A genetic algorithm based feature selection from gene expression profiling for cancer classification using machine learning

Meijian Guan, Wake Forest University

1 Introduction

Cancer is one of the leading causes of morbidity and mortality in the US, with nearly 1,685,210 new cases in all sites and 595,690 deaths in 2016 [Howlader N et al. 2016]. Worldwide, cancer is the second leading cause of death, caused approximately 8.8 million deaths in 2015, and nearly 1 in 6 deaths is due to cancer [World Health Organization 2017]. Key strategies in reduction of mortality rate among cancer carriers are early detection, accurate determination of cancer type. Errors in cancer type or malignant growth type determination could cause reduced treatment efficiency, because anti-cancer strategies largely depend on tumor morphology [Podolsky et al. 2016]. The gene expression profile is a process that determines the time and location of gene expression. DNA mutation may change the gene expression, resulting in tumor or cancer growing. Gene expression profile has shown great value in classifying complex diseases such as cancer. A number of studies have demonstrated possibility to extract compelling information from microarray data to support clinical decisions on cancer diagnosis, prognosis and response to treatment [Patsialou et al. 2012; van 't Veer et al. 2002; Wang et al. 2005]. Despite the great opportunity, microarray data pose great challenges as well to accurately predict cancer types. First challenge comes from large amount of inherent noise and variability in samples, due to biological variations and experimental conditions. Furthermore, high dimensional features (tens of thousands of genes), as compared to a relatively small sample size could lead to the risk of over-fitting in most of machine learning methods. In practice, it's very difficult even for domain experts to determine an optimum feature set. Therefore, a feature selection method to eliminate irrelevant features and inherent noise is necessary before apply any machine learning algorithms on the data.

Genetic algorithms (GAs) are a group of techniques that inspired by biological evolution in order to solve optimization problems [Melanie Mitchell 1996]. They are good candidates for feature selection tasks since GAs are most useful in high-dimensionality and multi-class problems where heuristic knowledge is sparse or incomplete [Pei et al. 1995]. The objective of this study is to develop a GA-based approach, utilizing a feedback linkage between feature selection and classification. That is, a GA-based method carries out feature selection and simultaneously a classifier performs a classification with the selected features. The prediction accuracy is calculated based on the performance of the classifier, and it is used as the fitness function in evolution process of GA, with higher accuracy results in higher fitness. The goal of this method is to find a reduced subset among the original features such that redundant information and noise is excluded.

2 Experimental and Computational Details

2.1 Data Simulation

Simulated gene expression data for cancer patients were used to evaluate our GA model. We simulated gene expression data to mimic log2 (T/R) transformation with expression level ranged from -5 to 5 for 20 genes in 200 samples. Age, gender and cancer types (type 0

to type 3) were also simulated. Broadly, each cancer category has similar number of individuals, proportion of females and average age (Table 1). Gene expression data are accessed by the ordered gene indices from 1 to 20, with 0 means dropping the corresponding gene. We also simulated a larger dataset with 200 genes and 1000 samples. This led to an increased population size to 200.

Table 1. Sample characteristics of simulated gene expression data

Cancer type	Type0	Type1	Type2	Type3
N	57	44	54	45
Female (%)	43.9	38.6	55.6	44.4
Age*	60.21±8.82	60.25±9.08	60.87±8.44	59.76±8.59

*Mean ± standard deviation

2.2 Initial GA Design

The general study design is shown in Figure 1. Indices were assigned to the genes such that they can be used to code chromosomes in an incremental order and access gene expression data. In the initial population, a total of 20 chromosomes were coded in a "leave-one-out" manner, that is, each chromosome has one distinct gene index been coded as 0. Each chromosome was then used to extract gene expression data and perform cancer classification. The performance of each chromosome was evaluated via the proportion of correct prediction, and those achieved high classification accuracy were more likely to be selected to generate next generation. Single point crossover and single point mutation were also applied during the reproduction.

