Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 8, No. 3, 2023

The Application of Artificial Intelligence to The Bayesian Model Algorithm for Combining Genome Data

Zepeng Shen[1, *], Kuo Wei[2], Hengyi Zang[3], Linxiao Li[4], Guanghui Wang[5]

[1] Network Engineering, Shaanxi University of Technology, Shaanxi 723001, China
[2] Computer Science, Individual Contributor, Shenzhen, China
[3] Big Data and Business Intelligence, International University, Isabella I of Castile, Burgos, Castile and León, Spain
[4] Communication Engineering, Peking University, Beijing, China
[5] Computer Science, East China University of Science and Technology, Shanghai, China

[*] Corresponding author: Farid Uddin Ahmed (Email: lubyliu45@gmail.com)

Abstract: The combination of bioinformatics and artificial intelligence (AI) has made significant progress in the current phase. AI technology, especially deep learning, has been widely used in biology, resulting in many innovations. Currently, AI plays a key role in genomics, proteomics, and drug discovery. Deep learning models are used to predict protein structures, discover potential drug compounds, interpret genomic sequences, analyze medical images, and make personalized medical recommendations. In addition, AI can also help accelerate the processing and interpretation of biological big data, helping biologists to understand complex problems in the life sciences more deeply. Compared to traditional genetic data analysis methods, AI combined with bioinformatics methods are often faster and more accurate, and are capable of processing large-scale, high-dimensional biological data, opening up unprecedented opportunities for life science research. In this paper, the whole genome data and biomedical imaging data are used from the perspective of Bayesian hypothesis testing. Genome-wide association analysis, led by large-scale multiple tests, is a very popular tool for identifying genetic variation points in new complex diseases. In genome-wide association analysis, tens of thousands of SNPS need to be tested simultaneously to find out some SNPS related to traits. These tests are related due to factors such as linkage imbalances in the genetic process, and the test questions are set against the background of high-dimensional data (p >=n).

Keywords: Bayesian hypothesis; Large-scale genomes; Artificial intelligence algorithm.

1. Introduction

Bayesian network is a probabilistic graph model. On the one hand, it uses the structure of directed acyclic graph to describe the intrinsic relation between random variables from a qualitative point of view. On the other hand, on the basis of directed acyclic graph structure, the problem is decomposed by using the knowledge of probability theory, and the conditional probability parameters of each parent and child node are used to quantitatively describe the degree of influence between each variable, and the complex joint distribution is simplified into a series of simple local distribution! . Bayesian networks can not only well mine the complex dependence relationship between variables, but also have powerful reasoning ability, and the whole theory is easy to understand and learn, and has become a very beneficial tool to understand the uncertainty problem. In recent years, researches on Bayesian networks have emerged in an endless stream, and Bayesian network models have been applied to various fields, such as medicine and health technology [1], engineering [2], and environmental science [3].

The Bayesian network model of large-scale genomic data based on AI technology to solve practical problems often includes three steps, namely structure learning, parameter learning and knowledge reasoning. The process of structure learning is the process of training a directed acyclic graph that can fit the relationship between variables to the greatest extent through some structure learning algorithm. Parameter learning is also called parameter estimation[4]. After completing structure learning, conditional probability parameters in the conditional probability distribution table of each node need to be learned to describe the correlation strength between variables from a quantitative perspective. Thus, Bayesian network learning is completed.

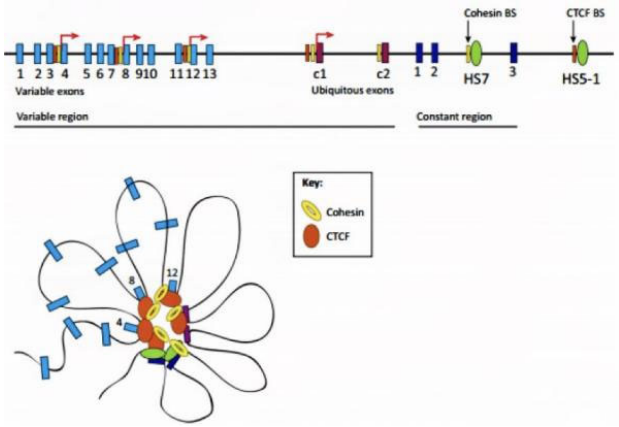

Figure 1. Construct genome-wide interaction network and gene expression regulation network; Construct a 3D structure map of chromatin

Therefore, based on the expression profile data, a network model of gene interaction can be established, which can be called reverse engineer ing[5]. Common gene regulatory network models include Boolean network model, linear combination model, Markov model, weighted matrix model, mutual information association model and Bayesian network model.

