Is Your Marriage Reliable? Divorce Analysis with Machine Learning Algorithms

Jue Kong†
Chang'an university
ShangYuan Road
Xi'an, China, 710021
+8618994535561
2017901295@chd.edu.cn

Tianrui Chai†*
Beihang University
XueYuan Road No.37
Beijing, China, 100191
+8618914739576
trchai@buaa.edu.cn

†: These authors contributed equally to this work.

*: corresponding author

ABSTRACT

In recent years, global divorce rate is still high. What kind of couple will divorce and what factors lead to divorce are important problems that worth studying. In this paper, we apply three machine learning algorithms (Support Vector Machine (SVM), Random forest (RF) and Natural Gradient Boosting (NGBoost)) on a divorce prediction dataset. The dataset consists of 170 samples, each of which contains 54 questions about the couple's emotional status. We regard the scores of 54 questions as the features of each sample to apply our machine learning algorithms. Compared with SVM and RF, NGBoost has superior performance as NGBoost can achieve 0.9833 accuracy, 0.9769 precision and 0.9828 F1 score. In addition, we also show the most important features in the model of RF and NGBoost to find the most important factors which lead to divorce.

CCS Concepts

• Computing methodologies→Machine learning→Machine learning approaches→Classification and regression trees.

Keywords

Divorce; machine learning; random forest; natural gradient boosting; data mining.

1. INTRODUCTION

Husband and wife are two different sides of a family and tied together to live for a better live. But Some couples divorced for various reasons. According to [3], in America, 40% to 50% of all first marriages, and 60% of second marriages, will end in divorce. Although divorce rates have declined in some areas [2], the global divorce rate is on the rise as a whole. As everyone knows, high divorce rates can lead to social unrest in many ways.

Divorce do harm to children, especially for teenagers. Children with divorced parents are more likely to drink, take drugs, injure themselves or do something harmful to others [1]. Except for this, divorce also hurts both husband and wife to some extent. Therefore, it is a very important and significant issue to explore the causes of divorce and predict what kind of couples will divorce.

Over the past few years, researchers have studied this issue from various perspectives. A. Jarynowski and P. Nyczka applied Dynamic Network to analyze divorce [6]. J. Li, G. Zhang, H. Yan et al. proposed a markov logic networks based method to predict divorce [7]. Yöntem, M , Adem, K , İlhan, T, et al. used neural network to predict divorce and analyze the important factors affecting divorce [8].

In this paper, we apply three machine learning algorithms, Support Vector Machine (SVM) [9], Random forest (RF) [10] and Natural Gradient Boosting (NGBoost) [11] to identify potential divorcing couples. Based on these algorithms, we could also determine important factors associated with divorce. In Section 2, we introduced our materials and three algorithms we used to detect divorcing couples. In Section 3 and 4, we put forward our experimental result and our conclusion.

2. MATERIALS AND METHODS

2.1 Dataset

Table 1. Some sample questions in the questionnaire

No.	Questions
1	If one of us apologizes when our discussion deteriorates, the discussion ends.
11	I think that one day in the future, when I look back, I see that my spouse and I have been in harmony with each other.
17	We share the same views about being happy in our life with my spouse
20	My spouse and I have similar values in trust.
26	I know my spouse's basic anxieties.
31	I feel aggressive when I argue with my spouse.
33	I can use negative statements about my spouse's personality during our discussions.
38	I hate my spouse's way of open a subject.
40	We're just starting a discussion before I know what's going on.
49	I have nothing to do with what I've been accused of.

ICCAI '20, April 23–26, 2020, Tianjin, China

ACM ISBN 978-1-4503-7708-9/20/04...$15.00
DOI: https://doi.org/10.1145/3404555.3404559

Our dataset was obtained from [8]. There are 170 samples in the dataset. Of the samples, 84 (49%) were divorced and 86 (51%) were married couples. According to [8], couples were invited to finish a questionnaire with 54 questions about their marriage. They scored themselves on a five-point scale according to the degree of conformity with the questions. We listed some sample questions in Table 1. In this paper, we regard the scores of 54 questions as 54 features.