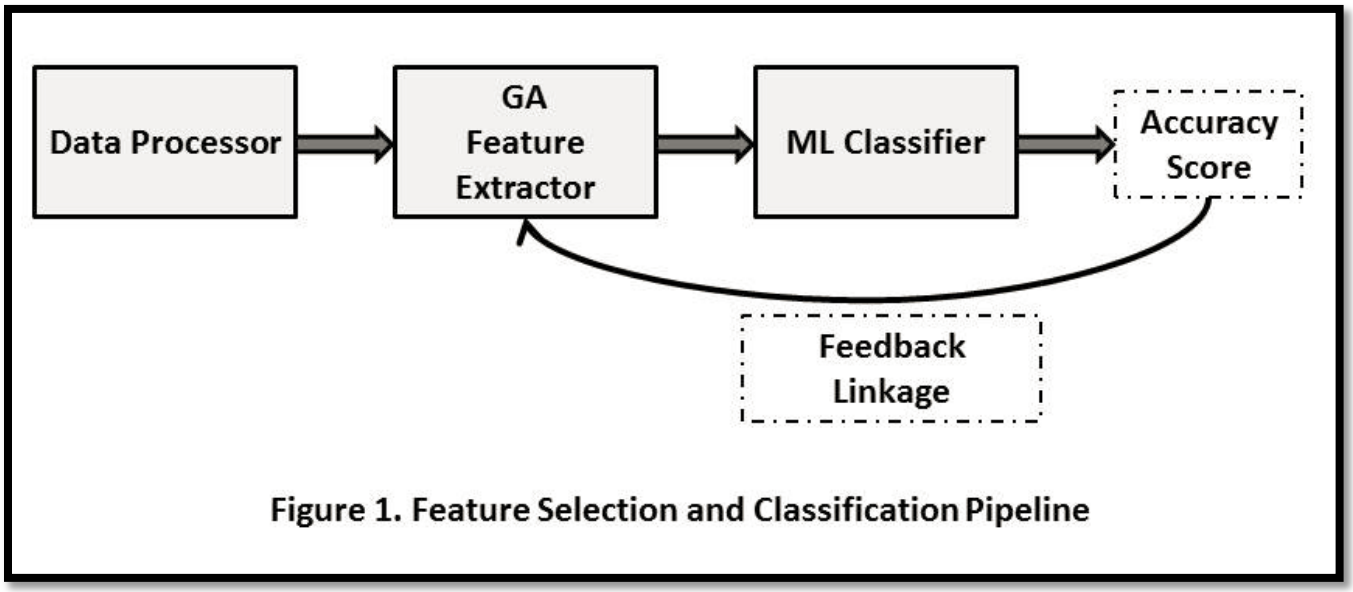

Figure 1. Feature Selection and Classification Pipeline

2.2.1 Classifier

We selected seven machine learning methods as candidate classifiers, including logistic regression (LR), linear discriminant analysis (LDA), K nearest neighbors (KNN), classification and regression tree (CART), gaussian naive bayes (NB), support vector machine (SVM), and artificial neural network (ANN). We performed ten-fold cross-validation with initial population to compare the performances of the seven algorithms. CART demonstrated the best performance and was

selected for the subsequent analysis (Figure 2). CART uses a decision tree to construct a predictive model which maps the data features to the value of a target variable [Leo Breiman et al. 1984]. The predictive model is obtained by recursively partitioning the data space and fitting a simple predictive model within each partition. CART consists of two categories of decision trees, classification trees, where the target variables take discrete values; and regression trees, where the target variables take continuous values. We only use classification trees for this study.