2. Related Work

2.1. Bayesian method

The Bayesian network parameter is to calculate the posterior probability P(θ|S) and use it as the premise for estimating the parameter. The basic idea of Bayes is a relatively complete set of instance data D and a topological structure S and a distribution with unknown parameters, θ represents a random variable, its prior distribution is denoted P(θ), and P(θ) follows uniform distribution, P(θ|S) is called the posterior probability of the parameter θ. If the Dirichlet distribution is used as the prior distribution, then:

$$P(\theta_{ij} \mid S) = Dir(\alpha_{ij1}, \alpha_{ij2}, ...\alpha_{ij3}) = \frac{\Gamma(\alpha_{ij})}{\prod_k \Gamma(\alpha_{ijk})} \prod_k [\theta_{ijk}]^{\alpha_{ijk}} \quad (1)$$

If the above formula is used in the case of incomplete genomic data, θ as a parameter transforms into the following algorithm:

$$P(\theta_{ij} \mid S, D) = \frac{P(\theta_{ij} \mid S) P(D \mid S, \theta_{ij})}{P(D)} = \frac{\Gamma(\alpha_{ij} + n_{ij})}{\prod_k \Gamma(\alpha_{ijk} + n_{ijk})} \prod_k [\theta_{ijk}]^{\alpha_{ijk} + n_{ijk}} = Dir(\alpha_{ij1} + n_{ij1}, \alpha_{ij3} + n_{ij2}, ...\alpha_{ijk} + n_{ijk},) \quad (2)$$

Where nijk meets the condition in sample set D :Xi=xik, and the verified Bayesian grid result is:

$$\hat{\theta}_{ijk} = \frac{\alpha_{ijk} + n_{ijk}}{\sum_k [\alpha_{ijk} + n_{ijk}]} = \frac{\alpha_{ijk} + n_{ijk}}{\alpha_{ij} + n_{ij}} \quad (3)$$

For the probability parameters of incomplete genetic data, their main idea is the same, both are to use the existing data to infer the missing data, and then continuously repair the missing data to improve the database, so as to achieve cyclic iteration to gradual refinement.

2.2. ESL-GA algorithm

ESL-GA algorithm uses an elite-guided adaptive genetic algorithm to train the optimal Bayesian network structure, which reduces the time complexity of the algorithm to a certain extent, and also provides a practical method for better learning Bayesian networks of different sizes under small training data, which makes me see a breakthrough to solve the above pain points[6-8]. Therefore, this paper introduces the idea of ensemble learning, takes ESL-GA algorithm as the base learner and uses the fusion strategy of weighted average method to construct the Bayesian network hybrid structure learning algorithm EN-ESL-GA.

In order to further improve the accuracy and reliability of ESL-GA algorithm in learning the final structure of Bayesian networks, and make the ESL-GA algorithm better applied to practical problems, a new Bayesian network structure learning algorithm EN-ESL-GA is proposed by introducing parallel integration idea into ESL-GA algorithm.

$$w^t_{i,j} = \frac{\sum_{k=1}^{m} f(x_{tk}) \cdot \delta(x_{tk,j} = i)}{\sum_{k=1}^{m} f(x_{tk})}. \quad (4)$$

The EN-ESL-GA algorithm performs one ensemble learning on the Bayesian network structure learned multiple times based on the ESL-GA algorithm, and the training number of the base learner is usually set to 20. Firstly, 20 sample datasets are randomly sampled from the original datasets. ESL-GA algorithm is used to learn the structure of Bayesian networks on each sample dataset, and the adjacency matrix and BDeu score of each learned Bayesian network are recorded. Then, the maximum and minimum BDeu scores of each network are normalized as the weights of the corresponding adjacency matrix, and the weighted average of all adjacency matrices is performed to obtain a transition matrix representing the strength of causality between variables[9]. The edge with weak causality in the transition matrix is filtered out according to the pre-set rush value, so as to obtain a fusion matrix. Finally, the fusion matrix is de-looped and the maximum number of parent nodes is restricted to obtain the integrated Bayesian network structure.

