Genetic Algorithm-based Feature Selection for Depression Scale Prediction

Seung-Ju Lee
Dept. Comp. Eng, Gachon University
Gyeonggi-do, Korea
poketred12@gc.gachon.kr

Hyun-Ji Moon
Dept. Comp. Eng, Gachon University
Gyeonggi-do, Korea
hyunjm95@gc.gachon.kr

Da-Jung Kim
Dept. Comp. Eng, Gachon University
Gyeonggi-do, Korea
rew324@gc.gachon.kr

Yourim Yoon
Dept. Comp. Eng, Gachon University
Gyeonggi-do, Korea
yryoon@gachon.ac.kr

ABSTRACT

This study aimed to improve the performance of machine learning prediction through feature selection using a genetic algorithm (GA) for predicting depression of elderly people based on the survey data of Korean Longitudinal Study of Aging (KLoSA) performed in South Korea. The proposed feature selection method finds an optimized feature set through a fitness function design that maximizes the correlations between the features selected and the label to be predicted while minimizing the correlations between the selected features by using GA. The effectiveness of the proposed GA was shown through comparative experiments.

CCS CONCEPTS

• **Computer methodologies→ Feature selection**; Heuristic function construction;

KEYWORDS

Genetic algorithms, Feature selection, Machine learning

ACM Reference format:

Seung-Ju Lee, Hyun-Ji Moon, Da-Jung Kim, and Y. Yoon, 2019. Genetic Algorithm-based Feature Selection for Depression Scale Prediction. In *Proceedings of ACM GECCO conference, Prague, Czech Republic, July 2019 (GECCO'19)*, 2 pages. DOI: 10.1145/3319619.3326779

1 INTRODUCTION

In this paper, a study is carried out to maximize the performance of machine learning prediction through feature selection using a genetic algorithm (GA) in terms of predicting the depression of elderly people based on the survey data of Korean Longitudinal Study of Aging (KLoSA) conducted in South Korea. Due to the characteristics of survey data that consist of questions and answers, data preprocessing is required and when the data labels are enormous and used as inputs of machine learning, appropriate preprocessing is essential for efficiency improvement. Moreover, a task of identifying significant data among large amount of data is required. This study derives significant data through the feature selection using GA, and based on this, proves that the performance and time of machine learning can be improved.

2 IMPROVED PREDICTION USING FEATURE SELECTION

This study used the survey results of "the 2nd KLoSA 2008" of Employment Research and Analysis System of South Korea, as a dataset [1]. The targets of KLoSA conducted by the Korea Employment Information Service were mid-aged or older individuals over 45 years of age (born before 1962) residing in South Korea excluding Jeju-island, and the study was performed for a sample size of about 10,000 people. In the data, CES-D10, a scale for distinguishing depression is given. It can be used a baseline for determining depression. It shows an integer from 0 to 10 as a sum of result values for ten questions on the psychological state of last one week in the survey.

To predict the CES-D10, which is one of depression scales, a machine learning method can be used. Because the original data is quite huge, a problem may occur in terms of efficiency if they are applied to the machine learning as they are. The feature selection is a processing process that selects useful attributes to obtain effective and improved solution in a given problem [2]. It is shown that the prediction performance of machine learning can be improved by performing the machine learning using the features selected by the GA.

3 FEATURE SELECTION BASED ON GENETIC ALGORITHM

In the feature selection algorithm using the GA proposed in this study, binary encoding was used to express the solutions. 1 was assigned when selected for the index of each feature, and 0 when not selected.