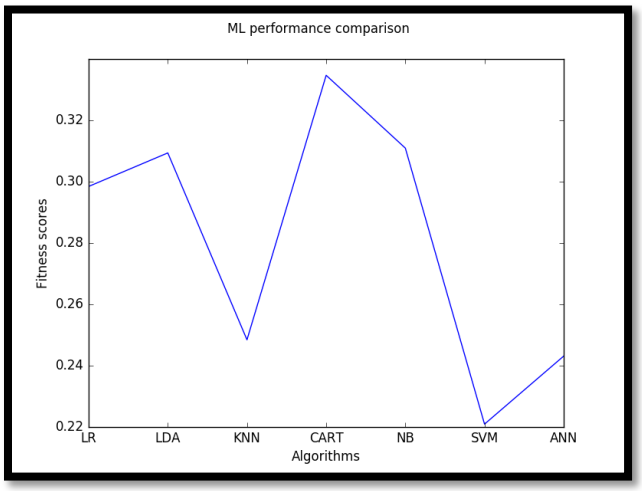

Figure 2. Performance comparison among seven machine learning algorithms on initial population

2.2.2 Fitness function

The fitness function was based on the prediction accuracy of the classifier for each chromosome, defined as $F = \frac{T}{N}$, where F denotes fitness score, T denotes true predictions and N is the total number of predictions. A feedback linkage between classification and GA feature selection was established. Specifically, we perform feature selection and cancer classification simultaneously, chromosomes with higher classification accuracy are more likely to be selected for reproduction.

2.2.3 Selection

A tournament selection method was implemented in our initial GA with a group size equals to four. Tournament selection involves conducting multiple “tournaments” among a group of chromosomes randomly chosen from the population. The chromosome with the best fitness score in each tournament is selected as parent chromosome for reproduction. At each generation, we perform tournament selection for 20 times in order to preserve the same population size in each population.

2.2.4 Crossover

Single point crossover between two parent chromosomes could occur during reproduction at a rate of P_c = 0.6. Crossover is a genetic operator that mimic biological crossover during reproduction used to exchange segments between parent chromosomes to generate child chromosomes. A crossover point is randomly selected and all the indices after that point were exchanged between two chromosomes if crossover occurs.

2.2.5 Mutation

Mutation is a genetic operator to maintain the genetic diversity from one generation of population to the next in GA. It is analogous to biological mutation. In this study, mutation was applied at a mutation rate of P_m=0.05, which

results in a switch between a single gene index and zero. Each base of a chromosome has equivalent chance to be mutated.

2.2.6 GA variations

Since we observed unstable performance of our baseline GA model, we applied modifications on crossover rate, mutation rate or tournament group size in order to evaluate their influence on GA performance. Three new variations were generated with, 1) decreased crossover rate from 0.6 to 0.3, 2) decreased mutation rate from 0.05 to 0.005, and 3) increased tournament group size from 4 to 8. In addition, we increased gene number to 200 and sample size to 1000 to evaluate whether population size and sample size have impacts on the GA performance.

2.3 Metropolis-Hastings

We sought to find an alternative algorithm to solve this problem. One candidate is Metropolis-Hastings algorithm (MH), which is a sampling algorithm to obtain random samples from a target probability distribution based on Markov chain Monte Carlo (MCMC) [Hastings 1970]. During the sampling procedure, MH generates a candidate x' from a proposal distribution for the next value given current data point x. An acceptance ratio $\alpha = f(x')/f(x)$ is then calculated to decide whether to accept or reject the candidate if assume symmetric proposal distribution. A sequence of data points from a probability distribution will be generated in this manner. Here f denotes the density function of target distribution which is considered as a fitness function of each sample. Along with other MCMC methods, MH doesn't suffer from the "curse of dimensionality", which makes it the best choice when sampling from a high dimensional space is needed. Therefore we implemented a MH method to find the best set of gene expression profiling to better predict cancer types. The target sample distribution clearly becomes an interval between [0,1] since we use classification accuracy to measure the performance of each sample. We assumed a symmetric proposal distribution in this study for simplicity. The initial sample is a list of gene indices from 1 to 20, which are used to extract the simulated gene expression data for the subsequent classification using CART. To generate a new sample, a "mutation" is applied to each position at a certain mutation rate to drop or add back genes. The prediction accuracy f(x) is estimated for each sample and an acceptance rate is calculated to decide whether accept or reject a new sample.

Our GA model is broadly similar to the sampling procedure of MH, that is, GA generates a candidate population based on current population with acceptance ratio always equals 1.