3. Methodology

3.1. Evaluation index

To evaluate the performance of a Bayesian network structure learning algorithm, Fi, Sensitivity and Specificity were used to evaluate the learned Bayesian network structure. The relevant formulas are shown in formulas (5 ~ 7). Where, F balances Precision and Recall; Sensitivity (recall rate), that is, the true positive rate, represents the ratio of the number of correct edges learned in the current network to the number of edges present in the general network[10-11]; Specificity represents the ratio of the number of edges that do not exist in both the current network and the end of the universal network to the number of edges that do not exist in the general network, that is, the true negative rate. The closer these three indicators are to 1 at the same time, the more accurate the final structure of the Bayesian network we learn.

$$Sensitivity = Recall = \frac{TP}{TP+FN}. \quad (5)$$

$$Precision = \frac{TP}{TP+FP}. \quad (6)$$

$$F_1 = \frac{2 \times Recall \times Precision}{Recall + Precision}. \quad (7)$$

$$Specificity = \frac{TN}{FP+TN}. \quad (8)$$

3.2. Experiment and result

In order to explore the Bayesian network structure learning performance of EN-ESL-GA algorithm, K2 and Max are used

the Hybrid Structure Learner Using Genetic Algorithms, Maximum Weight Spanning Tree (MWST), The Hybrid Structure Learner using Genetic Algorithms, the Diversity-guided Site-specific Rate Genetic Algorithm, HSL-GA, The diversity-guided Site-specific rate genetic Algorithm, DiGSiR-GA) [12] DiGSiR-GA (the DiGSiR-GA Using the Parent Set Crossover Operator, DiGSiR-GA-PSX) and ESL-GA[13] are used as comparison algorithms. Taking into account the randomness of the data set as well as the algorithms themselves, all algorithms were run on 20 different random samples of the same size, recording the average performance metrics of each algorithm. In addition, since both K2 algorithm and MWST algorithm are deterministic algorithms, in order to ensure fairness, the node order is randomly determined as the input of K2 algorithm in each experiment without prior knowledge. Similarly, because the MWST algorithm needs to give the root node in advance, it uses the randomly generated root node to learn the structure 20 times. The method parameter Settings of other genetic algorithms on genetic databases are consistent.

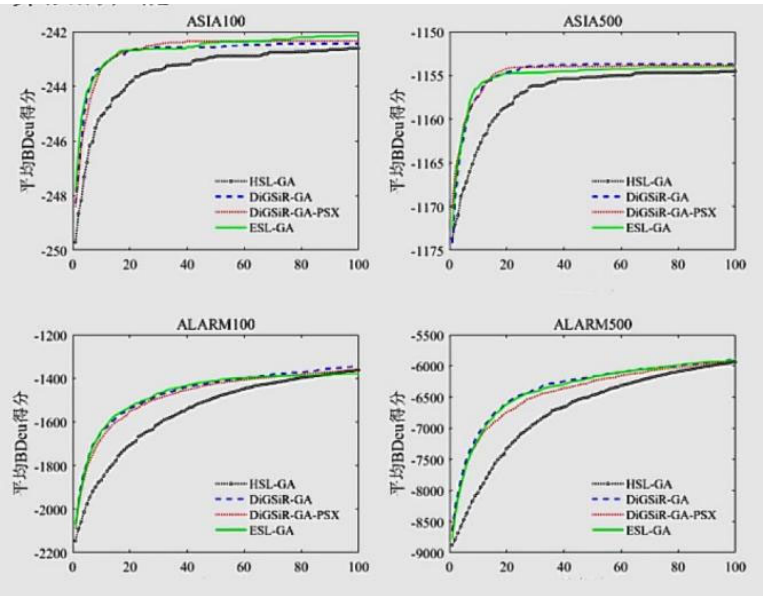

Figure 2. Variation curve of average BDeu score of Bayesian network structures under different data sets

Although EN-ESL-GA algorithm solves the problem of low accuracy and reliability in learning Bayesian network structure with small training data set, its time complexity also increases greatly with the increase of network scale, which is mainly spent on calculating multiple BDeu scores[14]. Therefore, this algorithm is only suitable for learning Bayesian network structure in small and medium-sized networks with limited hardware conditions.