GECCO'19, July 13-17, 2019, Prague, Czech Republic
 978-1-4503-6748-6/19/07. . . $15.00
DOI: 10.1145/3319619.3326779

The number of solutions of initial population was set as 100, and the solutions were generated randomly. Furthermore, one-point crossover was selected as the crossover and a random location was selected as the mutation so that the value of corresponding location would be inversed. The mutation rate was set at 10%. The number of selected features was fixed and in a case of the number of selected features exceeding the fixed number, a repair task was carried out that deletes the exceeded number randomly from the selected features. The fitness function of proposed GA was designed as follows by using the correlation.

$$p * \overline{|p(X,Y)|} + |p'(X,C)|^{-1}$$

Here, p is a parameter that shows the weight of correlation and can be adjusted according to the problem. In this study, a value, 0.2 was used. $p(X,Y)$ is correlation coefficient between the features X and Y, and $p'(X,C)$ is a correlation coefficient between X and label C that shows the CES-D10 value. As the fitness result value of the equation decreases, the probability of being selected increases. Therefore, according to the proposed equation, the correlation between a selected feature and the label to be predicted is maximized while the correlation between the selected features is minimized. It has been introduced in many conventional studies that such a feature selection strategy shows good performance [2, 3].

4 EXPERIMENTAL RESULTS

By choosing five different numbers (10, 20, 30, 40, and 50) for the number of selected features, the results were compared between the proposed GA-based feature selection method and the method that selects features randomly among the total of about 3,000 ones. The feature sets selected by the two methods were learned by using four machine learning methods: random forest, linear regression, multilayer perceptron, and SMO Regression provided by the WEKA library [5]; and the results were compared. To increase the integrity of training data, a 10-fold cross-validation was used. For the evaluation scale of test-set, the RMSE values were used. The RMSE value is obtained as follows:

$$\boldsymbol{RMSE} = \sqrt{\frac{\sum(\boldsymbol{y}-\hat{\boldsymbol{y}})^2}{\boldsymbol{n}}},$$

where y is the real value, $\hat{y}$ is the the predicted value, and n is the number of data. It is a scale commonly used when dealing with the difference between the predicted value and the real value.

Table 1 shows the results of calculating RMSE by predicting the CES-D10 depression scale through each machine learning method based on the features selected by performing the GA-based feature selection and the features selected randomly according to each number of features mentioned above. As in the table, the number of features and four machine learning methods all showed that the GA-based feature selection reduced the prediction error compared to the random selection, i.e., the feature selection using the GA was more effective. Among the four machine learning methods, the random forest was most effective for the prediction. This seems to be because the structure of random forest method is most suitable for the characteristics of survey data.

Table 1 : The prediction error (RMSE) of depression scale according to the number of selected features and machine learning method

feature selection \ prediction method		Random Forest	Linear Regression	Multilayer Perceptron	SMOreg
GA	10	2.133	2.251	2.442	2.300
Random		2.231	2.383	2.552	2.383
GA	20	2.151	2.226	2.473	2.265
Random		2.482	2.517	2.815	2.569
GA	30	1.839	2.021	2.335	2.08
Random		2.131	2.216	2.539	2.244
GA	40	1.336	1.697	1.883	1.735
Random		1.916	2.09	2.475	2.130
GA	50	**1.007**	**1.525**	**1.72**	**1.584**
Random		1.448	1.757	2.135	1.786
No feature Selection		1.822	2.026	2.302	2.071

5 CONCLUSION

When supervised learning is performed by using the machine learning, the quantity and quality of training data are very important. The survey data used contained a large number of features and because there were too many features of data to apply the machine learning whereby the need to identify significant features was large, the time and performance deteriorate. Therefore, the performance can be improved by choosing the valid features by using appropriate feature selection method. This study compared the method of selecting features randomly and the method of selecting features by using the GA. As a result, the GA-based method showed a lower error calculated with RMSE than the random selection method. To sum up, the GA-based feature selection selected the features better for the prediction through the machine learning, and improved the performance of machine learning.

ACKNOWLEDGMENTS

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(Ministry of Science, ICT & Future Planning) (No. 2017R1C1B1010768).

REFERENCES

[1] http://survey.keis.or.kr/klosa/klosaguide/List.jsp
[2] Guyon, I., and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, *3*(Mar), 1157-1182.
[3] Kim, Y. H., and Yoon, Y. (2015). A genetic filter for cancer classification on gene expression data. *Bio-medical materials and engineering*, *26*(s1), S1993-S2002.
[4] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. ACM SIGKDD explorations newsletter, 11(1), 10-18.