3 Results

We initially performed our model on the simulated dataset with P_c=0.6, P_m=0.05 and tournament group size=4 for 1000 generations. An increased population fitness score was observed (Figure 3). The maximum fitness score for a single chromosome was 0.57. Figure 4 showed the average population fitness, the highest fitness, as well as the minimum fitness of each generation respectively for the first 200 generations. A quick improvement of fitness was observed for both average and maximum fitness score. In addition, GA successfully dropped three genes from the population in the final generation. Overall, under initial parameters increased population fitness was observed, however the performance was not stable and needed to be further improved.

We hypothesized that crossover rate, mutation rate or tournament group size may have an

impact on the performance of our GA model. Thus, we applied modification on these three parameters to generate three GA variations, where variation 1 (V1) had P_c=0.3, variation 2 (V2) had P_m=0.005 and variation 3 (V3) had tournament group size=8. No significant improvement was observed for these three variations (Figure 5 - 10). We also hypothesized that increased population size and sample size would reduce the data variability, and could lead to better GA performance. The result for the first 1000 and 200 generations has been shown in Figure 11 and Figure 12. All three fitness scores demonstrated more stable performance, which confirmed our hypothesis. We next implemented MH as a feature selection method on the original dataset. We started with mutation rate =0.05 and acceptance rate=0.8 for 1000 iterations. No obvious improvement on prediction accuracy was observed. We then changed the mutation rate to 0.01, which led to a similar result (Figure 13, 14). We increased the number of generations to 10,000 to avoid potential "burn-in" period. However, no improvement was obtained (Data not shown). Various acceptance rates were also applied, but we did not observe better results.