2.2 Machine Learning Algorithm

2.2.1 Support Vector Machine (SVM)

SVM is a effective machine learning method with sufficient theoretical explanation, and it is mainly used for classification recognition and regression modeling [2], [9]. It calculates a decision boundary to maximizes the distance of the nearest points of any class to the decision boundary. Given N points $\{X_i, Y_i\}_{i=1}^{N}$ where $X_i \in R^n$ is the *i*-th input and Y_i is the *i*-th input's label. In this paper, $n = 54$ and $Y_i \in \{0,1\}$. SVM aims to create a classifier in the form of:

$$f(x) = sign(\sum_{i=1}^{n} y_i a_i \psi(x, x_i) + b) \quad (1)$$

where a_i are positive real constants and b is a real constant. $\psi(\cdot,\cdot)$ is a kernel function [12].

2.2.2 Random Forest (RF)

Random forest is an ensemble of decision trees trained with the bagging method used for Classification, Regression and other tasks [4], [10]. Trees in forest are trained on random samples of training data and the split of each node on a tree is based on a random subset of features. The final result of RF is determined according to ensemble of the all classification results of decision trees by average voting. Random forest has a better ability to control the overfitting compared with single decision tree.

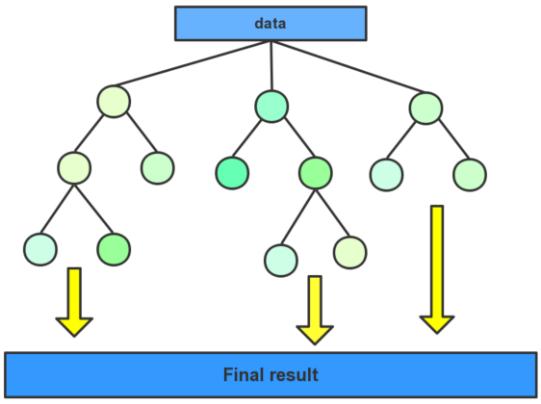

Figure 1. Illustration of Random forest.

2.2.3 Natural Gradient Boosting (NGBoost)

The NGBoost algorithm is a supervised learning method for probabilistic forecasting proposed by Tony Duan, Anand Avati, Daisy Yi Ding, et al. in 2019. It is a gradient boosting approach which uses natural gradients to solve the problem of simultaneous boosting of multiple parameters from the base learners [11]. In this paper, we use the python library "ngboost" created by stanfordmlgroup and choose decision trees as the base leaners.

2.3 Evaluation

In order to evaluate algorithms on the dataset objectively, we used 10-fold cross-validation. On 10-fold cross-validation, the data set was divided into ten subsets, and 9 of them were taken as training set and 1 subset was used as validation set. This process repeated 10 times, with every subset used as validation set exactly once.

In this paper, Accuracy, Precision and F1 score are utilized to evaluate performance. The definitions of them are:

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad (2)$$

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

$$F1\ score = 2 * \frac{precision*recall}{precision+recall} \quad (4)$$

where TP, TN, FP, FN means True Positive, True Negative, False Positive, False Negative and

$$recall = \frac{TP}{TP+FN} \quad (5)$$

3. RESULTS

3.1 SVM

In this paper, we choose the Radial Basis Function as the kernel function and set key parameters C = 1.0, Gamma = 0.03. The performance of SVM is shown in Table 2. All results are calculated as the average of 10-fold cross-validation.

Table 2. Performance of SVM

Evaluation Metrics	Evaluation Scores
Accuracy	0.9715
Precision	0.9769
F1 score	0.9710

As shown in Table 2, SVM has achieved very good performance as all evaluation scores are higher than 97%.

3.2 Random Forest

The key parameters of Random forest set in this paper are shown in Table 3.

