Analysis of Nine Instance-Based Genetic Algorithm Classifiers Using Small Datasets

Hossin, M., Mahudin, F., Din, I and Mat, A.R
Faculty of Computer Science and Information Technology,
Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia.
hmohamma@unimas.my

***Abstract*—The application of genetic algorithm (GA) has emerged covering various areas including data classification. In data classification, most studies of GA were focused on the enhancement of GA and development of different types of GA classifiers. To the best of our knowledge, there is no study has been conducted to examine the influence of GA operators based on the size of data set towards training time and generalization ability. Therefore, this study develops and compares nine Instance-based genetic algorithm (IbGA) classifiers with different combinations of GA operators. The goal of this comparison is to examine and identify the best combination of GA operators which have performed better on generalization ability and training time efficiency. Nineteen benchmark data sets were used in this study. The non-parametric statistical tests were applied to justify the comparison results. The statistical tests suggest that the combination of roulette wheel selection and uniform crossover operator is the best combination of IbGA model although the training time is a bit lengthier than compared to other IbGA models..**

***Index Terms*—Data Classification; Genetic Algorithm; Instance-Based Classifier.**

I. Introduction

Typically, genetic algorithm (GA) was used as an optimizer to solve complex problems. Since its inception, the use of GA has been expanded to solve data classification. There are two types of GA classifiers; rule-based GA (RbGA) and instance-based GA (IbGA) [1]. The IbGA classifier was inspired from the drawback of the nearest neighbor (NN) algorithm. The large storage of prototypes and long response time classification are two major drawbacks of NN classifier. Due to these disadvantages, IbGA was proposed to reduce the number of prototypes as much as possible while preserving the NN classifier performance. On the other hands, the RbGA classifier was inspired from the rule-based approach. In RbGA, each chromosome is represented by different rules that generated randomly. Each allele (or known as gene) represents each data attribute and represented by binary string (0 and 1) based on the possible values for each attribute. Normally, each allele has different length of binary bits. Then, the rule is generated by employing information measure such like entropy [2], or ranking with correlation coefficients [3].

Interestingly, many studies were done on the RbGA as compared to IbGA. From the review, the RbGA has been applied to solve large data sets [4]. Meanwhile, IbGA seems less attractive to researchers due to its complicated framework representation and optimization process. In IbGA, the process of building the classifier is a stochastic process where the optimal n reference set is searched using optimization process. Due to optimization process, the finding process has become computational costly when large data is involved. Although IbGA is less attractive, the reported generalization performance of IbGA was superior or at par as compared to other instance-based classifiers or other types of classifiers for many benchmark data sets [5, 6].

In order to design the best classification algorithm, many studies focus on both data and algorithmic level had been conducted. For algorithmic level, it includes the advanced design of algorithm and improvement in order to get better results for a specific domain. In contrast, this study attempts to analyse the algorithmic level of operators used in GA towards training time and generalisation ability. According to Abdoun and Abouchabaka [7], this analysis is important because the performance of GA is totally dependent on the selection of appropriate genetic operators. To the best of our knowledge, there are no known studies that focus on the effect of different combination GA operators towards generalization ability and training time efficiency. However, Andrade et al., [8] did conduct a comprehensive analysis to examine the effect of GA operator combinations in route searching problem in IP network domain. They noticed that each combination of GA operators did influence the performance of GA in routing searching problem. They also conclude that Stochastic Random Sampling (SRS selection) and uniform crossover combination was able to achieve less processing time as compared to other GA operator combinations. Thus, it is important to investigate the influence of each GA operator combinations towards the performance of IbGA in terms of training time and generalization ability. Through this study, two research questions have been identified:

- How different combinations of GA operators influence the performance of the IbGA towards training time efficiency and generalization ability based on several benchmark data sets?
- What is the best combination of GA operators that give better generalization performance and produce less training time based on several benchmark data sets?

The scope of this study is confined to the modification of classical IbGA classifier. This study will examine the influence of different combinations GA operators towards training time efficiency and generalization ability. Three selection techniques, three crossover operators, and one mutation technique were used for this particular study. In total, nine various IbGA classifiers were developed. This study employs a standard accuracy measure and training time (in second) to measure the performance of each proposed IbGA classifier on various benchmark data sets. 19 benchmark data sets which are binary-class datasets were