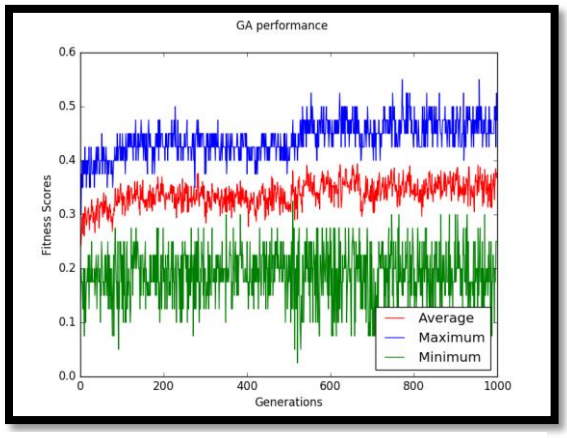

Figure 3. Fitness score over 1000 generations for baseline GA

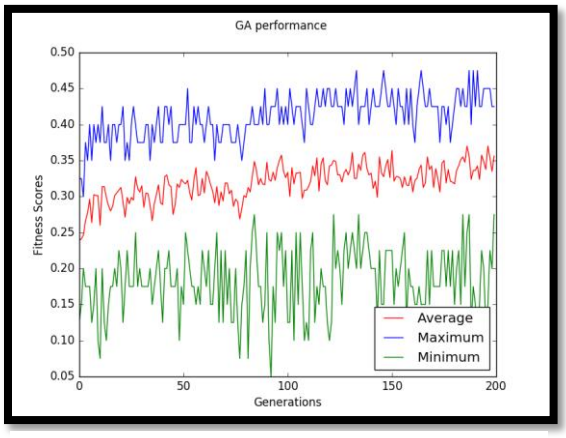

Figure 4. Fitness score over 200 generations for baseline GA

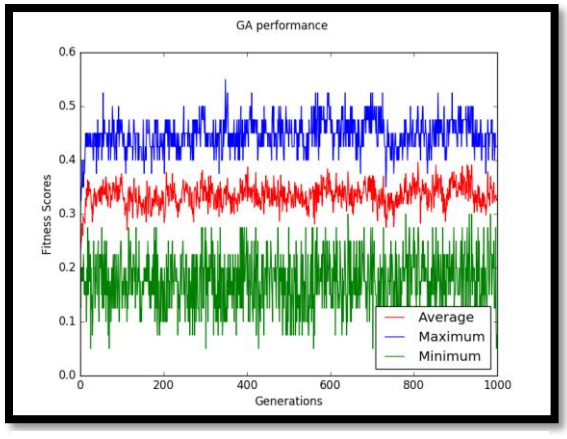

Figure 5. Fitness score over 1000 generations for GA V1

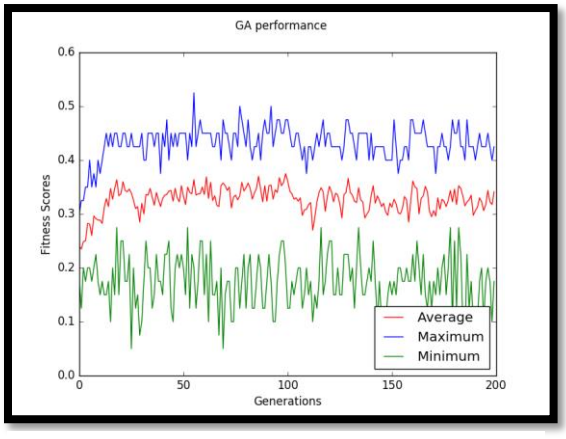

Figure 6. Fitness score over 200 generations for GA V1

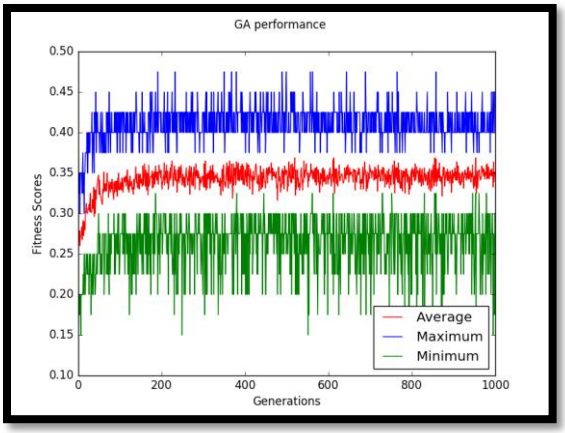

Figure 7. Fitness score over 1000 generations for GA V2

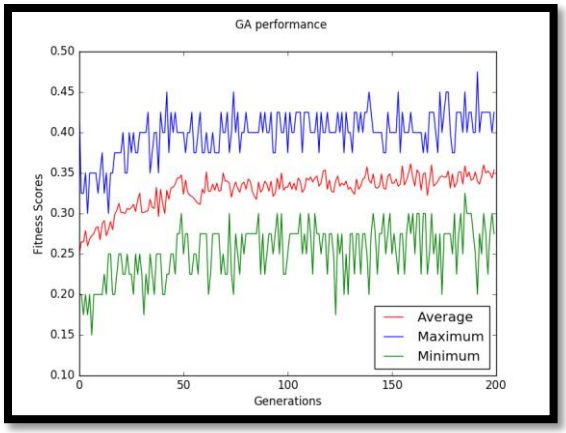

