Advances in Computer Science Research (ACSR), volume 61
7th International Conference on Education, Management, Computer and Society (EMCS 2017)

Feature Selection for Single Nucleotide Polymorphisms using Parallal Genetic Algorithm

Ming Zheng[1, a] and Mugui Zhuo[1, b*]

[1]Guangxi Colleges and Universities Key Laboratory of Professional Software Technology, Wuzhou University, Wuzhou, China

[a]370505375@qq.com, [b]756456050@qq.com

*The corresponding author

Keywords: Feature selection; Mutual information; Genetic algorithm; Parallel; Single nucleotide polymorphism

Abstract. The application of statistical machine learning methods to study the association between large scale single nucleotide polymorphism (SNP) and complex diseases is facing the "Curse of dimensionality". The first task is to reduce the large scale SNP to smaller sets. For this reason, a multiple genetic algorithm is proposed for feature selection of single nucleotide polymorphism. For the first time, a new method is proposed to measure the degree of association between SNP and disease by mutual information, and as the fitness value of genetic algorithm (GA), the candidate feature SNP set is obtained by the use of genetic algorithm. In the SNP experiment on simulation data and the maximum entropy (ME) method for performance comparison shows that, this method can reduce the SNP collection and disease related SNP, while retaining the disease associated with SNP, SNP provides data for further research on the suitable scale, this method can be used in medium or large scale SNP set.

Introduction

Single nucleotide polymorphism (SNP)[1] is one of the most important tools to identify the susceptibility loci of complex diseases in human genome. At present, most of the research in this area is determined by biological experiments, and the biological research methods based on SNP are mainly based on clinical case control study and linkage disequilibrium test (LD). However, these methods have some disadvantages, such as high cost, long time period and complex experiment, so it is not feasible to study the relationship between SNP and disease in large scale genome. The use of machine learning methods can effectively remedy the defect, in the balance of individual and SNP size constraints, through the mountain climbing method is studied and the genetic algorithm to solve the small sample[2]; Some researcher Proposed by Monte Carlo logic regression method to identify SNP associated with the disease[3]; [4] used logistic regression combined with high risk identification the SNP, and puts forward the importance of the two measures to quantify portfolio SNP; [5] experimental results show GPAS method in dozens of SNP, high order the interaction can identify the high risk of breast cancer in patients with SNP; many SNP choose a software development method based on the complex algorithm implementation and configuration of the proposed integration[6]; A statistical model combining supervised classification and variable selection methods can reduce the number of thousands of genetic features as much as possible to form a simple classification rule[7].

In view of some correlations between some SNP and complex diseases, the author hopes that the study provides a feasible method, found all kinds of pathogenic factors of complex diseases at the level of SNP, the relationship model of these factors and diseases, provide tools for the early prediction and diagnosis of the disease, provide the basis for the pathogenic mechanism of disease. However, extremely serious "dimension disaster caused by the large scale SNP data" (Dimensionality Curse, DC) the phenomenon, from large-scale SNP collection to extract a small part of the SNP, which reduce the size of SNP, do not leave the SNP associated with complex diseases has become a key step to solve the problem. With the existing label SNP (tSNP) for most of the literature, the authors studied first stages

(Phase I) is characteristic of crude SNP extraction phase, reducing the number of overall SNP, solve the "dimension disaster" problem is the main objective of this phase. The goal of the second stage is to realize the accurate extraction of feature SNP in the case of a small number of SNP. In this paper, a new method is proposed for the first stage.

As a statistical optimization method, genetic algorithm has been successfully applied in various fields. The author proposes the use of mutual information (MI)[8] the possibility to measure the size and the associated SNP pathogenicity as the fitness of genetic algorithm, the genetic algorithm as the associated SNP chromosome, through iterative evolution to search for the optimal chromosome that is characteristic of SNP. In order to find out the optimal chromosome hidden in optimal chromosome under, through repeated use of genetic algorithms, the multiple optimization results and obtain a set of candidate features with frequency of SNP (Candidate SNP CSNP), according to the characteristics of SNP need to select a certain frequency, for the second phase of the research use. The author proposed multiple genetic algorithm based on mutual information (MGA)[9] to get a small number of CSNP collection, this collection as much as possible to remove the disease and unrelated to SNP, at the same time as much as possible to retain the true disease associated with SNP (GroundTruthSNP, GTSNP), to reduce the scale of SNP set objective. In the same way as the author's work, Miller proposes a method of finding the characteristic SNP set with the maximum entropy method (ME)[10] under certain constraints. The ME method first identifies 1 SNP, and then combines the SNP with the other SNP to find the best combination of SNP, by gradually increasing the length of the SNP set to determine the final feature set. The performance of the proposed MGA method is better than that of the ME method through the comparison of the performance index, such as accuracy, sensitivity, specificity, compression rate and classification error rate.

