Hypothesis Combination using Genetic Algorithm

Kuan-Yu Lai, Jhin-Wun Wang and Chien-Liang Liu
Department of Industrial Engineering and Management
National Chiao Tung University
1001 University Road, Hsinchu, Taiwan 300, ROC
sasuke11291@gmail.com, bee806806806@gmail.com, clliu@mail.nctu.edu.tw

Abstract

Improvement of machine learning performance is always an important issue in machine learning community. Ensemble learning has been shown to be able to normally improve model performance by combining various algorithms in the model. Most winners of large-scale data science competitions use ensemble learning technique to get high scores. Given available machine learning algorithms, ensemble learning involves two tasks, hypothesis selection and hypothesis combination. Hypothesis selection is about selecting the algorithms that are beneficial to model performance into the ensemble learning model, while hypothesis combination is related to determine the combination coefficients of the algorithms in the model. This work focuses on hypothesis combination problem, and we formulate it into an optimization problem. We propose to use genetic algorithm (GA) to tackle this optimization problem.
The GA is a search heuristic for the purpose of solving optimization problem, which is a part of evolutionary computation. GA is inspired by the theory of natural evolution, which comprises several components, including selection, crossover and mutation. There are five phases involved in a typical GA algorithm: initial population, fitness function, selection, crossover, and mutation. GA begins with an initial population which can be regarded as a set of solutions. Next, a fitness function has to be defined to determine the fitness of each individual. Once the fitness score for each individual is computed, the selection and crossover are used to generate the fitting offspring. For a small amount of new offspring, some of their genes will mutate with a low probability. The whole process will repeat until the population converges.
This work follows GA process to encode the problem, and defines appropriate fitness function according to the characteristics of the problem. We conduct experiments to evaluate the proposed approach in ensemble learning on a multi-class classification dataset, Red Wine Quality. We transform this problem into a binary classification problem, and use ten classification algorithms in the hypothesis pool, including Support Vector Machine, Random Forest, Decision Tree, Gradient Boosting, AdaBoost, Gaussian Naïve Bayes, Logistic Regression, Nu-Support Vector Machine, Stochastic Gradient Descent and Nearest Centroid. To determine the combination coefficients of each model, the continuous encoding of chromosomes is represented as binary chromosomes of approximation, and F1 is used as the fitness function. We use F1 score as the evaluation metric, and compare the proposed method with several alternatives. The experimental results indicate that the proposed method is comparative in determining the combination coefficients of algorithms in the ensemble learning.

Keywords
Ensemble Learning, Hypothesis Combination, Combination Coefficients, Genetic Algorithm

Acknowledgements
This work was supported in part by Ministry of Science and Technology, Taiwan, under Grant no. MOST 107-2221-E-009-109-MY2.

Biography / Biographies

Kuan-Yu Lai received the B.S. degree in Department of Transportation & Logistics Management from National Chiao Tung University, Taiwan in 2017. He is currently a graduate student in Department of Industrial Engineering and Management from National Chiao Tung University. His research interests include machine learning and data mining.

Jhin-Wun Wang received the B.S. degree in Department of Industrial Engineering and Management from National Chiao Tung University, Taiwan in 2018. He is currently a graduate student in Department of Industrial Engineering and Management from National Chiao Tung University. His research interests include machine learning and data mining.

Chien-Liang Liu received the M.S. and Ph.D. degree in Department of Computer Science from National Chiao Tung University, Taiwan, in 2000 and 2005, respectively. He is currently an assistant professor in Department of Industrial Engineering and Management at National Chiao Tung University, Taiwan. His research interests include machine learning, data mining, and big data analytics.