Figure 8. Fitness score over 200 generations for GA V2

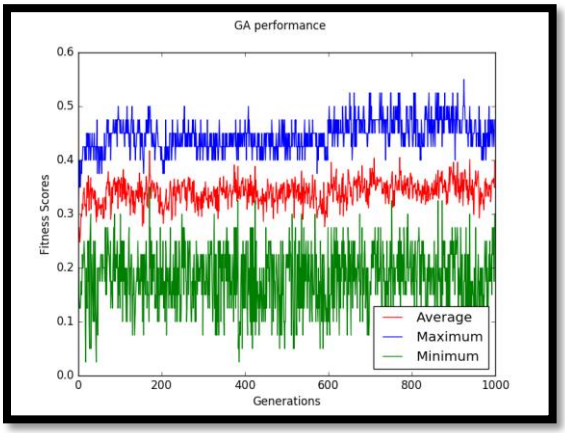

Figure 9. Fitness score over 1000 generations for GA V3

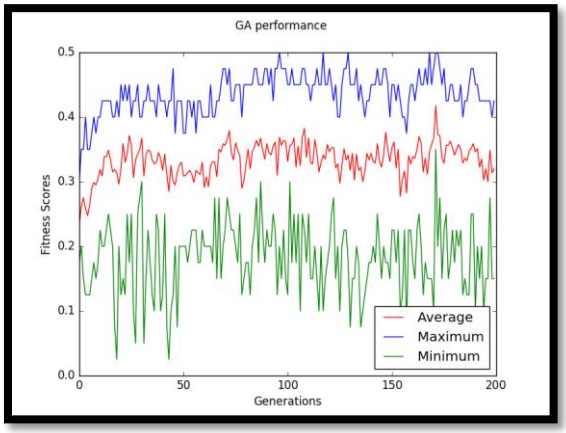

Figure 10. Fitness score over 1000 generations for GA V3

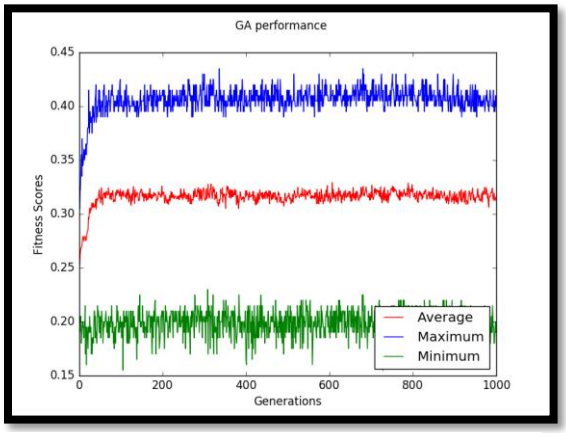

Figure 11. Fitness score over 1000 generations for larger sample and population size

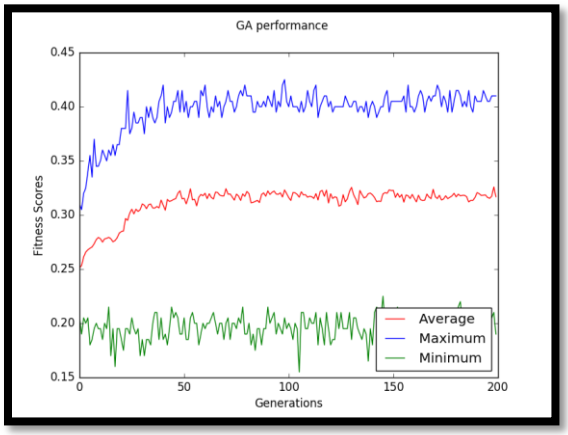

Figure 12. Fitness score over 200 generations for larger sample and population size

