빅데이터를 이용한 재난 생존자 예측

정길준, 강주성, 이흥노*
광주과학기술원

jinpeg112@gist.ac.kr, k92492@gist.ac.kr, heungno@gist.ac.kr

Prediction of Survival from Disaster with Bigdata

Giljun Jung, Jusung Kang, Heung-No Lee*
Gwangju Institute of Science and Technology (GIST)

Abstract

In this paper, we analyze the survival on shipwreck of the RMS Titanic. In order to predict the passengers are alive or not, we select features according to the values of correlation between each feature. After we submit to Kaggle, we get the accuracy of test data using deep neural network. We compare various cases by using different features, and we predict the survival of each passenger aboard the Titanic properly.

Ⅰ. Introduction

On April 15, 1912, there was one of the most infamous shipwrecks which was the sinking of the RMS Titanic [1]. The Titanic collided with an iceberg and sank in the North Atlantic Ocean. From this disaster, 1502 out of 2224 passengers were killed. In this paper, we analyze the survival from the disaster which is about RMS Titanic through Kaggle. Kaggle is an online community website including plenty of open competitions. For example, there is an ongoing competition which is 'Titanic: Machine Learning from Disaster'. The dataset is given in Kaggle, and until Dec 31, 2018, more than 10430 teams have participated in the competition.

Tryambak Chatterjee used two methods to find the prediction of survival [2]. One is Multiple Linear Regression, the other is Logistic Regression. He divided given training data into 4 ways to validate the results of training. The maximum accuracy that he got through the experiments was 80.756%. Aakriti Singh, Shipra Saraswat, and Neetu Faujdar predicted the survival using 4 methods [3]. These are Naïve Bayes, Logistic Regression, Decision Tree, and Random Forest. They got the accuracy of 91.3% with Naïve Bayes, 94.26% with Logistic Regression, 93.06% with Decision Tree, and 91.86% with Random Forest.

There are various methods to solve real world problems using deep learning [4]. In this paper, deep neural network (DNN) is supposed to analyze survival of passengers aboard the Titanic. The final outcome of analysis with DNN is the prediction that the passengers are alive or not.

Ⅱ. Approach

The following dataset is given on KAGGLE. One is 'train.csv', and the other is 'test.csv'. Both files have variables on the first row, and they have elements of each variable from the second row. The given dataset has 12 features as shown in Table 1. The common variables are PassengerId, Pclass, Name, Sex, Age, SibSp, ParCh, Ticket, Fare, Cabin, and Embarked. 'train.csv' has one more variable, Survived which is not in 'test.csv'. If Survived has 1, it means survival. Embarked has the first character of Cherbourg, Queenstown, or Southampton.

Table 1. The given features on Kaggle.

Variable	Description
PassengerId	Index of passengers
Survived	Survival (0 or 1)
Pclass	Ticket Class (1 or 2 or 3)
Name	Name
Sex	Sex (male or female)
Age	Age
SibSp	# of siblings / spouses aboard the Titanic
ParCh	# of parents / children aboard the Titanic
Ticket	Ticket number
Fare	Passenger fare
Cabin	Cabin number
Embarked	Port of Embarkation (C or Q or S)

Before trying to train the models, choosing proper features is needed and most important. Our purpose is predicting a passenger is alive or not. A feature has more information to predict passengers are alive or not if the magnitude of correlation with Survived is

near to 1. We did plot the correlation table between each feature except for Name as shown in Figure 1. If the magnitude of the correlation is bigger, the color has more vivid. A total of 7 features are selected. There are Pclass, Sex, Age, SibSp, ParCh, Fare, and Title. Title is extracted from parsing Name. The prediction of survival is analyzing by the combination of these features.