Correlation Measure of SNP and Disease

SNP Data. Usually said SNP are two allelic polymorphisms, if a allele with "A", "a" represents another allele, there is no difference between the assumption of "Aa" and "aA", then a SNP site may have 3 results of "AA", "Aa with" AA "". The 3 results are encoded by the 0, 1 and 2, respectively. The SNP data can be expressed as:

Set $X_{population} = \{(x_i, y_i), i = 1,..., N, x_i \in \{0, 1, 2\}M, y_i \in \{0, 1\}\}$, the X_i is a SNP sample of i data, Y_i identifies the sample is normal samples (0) or (1 cases). The goal of SNP selection is to identify the X_{sub} of each of the SNP, which is statistically pathogenic, from the $X_{population}$, so that any subset of the SNP is associated with a statistically significant. In the results of each subset of the SNP, that is, to find the candidate feature after the rough extraction SNP.

SNP Subset and Disease Association Measure. The mutual information in information theory is introduced to measure the relevance of a SNP subset of X_{sub} (for the sake of convenience, the following is replaced by X) with the disease, see form (1).

$$\mathrm{MI}(Y;X) = H(Y) - H(Y|X) \tag{1}$$

Among them, H (Y) and H (Y|X) are entropy and conditional entropy, defined as follows:

$$\begin{cases} H(Y) = -\sum_{y\in\{0,1\}} p(y)\log p(y) \\ H(Y|X) = -\sum_{y\in\{0,1\}}\sum_{x\in\{0,1,2\}} p(y|x)\log p(y|x) \end{cases} \tag{2}$$

Among them, the probability of all normal samples, P cases (y), $y \in \{0, 1\}$ and $X \in \{0,1,2\}$ on the probability of the normal samples and samples of the cases (Y|X), $y \in \{0, 1\}$ can use simple data system counting numbers, without the probability density estimation of complex.

Feature Extraction

For a single SNP model, no matter how much the number of SNP, the mutual information can be obtained by direct computation. But as the candidate set and SNP Model contains the increase in the number of SNP at the same time, the cost of mutual information is obtained by using Model direct method of calculating the time cost will present exponential growth. Because of this, the ME algorithm takes a lot of time. Therefore, we consider the search of the search space by genetic algorithm to find multiple SNP sets with large mutual information.

First, set the size of the initial population (e.g., 100), and use different SNP combinations as chromosomes. The GTSNP contains n Model called n_Way Model, such as SNP45, SNP78 and SNP99 is 3_Way Model, the corresponding chromosome shown in Fig. 1:

SNP45	SNP78	SNP99

Figure 1. Chromosome of 3-Way Model

Experimental Design and Results

The preliminary experiments show that the 100 SNP set of the evolution of 2_Way chromosome on behalf of the Model algebra can be around 600 times, the evolution of the 3 _Way can be optimized to the best. With the increase of the scale of SNP, the number of iterations can be increased accordingly. The number of iterations of the genetic algorithm can not be set, but the mutual information between the chromosome and the last iteration in a number of times (such as 100) in the case of the same has been considered to have evolved to the best. 2_Way Model chromosome genetic algorithm for the 8 genetic algorithm to get the results of the optimal subset of features SNP, 3_Way is the 5. Finally, all the SNP subsets obtained from these two genetic algorithms are obtained and a candidate set is obtained: 31 SNP (2_Way:2 = 8 =; 3_Way:3 = 5). The algorithm is repeated 20 times, and a total of 620 SNP are obtained. Statistics of the frequency of each SNP in the 620 SNP set, according to the frequency from large to small sort, according to the need to remove a frequency of more than SNP as a candidate SNP set. A total of 17 GTSNP, with a selection of 2 _Way with a number of times of 3 _Way with a number of times of 5 is found for the maximum number of GTSNP not exceeding the set value of 17. In the case of GTSNP is unknown, the number can be envisioned by looking for the GTSNP to set the 2_Way and 3_Way.