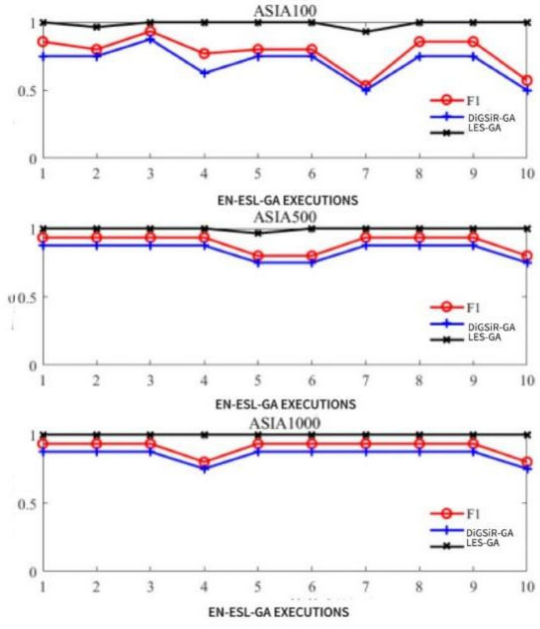

Figure 3. Genome data EN-ESL-GA algorithm and MA-EN-ESL-GA algorithm performance variation diagram

As can be seen from the experimental results, there are many different forms of genomic data algorithm analysis model under AI technology, among which EN-ESL-GA algorithm and traditional Bayesian algorithm are more popular. Combined with Bayesian network, it is a probability graph model, which uses Bayesian probability to represent the dependency relationship between variables. Bayesian parameter learning is usually used to estimate the parameters of the network[15]. ESL-GA is more focused on finding the best network structure through search algorithms. And Bayesian networks usually require prior information that explicitly specifies the probability distribution, which may require input of domain knowledge. ESL-GA does not necessarily need this prior information, it can search to find the best structure, and then can estimate the parameters through parameter learning. Combined with the global search capability of genetic algorithm and the probabilistic modeling capability of hybrid Bayesian networks, potential probabilistic relationships can be discovered in large-scale genomic data. Bayesian networks can often handle uncertainty and noise between variables, making them robust in genomic data analysis. Finally, genetic algorithms can optimize the network structure through the evolutionary process and help find the best model suitable for the data.

4. Conclusion

Artificial intelligence data mining technology is widely used in today's business and industry, but its application in biomedicine, especially in the field of genomic information analysis, is still in the initial stage. In order to have greater application value in medicine, we need to actively explore

data mining. In the face of the complexity and particularity of medical information acquisition, the mining algorithm also has high efficiency and robustness[16-17]. Therefore, in the future, genome data analysis algorithms based on AI may continue to develop, including the following directions: More complex deep learning models: Deep learning has shown great potential in genomic data analysis, and more complex deep learning models may appear in the future, which can better mine the information in genomic data; Integration methods, which integrate different algorithms (such as genetic algorithms, Bayesian networks, deep learning, etc.) together for better performance and robustness; Reinforcement learning, which uses reinforcement learning to optimize the decision-making process of genomic data analysis to achieve more precise results; Large-scale data processing, with the accumulation of genomic data, the development of efficient large-scale data processing algorithms will become more important[18-20].

Overall, the field of genomic data analysis remains an area full of challenges and opportunities, and future developments are likely to combine multiple AI techniques to better understand the complexity and biological significance of genomic data.

References

[1] WATTS DJSTROGATZ S H.Collective dynamicsofsmall-word networks [J]. Nature, 1998,(393):440-442.\

[2] BARABASI A L,ALBERT R.Emergence of scalingin random networks [J].Science,1999,(286):509-512.

[3] BARABASI A L,ALBERT R,JEONG H,etal.Power-Law distribution of the World Wide Web [J] Science, 2000, 287 (5461):2115.

[4] Gasser, Thomas, and Manu Sharma, "Computational Approaches for Understanding the Diagnosis and Treatment of Parkinson's Disease", Journal of Neurology, 2015.

[5] Chang Che, Bo Liu, Shulin Li, Jiaxin Huang, and Hao Hu. Deep learning for precise robot position prediction in logistics. Journal of Theory and Practice of Engineering Science, 3(10):36–41, 2023.DOI: 10.1021/acs.jctc.3c00031.