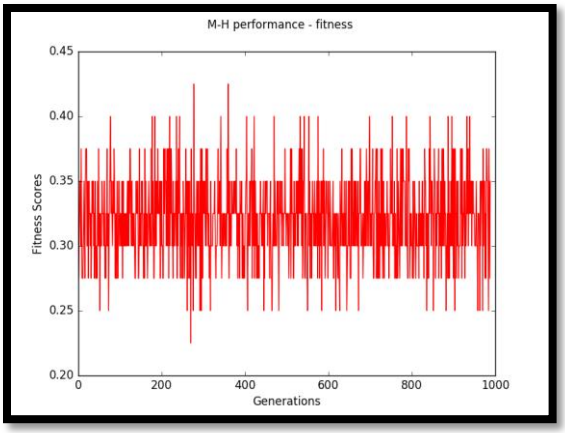

Figure 13. MH performance over 1000 generations with mutation rate 0.05

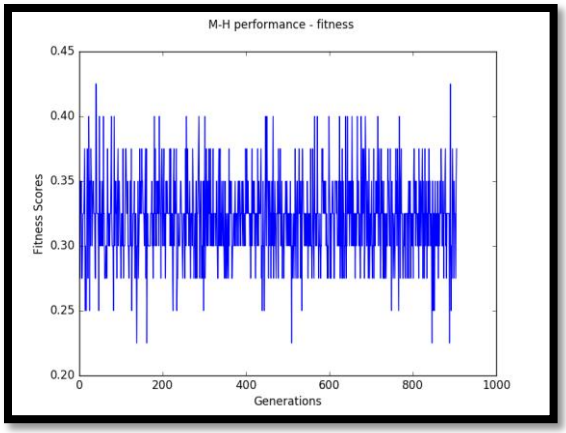

Figure 14. MH performance over 1000 generations with mutation rate 0.01

4 Conclusions

The GA is an adaptive heuristic search method that has been widely used to approximate the best solutions. It is particularly useful in feature selection due to its ability to solve high-dimensionality and multi-class problems. In this study we implemented a GA feature selection method to extract gene expression profile for cancer type prediction. A feedback linkage was established to evaluate the performance of each population through their performance on cancer prediction accuracy. Crossover, mutation and tournament selection were implemented in GA, and CART was used as the classifier. Our initial GA design revealed a quick improvement for both population fitness and individual maximum fitness. However it failed to "converge" to a narrow region. One possible reason is that the small population and sample sizes resulted in high level of variability. After increase both population size and sample size, the stability of all three measurements were significantly improved. Another explanation is that the mutation and crossover introduced too much diversity. A cooperative evolution method might help to control the diversity introduced to the population and help to stabilize the performance. Moreover, we may use proportionate selection method instead of tournament selection in the future as a less aggressive method.

To compare the performance of GA to other similar algorithms, we also implemented MH on exactly the original dataset. Just like GA, MH is good at dealing with multi-dimensional problems and is able to find good solutions over time. However, in this study MH could not achieve compelling results as GA did. No obvious improvement on classification performance was observed. One possible reason is the so called "burn-in" process, which is because of the initial sample was located in a region with very low density. As a result, the algorithm needs to run many iterations until this initial state is overcome.

In summary, we implemented a GA method to serve as a feature selector for cancer type classification using gene expression data. It successfully improved prediction accuracy and dropped irrelevant features. One important discovery is that large population and sample size could help to improve the GA performance. However, GA itself still needs to be further improved. We also applied MH on the same dataset as a comparison. Unfortunately, it failed to achieve comparable results. Further work is warranted to improve this algorithm.

References

Anon. World Health Organization. 2017. See http://www.who.int/cancer/en/.

W.K. Hastings. 1970. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika* 57, 1 (1970), 97–109. DOI:https://doi.org/10.2307/2334940

Howlader N et al. 2016. SEER Cancer Statistics Review, 1975-2013, National Cancer Institute. (April 2016).

Leo Breiman, Jerome H Friedman, Richard A Olshen, and Charles J Stone. 1984. *Classification and Regression Trees*, CRC press.

Melanie Mitchell. 1996. *An Introduction to Genetic Algorithms*, MIT Press.

Antonia Patsialou et al. 2012. Selective gene-expression profiling of migratory tumor cells in vivo predicts clinical outcome in breast cancer patients. *Breast Cancer Res. BCR* 14, 5 (2012), R139. DOI:https://doi.org/10.1186/bcr3344

Min Pei, Erik D. Goodman, William F. Punch Iii, and Ying Ding. 1995. Genetic algorithms for classification and feature extraction. In *Annual Meeting, Classification Society of North America*.

Maxim D. Podolsky, Anton A. Barchuk, Vladimir I. Kuznetcov, Natalia F. Gusarova, Vadim S. Gaidukov, and Segrey A. Tarakanov. 2016. Evaluation of Machine Learning Algorithm Utilization for Lung Cancer Classification Based on Gene Expression Levels. *Asian Pac. J. Cancer Prev. APJCP* 17, 2 (2016), 835–838.

Laura J. van ’t Veer et al. 2002. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 415, 6871 (January 2002), 530–536. DOI:https://doi.org/10.1038/415530a

Yixin Wang et al. 2005. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet Lond. Engl.* 365, 9460 (February 2005), 671–679. DOI:https://doi.org/10.1016/S0140-6736(05)17947-1