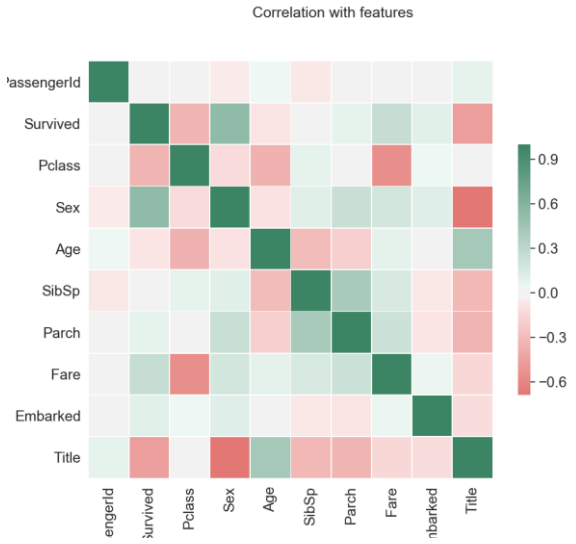

Figure 1. Correlation table between each feature.

Some passengers do not have full elements on each variable. In addition, the type of some data might be a string. We need two kinds of data pre-processing to make a calculation of weights in DNN model be easier. One is to fill the empty values, the other is to convert string values into numeric values. There are some missing values on Age, Fare, and Title. Age is correlated with Pclass according to Figure 1, so the empty values on Age are filled with the average values of each Pclass. Fare has the range from $4.0125 to $512.3292 in training dataset. The lowest price is allocated to 15 null elements. To make a feature of Title, we use enum function to Mr., Miss., Master, Mrs., and others. Additionally, we take only quotients of Age and Fare divided by proper number to make our model simpler.

In our DNN, the models have 2 Layers of a fully connected layer. To predict Survived, we utilize one-hot encoding. While training the given training dataset, we are exposed to an overfitting problem. To reduce overfitting, we use ReLU activation function and apply dropout. In addition, we train multiple classifiers instead of learning single classifier in order to get higher accuracy and to overcome overfitting problem.

Ⅲ. Results

We measure the performance of DNN models for 5 cases with submission to Kaggle. Table 2 shows which features are used in each case and its accuracy. Child in Case 5 is an additional feature which means Age is under 14 or not. Accuracy is measured by the percentage of passengers who are correctly predicted to alive. The maximum accuracy of all cases is 79.90%.

Table 2. Used features and accuracy in each case.

Case	Used Features	Accuracy (%)
Case 1	Pclass, Sex, Age, SibSp, ParCh	79.90
Case 2	Pclass, Sex, Age, SibSp, ParCh, Title	79.43
Case 3	Pclass, Sex, Age, SibSp, ParCh, Title, Fare	77.99
Case 4	Pclass, Sex, Age, ParCh, Title	78.95
Case 5	Pclass, Sex, Age, SibSp, Child	77.03

The result is that the accuracy of test data with our DNN model converges to 80%. According to [3], the results look like higher accuracy, but all results of [2], [3] are measured not from test data. However, the accuracy of DNN is from test data so we can say our results have more meaningful.

Ⅳ. Conclusions

To predict the passengers who are aboard on Titanic are alive or not, selecting features properly is important. Under expectation that a higher accuracy is from using features more correlated with Survived, our DNN model has accuracy around 77~80%. There exists a limit of 80%. This result is from several reasons. The biggest reason is that the size of training data is not enough big, so the problem of overfitting has occurred. If we have more data, the accuracy will be much higher. In addition, if the results that predictions of survival from any disaster is applied for disaster relief, we expect many people more than now can be rescued.

ACKNOWLEDGMENT

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIP) [NRF-2018R1A2A1A19018665].

References

[1] "Titanic: Machine Learning from Disaster," *Kaggle.com*. [Online]. Available: https://www.kaggle.com/c/titanic. [Accessed: 31-Dec-2018].

[2] T. Chatterjee, "Prediction of survivors in titanic dataset: a comparative study using machine learning algorithms," *International Journal of Emerging Research in Management & Technology*, vol. 6, Jun. 2017.

[3] A. Singh, S. Saraswat, and N. Faujdar, "Analyzing Titanic Disaster using Machine Learning Algorithms," *2017 International Conference on Computing, Communication and Automation (ICCCA)*, Greater Noida, 2017, pp. 406-411. doi: 10.1109/CCAA.2017.8229835.

[4] J. Schmidhuber, "Deep learning in neural networks: An overview", *Neural Networks*, vol. 61, pp. 85-117, 2015. Available: 10.1016/j.neunet.2014.09.003.