Single Nucleotide Polymorphisms Data Analysis using Improved Ant Colony 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

Keywords: Single nucleotide polymorphisms; Ant colony algorithm; Data analysis; Data prediction; Optimal strategy.

Abstract. Improved classical ant colony clustering algorithm (LF algorithm), applied to salt sensitive hypertension SNPs data analysis, in order to explore the high throughput SNPs statistical analysis to provide new ideas. The LF algorithm was improved, and the improved algorithm was programmed with Mat 1ab8.0 software. The cluster analysis was performed on 335 samples of salt sensitive hypertension. The LF algorithm was successfully improved and the software interface was realized. Using the new algorithm, all samples are divided into 2 categories, the first class of 169 samples, second of 166 samples of consistency test and latent class analysis results, the Kappa value is 0.93, P<0.001, and the two kinds of differences in population SNPs probability distribution statistical test, we selected 3 SNPs:rs848307, rs1739843, rs1010069, clear it plays an important role in the classification of. Conclusion: ant colony clustering algorithm has the characteristics of unique thinking, automatic calculation, easy to improve, etc. it has broad application prospects in the field of high throughput SNPs data analysis and other related fields of genomics.

Introduction

Single nucleotide polymorphisms (SNPs)[1] refers to DNA sequence polymorphism in the genome level caused by a single nuclear bitter acid variation, in the human genome, the average per thousand bases of a SNPs[2]. The SNPs of patients with complex disease has the characteristics of multi-level and multi dimension. It is an important task how to use SNPs data to find various problems related to disease[3]. The ant colony optimization (ACO)[4] algorithm does not depend on the specific problems in the mathematical description, with global optimization, distributed, self-organizing and adaptive ability, then many researchers inspired by ant behavior, a series of ant colony algorithm design, which based on the ant heap of corpses piled up behavior, put forward a basic model (BM), and has been successfully applied to robot. The BM model is extended and the similarity measure of the data object is given, and the LF algorithm for data clustering is designed[5]. Clustering analysis of common SNPs data have their own limitations, compared with ant colony clustering algorithm has the unique thinking and operation automation features 3 and the improvement measures to improve the algorithm has good clustering effect of defects. For this purpose, the LF algorithm is improved and the software is realized on the computer, and then the results of clustering analysis are compared with the improved LF algorithm and the latent class analysis method. The data of this study from the Capital Medical University School of public health cooperation with Beijing City, some community health service centers, the salt sensitive hypertension SNPs information of 335 sample test results, including 20 genes, 29 susceptible SNPs. Each SNP is divided into heterozygous and homozygous forms[6].

Method

Procedure. Ant colony clustering algorithm LF algorithm is the most classic, the calculation process is as follows: first, initialization for clustering, all data in the scatter form, random distribution in the two-dimensional plane form, then add a certain amount of artificial ants. Then enter the iterative process, for any ant in each iteration, the current location of the data to calculate the neighborhood similarity. If the data is carried by ants, the probability of calculating the ants to lay down the data Pd, if the data is in the grid, the probability of calculating the ants pick up the data Pp. The two should be compared with the random probability Pr, to determine whether the data operation, the algorithm has been carried out until the set number of iterations.

The concept of LF algorithm is introduced into the local memory, each ant will have data object handling and put down the position information, into their own memory, once the data raised will be the first with their memory storage compared to find the nearest data points, jump directly to the data points the adjacent to the ant memory length is more limited and mobile principles can set all the ants, put down the operations on all data continue to lift or, after several iterative times after 3 dimensional form, similar to a relatively high degree of data will be transported to the same area, so as to realize the clustering process the data, the calculation process is shown in Fig. 1.

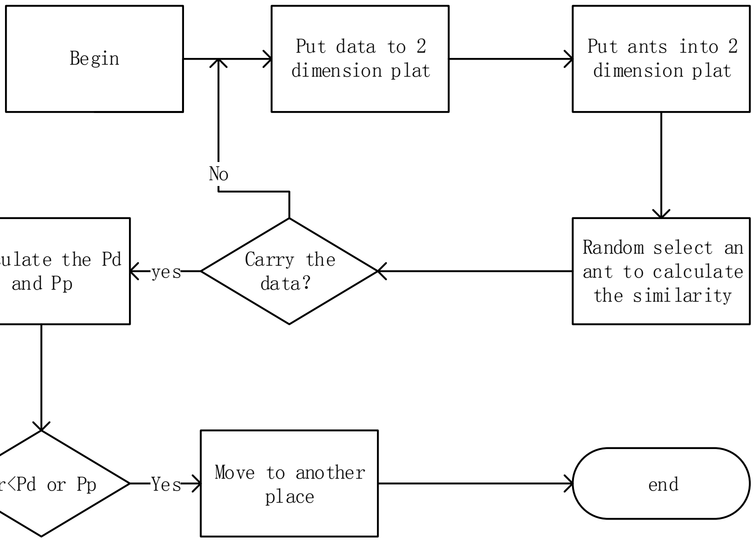
Begin
Put data to 2 dimension plat
Put ants into 2 dimension plat
No
Calculate the Pd and Pp
yes
Carry the data?
Random select an ant to calculate the similarity
No
Pr<Pd or Pp
Yes
Move to another place
end

Figure 1. Flow chart of LT algorithm

Experiment

In the improved scheme, the guidance of the global memory is established, and the capacity adjustable memory bank is set up, and the information of all the data that the ants put down is recorded in the database and S is shared by all the ants. By changing the method of neighborhood clustering in the improved algorithm in the process of experiment 171 verification method using neighborhood linear decreasing, can increase the calculation efficiency, this research will decrease to 4 from the 12 linear neighborhood[10].

