate spatial distribution in China *Discrete Dynamics in Nature and Society* **2018** 1–11

[7] Fu H and Li X 2008 The analysis of the contemporary Chinese divorce situation and the marriage forecast *Northwest Population J.* **29** 55–8 (in Chinese)

[8] Xu A and Mao Y 2001 Analysis on the characteristics and causes of high divorce rate in uygar inhabited areas of Xinjiang *Chin. J. Popul. Sci.* **2** 25–35 (in Chinese)

[9] Xu A 1991 The present situation and trend of urban marriage in China *Sociological Research* **3** 12–8 (in Chinese)

[10] Friedman J 2001 Greedy function approximation: a gradient boosting machine *Ann. Statist.* **29** 1189-232

[11] Hastie T and Friedman J 2009 The elements of statistical learning: data mining, inference, and prediction. 2nd ed. New York: Springer

[12] Zhang T and Johnson R 2014 Learning nonlinear functions using regularized greedy forest *IEEE Transactions on Pattern Analysis and Machine Intelligence* **36**

[13] Fisher R A 1936 The use of multiple measurements in taxonomic problems *Annals of Eugenics* **7** 179-88

[14] Pearl R and Reed L 1920 On the rate of growth of the population of the United States since 1790 and its mathematical representation *Proc. Natl. Acad. Sci.* **6** 275–88

[15] Berkson J 1994 Application of the logistic function to bio-assay *J. Amer. Statist. Assoc.* **39** 357-65

[16] Robert E Schapire 1999 A brief introduction to boosting Proceedings of the 16th international joint conference on Artificial intelligence **2** 1401–6

[17] Leo Breiman 1996 Bagging Predictors *Machine Learning* **24** 123-40

[18] Chen T and Guestrin C 2016 XGBoost: a scalable tree boosting system *arXiv:1603.02754v1*