Table 3. Key parameters of Random forest

Parameters	Description	Optimal Value
n_estimators	The number of trees in the forest.	2000
max_features	The number of features to consider when looking for the best split.	15
max_depth	The maximum depth of the tree.	not limited
min_samples_split	The minimum number of samples required to split an internal node.	2
min_samples_leaf	The minimum number of samples in newly created leaves.	1
min_weight_fraction_leaf	The minimum weighted fraction of the input samples required to be at a leaf node.	0
bootstrap	Whether bootstrap samples are used when building trees.	True

The performance of random forest is shown in Table 4. Random forest has better performance and all three evaluation scores are better than that of SVM's.

Table 4. Performance of Random Forest

Evaluation Metrics	Evaluation Scores
Accuracy	0.9778
Precision	0.9824
F1 score	0.9772

The top 10 important features of random forest are shown in Figure 2. As shown in Figure 2, question 18 is the most important question analyzed by random forest.

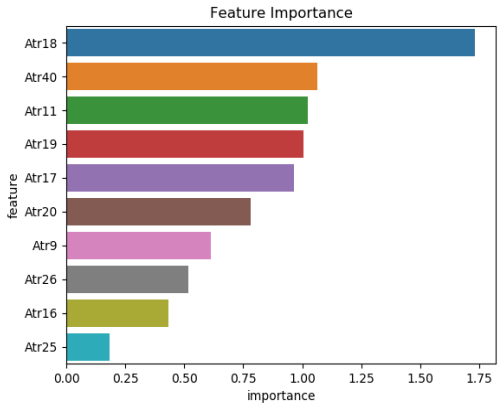

Figure 2. Top 10 importance features of Random forest.

3.3 Natural Gradient Boosting

The key parameters of Random forest set in this paper are shown in Table 5.

Table 5. Key parameters of Natural Gradient Boosting

Parameters	Description	Optimal Value
base leaner	The base leaner of natural gradient boosting machine	decision tree
n_estimators	The number of trees in the gradient boosting machine.	2000
max_depth	The maximum depth of the tree.	3
min_samples _split	The minimum number of samples required to split an internal node of the base decision tree.	2
min_samples _leaf	The minimum number of samples in newly created leaves of the base decision tree.	1
min_weight_f raction_leaf	The minimum weighted fraction of the input samples required to be at a leaf node of the base decision tree.	0
minibatch_fra c	Fraction of minibatch selected randomly without resampling.	0.9
Dist	Distribution configured in NGBoost	Bernoulli
Score	Scoring rule	MLE

The performance of Natural Gradient Boosting is shown in Table 6. It outperforms SVM and random forest on all three evaluation scores.

The top 10 important features of random forest are shown in Figure 3. As shown in Figure 3, question 18 is still the most important feature and is relatively more important than that of random forest.

Table 6. Performance of Natural Gradient Boosting

Evaluation Metrics	Evaluation Scores
Accuracy	0.9833
Precision	0.9875
F1 score	0.9828

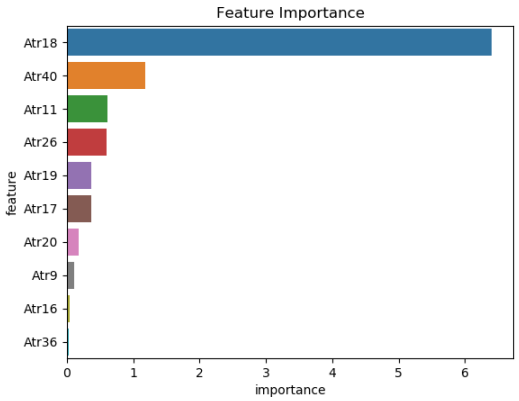

Figure 3. Top 10 importance features of Natural Gradient Boosting.