Using spherical coordinates, the four boundaries of the two-dimensional grid are connected together to form a spherical surface to ensure that the data points, regardless of location, the same neighborhood.

The clustering data into the improved LF algorithm procedures, using Matlab8.0 software programming. The program is small; and the sampling error, the program will run 10 times, of which a calculation results are shown in Table 2, shown in Figure 2, taking the average of the results. The results are as follows: all 335 samples are divided into 2 categories: the first category is and the other is the Z of the second category. The improved ant colony clustering results and the use of latent class analysis (latent class analysis, LCA) to compare the clustering results, both have the same classification results for 324 main samples, after Kappa test, the Kappa coefficient between the two results is 0.93, according to international standard, Kappa 0.75 can be considered quite satisfactory coefficient of maximal consistent degree.

Using the above data sets were compared to verify the.SNPRuler algorithm is to predict the rule reasoning and two stage based on this method and SNPRuler (Two stage) design strategy, learn the relationship between feature and class variables by pre rules, predict the class label and then in the test data. The epistasis detection using the rules of learning, the reason is: first the combination contains, some patterns or prediction rules; furthermore, looking for assessment rules easier and more efficient. Therefore, the SNPRuler method by mining prediction rules to find potential epistatic combinations. The result can be seen from Fig. 2 and Fig. 3 as below:

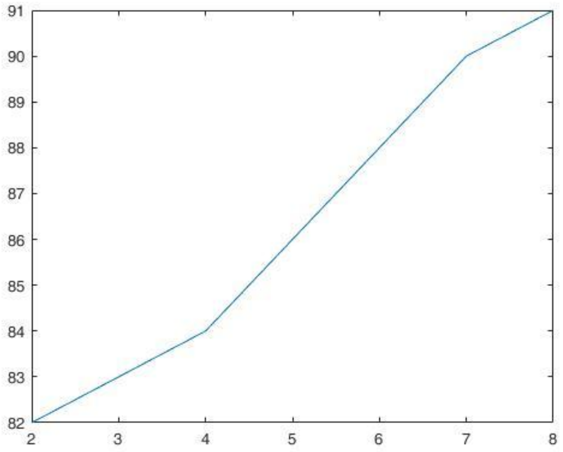

Figure 2. The prediction of dataset 1.

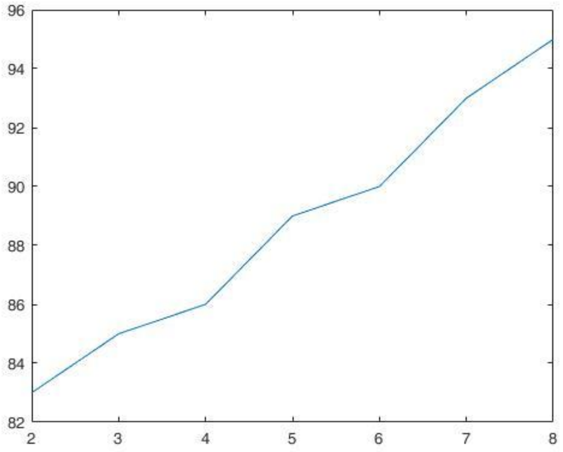

Figure 3. The prediction of dataset 2.

Conclusion

In this paper, one of the methods of cluster analysis as a data mining technique, the purpose of which is to divide the data without supervision, so that the final classification of the data in the highest degree of similarity of 3 and not the same data in the lowest degree of similarity. The data clustering analysis has two purposes, is to find the final classification for the characteristics of each type were analyzed; two is as a preprocessing method to reduce dimension data of other analysis, this study implements second objective clustering analysis by ant colony clustering algorithm, however in the study the algorithm can achieve the first objective of service. In a word, the ant colony clustering algorithm has become a kind of intelligent algorithm based on its unique conception, simple programming and easy to improve.

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] 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.

[2] 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.

[3] ASKAR M, MAJHAIL N S, RYBICKI L, et al. Single Nucleotide Gene Polymorphisms (SNP) in the Gamma Block of the Major Histocompatibility Complex (MHC) Are Independent Risk Factors for Severe Acute Graft Versus Host Disease (GVHD) in Unrelated Donor Hematopoietic Cell Transplantation (HCT)[J]. Biology of Blood and Marrow Transplantation, 2015, 21(2): S326-S327.

[4] ACHARY R, VITYANATHAN V, RAJ P, et al. Dynamic Job Scheduling Using Ant Colony Optimization for Mobile Cloud Computing[M]. // BUYYA R, THAMPI S M. Intelligent Distributed Computing. City, 2015: 71-82. <Go to ISI>://WOS:000347783600007.

[5] SOUSA M, LOPES W, MADEIRO F, et al. Cognitive LF-Ant: A Novel Protocol for Healthcare Wireless Sensor Networks[J]. Sensors, 2012, 12(8): 10463-10486.

[6] 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.

[7] DORIGO M, BONABEAU E, THERAULAZ G. Ant algorithms and stigmergy[J]. Future Generation Computer Systems-the International Journal of Escience, 2000, 16(8): 851-871.

[8] KOROSEC P, SILC J. The continuous differential ant-stigmergy algorithm for numerical optimization[J]. Computational Optimization and Applications, 2013, 56(2): 481-502.

[9] KOROSEC P, SILC J, FILIPIC B. The differential ant-stigmergy algorithm[J]. Information Sciences, 2012, 192: 82-97.

[10] 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.