[6] Hao Hu, Shulin Li, Jiaxin Huang, Bo Liu, and Change Che. Casting product image data for quality inspection with xception and data augmentation. Journal of Theory and Practice of Engineering Science, 3(10):42–46, 2023. https://doi.org/10.53469/jtpes.2023.03(10).06

[7] Chang Che, Qunwei Lin, Xinyu Zhao, Jiaxin Huang, and Liqiang Yu. 2023. Enhancing Multimodal Understanding with CLIP-Based Image-to-Text Transformation. In Proceedings of the 2023 6th International Conference on Big Data Technologies (ICBDT '23). Association for Computing Machinery, New York, NY, USA, 414–418. https://doi.org/10.1145/3627377.3627442

[8] Y. Wang, K. Yang, W. Wan, Y. Zhang and Q. Liu, "Energy-Efficient Data and Energy Integrated Management Strategy for IoT Devices Based on RF Energy Harvesting," in *IEEE Internet of Things Journal*, vol. 8, no. 17, pp. 13640-13651, 1 Sept.1, 2021, doi: 10.1109/JIOT.2021.3068040.

[9] Y. Wang, K. Yang, W. Wan, Y. Zhang and Q. Liu, "Energy-Efficient Data and Energy Integrated Management Strategy for IoT Devices Based on RF Energy Harvesting," in IEEE Internet of Things Journal, vol. 8, no. 17, pp. 13640-13651, 1 Sept.1, 2021, DOI: 10.1109/JIOT.2021.3068040.

[10] Wang, Y, Yang, K, Wan, W, Mei, H. Adaptive energy saving algorithms for Internet of Things devices integrating end and edge strategies. Trans Emerging Tel Tech. 2021; 32:e4122.DOI: https://doi.org/10.1002/ett.4122

[11] Xu, J., Pan, L., Zeng, Q., Sun, W., & Wan, W. Based on TPUGRAPHS Predicting Model Runtimes Using Graph Neural Networks. https://api.semanticscholar.org/Corpus

[12] Yao, J., Zou, Y., Du, S., Wu, H., & Yuan, B. Progress in the Application of Artificial Intelligence in Ultrasound Diagnosis of Breast Cancer. DOI:https://api.semanticscholar.org/Corpus

[13] Zhou Y, Chen S, Wu Y, Li L, Lou Q, Chen Y, Xu S. Multi-clinical index classifier combined with AI algorithm model to predict the prognosis of gallbladder cancer. Front Oncol. 2023 May 10;13:1171837. DOI: 10.3389/fonc.2023.1171837. PMID: 37234992; PMCID: PMC10206143.

[14] Li L, Xu C, Wu W, et al. Zero-resource knowledge-grounded dialogue generation[J]. Advances in Neural Information Processing Systems, 2020, 33: 8475-8485. DOI: https://doi.org/10.48550/arXiv.2008.12918

[15] Lin, Q., Che, C., Hu, H., Zhao, X., & Li, S. (2023). A Comprehensive Study on Early Alzheimer's Disease Detection through Advanced Machine Learning Techniques on MRI Data. Academic Journal of Science and Technology, 8(1), 281–285.DOI: 10.1111/jgs.18617

[16] Che, C., Hu, H., Zhao, X., Li, S., & Lin, Q. (2023). Advancing Cancer Document Classification with R andom Forest. Academic Journal of Science and Technology, 8(1), 278–280. https://doi.org/10.54097/ajst.v8i1.14333

[17] Zheng Yang, Tien Tuan Anh Dinh, Chao Yin, Yingying Yao, Dianshi Yang, Xiaolin Chang, and Jianying Zhou. "LARP: A Lightweight Auto-Refreshing Pseudonym Protocol for V2X." Proceedings of the 27th ACM on Symposium on Access Control Models and Technologies, 2022, pp. 49-60. DOI:https://doi.org/10.1145/3532105.3535027

[18] KWAK J G， LEE J. Thermoresponsiveinvertedcolloidal crystal hydrogel scaffolds for lymphoid tissueengineering [J]. Advanced Healthcare Materials ,2020.9(6): 1901556:1-9.

[19] ALBERT RBARABASI LStatistical mechanicsof complex networks[J].Reviews of Modern Physics,2002(74):47-97.

[20] GIRVAN M,NEWMAN M E J.Community structure in social and biological networks[J].Proc of the National Academy of Science,2002,9(12):7821-7826.