Table 7. Top 10 important features of each algorithm

RF		NGBoost	
No.	Question	No.	Question
18	My spouse and I have similar ideas about how marriage should be.	18	My spouse and I have similar ideas about how marriage should be.
40	We're just starting a discussion before I know what's going on.	40	We're just starting a discussion before I know what's going on.
11	I think that one day in the future, when I look back, I see that my spouse and I have been in harmony with each other.	11	I think that one day in the future, when I look back, I see that my spouse and I have been in harmony with each other.
26	I know my spouse's basic anxieties.	19	My spouse and I have similar ideas about how roles should be in marriage.
19	My spouse and I have similar ideas about how roles should be in marriage.	17	We share the same views about being happy in our life with my spouse.
17	We share the same views about being happy in our life with my spouse.	20	My spouse and I have similar values in trust.
20	My spouse and I have similar values in trust.	9	I enjoy traveling with my wife.
9	I enjoy traveling with my wife.	26	I know my spouse's basic anxieties.
16	We're compatible with my spouse about what love should be.	16	We're compatible with my spouse about what love should be.
36	I can be humiliating when we discussions.	25	I have knowledge of my spouse's inner world.

3.4 Important Features

In order to effectively solve the problem of high divorce rate, we output the importance of each feature. The ten most important features of RF and NGBoost are shown in Table 7.

4. CONCLUSION AND DISCUSSION

In this work, three machine learning algorithms are applied to predict potential divorce. SVM, RF and NGBoost all have good performance. NGBoost has a slight advantage than SVM and RF in predicting. In addition, we find out the most important factors lead to divorce calculated by RF and NGBoost. As shown in Table 7, there is only one difference in the top ten questions. This shows that the research is very convincing. In the future, we will collect more data and add some personal information to analyze.

5. REFERENCES

[1] A. P. Panatagama, G. P. Pratama and D. Y. Wibawa, "SocioEmpathy: A Social-Sensitivity Application to Reduce Stress and Depression of Divorce or Domestic Violence Victims," 2018 6th International Conference on Information and Communication Technology (ICoICT), Bandung, 2018, pp. 92-97.

[2] Zhang, C., Zhou, Y., Guo, J. et al. Research on classification method of high-dimensional class-imbalanced datasets based on SVM. Int. J. Mach. Learn. & Cyber. 10, 1765–1778 (2019)

[3] Cella, Annabelle, Ford Martin, and Marielle Cohen. "Developmental Psychology at Vanderbilt."

[4] Lakshmanaprabu, S.K., Shankar, K., Ilayaraja, M. et al. Random forest for big data classification in the internet of things using optimal features. Int. J. Mach. Learn. & Cyber. 10, 2609–2618 (2019)

[5] S. Sohail, S. Aziz, F. Tahir, S. Haqqui and A. Hussain, "Implementation of machine learning algorithm on factors effecting divorce rate," *2018 International Conference on Engineering and Emerging Technologies (ICEET)*, Lahore, 2018, pp. 1-5.

[6] A. Jarynowski and P. Nyczka, "Dynamic Network Approach to Marriage/Divorces Problem," 2014 European Network Intelligence Conference, Wroclaw, 2014, pp. 122-125.

[7] J. Li, G. Zhang, H. Yan, L. Yu and T. Meng, "A Markov Logic Networks Based Method to Predict Judicial Decisions of Divorce Cases," 2018 IEEE International Conference on Smart Cloud (SmartCloud), New York, NY, 2018, pp. 129-132.

[8] Yöntem, M , Adem, K , İlhan, T , Kılıçarslan, S . (2019). Divorce Prediction Using Correlation Based Feature Selection and Artificial Neural Networks. Nevşehir Hacı Bektaş Veli Üniversitesi SBE Dergisi , 9 (1) , 259-273 .

[9] Cortes, Corinna, and Vladimir Vapnik. "Support-vector networks." Machine learning 20.3 (1995): 273-297.

[10] Breiman, Leo. "Random forests." Machine learning 45.1 (2001): 5-32.

[11] Duan, Tony, et al. "NGBoost: Natural Gradient Boosting for Probabilistic Prediction." arXiv preprint arXiv:1910.03225 (2019).

[12] Haiqing Yu, Yanling Li, Shujun Zhang, and Chunyan Liang. 2019. Popularity Prediction for Artists Based on User Songs Dataset. In Proceedings of the 2019 5th International Conference on Computing and Artificial Intelligence (ICCAI '19). ACM, New York, NY, USA, 17-24.