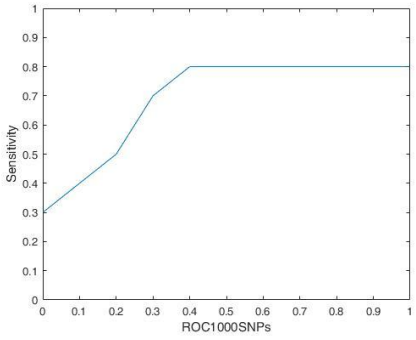

Figure 2. Receiver operating characteristic curve

It can be seen that the MGA method has higher accuracy under the ROC curve. This shows that in the first few features, each additional feature, the classification ability is significantly increased, the classification error rate is significantly reduced. As can be seen from Fig. 2, compared with the ME method, the MGA method in the case of the same number of features, with the exception of a few points, the feature classification ability is higher than the ME method.

Summary

The feature extraction of SNP multiple genetic algorithm based on mutual information collection, compression mass SNP, extracted from the set of candidate features smaller SNP, including GTSNP and GTModel as much as possible, to lay a good foundation for the next step of the research on Phase I. The time complexity of the MGA method proposed by the authors does not increase with the increase of the number of SNP.

Acknowledgements

This work was supported by grants from The National Natural Science Foundation of Chi-na (No. 61502343, No. 61373051, and No. 61402423), China Postdoctoral Science Foun-dation funded(No. 2016M590260), the Guangxi Natural Science Foundation (No. 2015GXNSFBA139262), the Science Research Funds for the Guangxi Universities (No. KY2015ZD122), Guangxi Colleges and Universities Key Laboratory of Professional Software Technology, Wuzhou University.

References

[1] JOLLIFFE D A, WALTON R T, GRIFFITHS C J, et al. Single nucleotide polymorphisms in the vitamin D pathway associating with circulating concentrations of vitamin D metabolites and non-skeletal health outcomes: Review of genetic association studies[J]. Journal of Steroid Biochemistry and Molecular Biology, 2016, 164: 18-29.

[2] YANG C H, CHUANG L Y, CHENG Y H, et al. Single nucleotide polymorphism barcoding to evaluate oral cancer risk using odds ratio-based genetic algorithms[J]. Kaohsiung Journal of Medical Sciences, 2012, 28(7): 362-368.

[3] LIU X Y, WANG Y P, SRIRAM T N. Determination of sample size for a multi-class classifier based on single-nucleotide polymorphisms: a volume under the surface approach[J]. Bmc Bioinformatics, 2014, 15.

[4] KASTELIC V, DROBNIC K. A single-nucleotide polymorphism (SNP) multiplex system: the association of five SNPs with human eye and hair color in the Slovenian population and comparison using a Bayesian network and logistic regression model[J]. Croatian Medical Journal, 2012, 53(5): 401-408.

[5] NUNKESSER R, BERNHOLT T, SCHWENDER H, et al. Detecting high-order interactions of single nucleotide polymorphisms using genetic programming[J]. Bioinformatics, 2007, 23(24): 3280-3288.

[6] CHEN C D, YU Z Q, CHEN X L, et al. Evaluating the Association between Pathological Myopia and SNPs in RASGRFI, ACTCI and GJD2 Genes at Chromosome 15q14 and 15q25 in a Chinese Population[J]. Ophthalmic Genetics, 2015, 36(1): 1-7.

[7] MEHTA A, RAMACHANDRA C J A, MURA M, et al. Single nucleotide polymorphisms discriminates between symptomatic and asymptomatic LQTS2 patients: A DNA-based patient classification[J]. European Heart Journal, 2015, 36: 1070-1070.

[8] VERGARA J R, ESTEVEZ P A. A review of feature selection methods based on mutual information[J]. Neural Computing & Applications, 2014, 24(1): 175-186.

[9] BEIKNEJAD D, CHAICHI M J, FATEMI M H. Prediction of photolysis half-lives of dihydroindolizines by genetic algorithm-multiple linear regression (GA-MLR)[J]. Journal of Physical Organic Chemistry, 2016, 29(6): 312-320.

[10] PIRES C A L, PERDIGAO R A P. Minimum Mutual Information and Non-Gaussianity through the Maximum Entropy Method: Estimation from Finite Samples[J]. Entropy, 2013, 15(3): 721-752.