Efficient Ensemble Feature Selection Based Boolean Modelling For Genetic Network Inference

Hasini Nakulugamuwa Gamage, Madhu Chetty, Adrian Shatte
School of Engineering, IT and Physical Sciences
Federation University
Churchill, Victoria 3842, Australia

Jennifer Hallinan
BioThink Pty Ltd.
Brisbane, Queensland, Australia

Correspondence: hasininakulugamuwagamage@students.federation.edu.au

Abstract*—The reconstruction of Gene Regulatory Networks (GRNs) is important in systems biology, because GRNs can provide insight into regulatory interactions between genes. Various computational methods have been developed for this task, but most have low computational efficiency. In this paper, we introduce an ensemble feature selection approach used with Boolean network modelling for efficient and accurate inference of GRNs. Using discretized microarray expression data, the regulatory genes for each target gene are inferred using an estimated multivariate mutual information-based feature selection method. To remove irrelevant features, pair-wise mutual information score-based thresholding is used, and gene-wise precision and dynamic accuracy-based stopping criteria is used for the determination of the maximum indegree of a target gene. Further inference of regulatory genes is performed by ReliefF, an instance-based feature ranking method. We also introduce a new Append function, to obtain a single optimal set of regulatory genes by combining the selected sets of genes from the MRMR and ReliefF based on performance evaluation criteria. Our previous research finding, a Pearson correlation coefficient based Boolean modelling approach is utilized in this research for the efficient observation of the optimal regulatory rules associated with target genes and the selected regulatory genes. Experiments, evaluating the proposed approach are ongoing. To date we have obtained improved results in terms of structural accuracy and efficiency.

***Keywords*—Gene Regulatory Networks, ensemble feature selection, Boolean network, Min-Redundancy Max-Relevance, stopping criteria, ReliefF, Append function**

I. Introduction

Biological processes are regulated through connections between genes and their products, forming *Gene Regulatory Networks* (GRNs). Accurate and efficient identification of the structure and the dynamics of these GRNs is one of the key challenges in systems biology. To tackle this problem, different mathematical and computational models have been developed based on time series gene expression data, generated using microarray and sequencing technology. Among these models, *Boolean network* [1] based inference methods are promising, because they are simple, and have the ability to efficiently represent both the topology and the functions of networks.

However, the reconstruction of regulatory interactions of GRNs that are both accurate and efficient remains a problem. In order to find a solution, many researchers focus on the application of *feature selection* (FS) methods for regulatory gene selection, primarily employing machine learning (ML). These methods can be divided in to two main categories, filter and wrapper, and combining these two categories, embedded, hybrid, and *ensemble feature selection* methods can be implemented [2]. While wrapper methods are based on learning algorithms, filter methods are based on measurement techniques such as mutual information, correlation, distance and consistency measurement, and they tend to be faster in execution. Among the numerous FS methods, the *Min-Redundancy Max-Relevance* (MRMR) [3] criterion is a commonly used filter feature selection method based on mutual information. *ReliefF* [4], another efficient instance-based filter method, assesses the quality of the features.

In the latest development in feature selection, ensemble models have attracted considerable attention, because they improve performance in classification and regression problems in ML by combining diverse FS methods, rather than by using a single model. GINIE3 [5], and BTNET [6] are well-known tree-based ensemble FS models. ARACNE [7], is another approach, which uses mutual information for gene ranking. As a further enhancement in gene inference approaches, FS models are used along with probabilistic modeling approaches such as Boolean or Bayesian networks. For instance, MIBNI [6] is a Boolean network inference method based on an efficient univariate filter technique, mutual information. However, these GRN inference methods, along with feature selection techniques, have limitations when attempting to obtain both efficient and accurate inference of network structure and network dynamics simultaneously within one computational method.

Therefore, in the proposed approach we implemented MRMR and ReliefF based ensemble filter FS approach together, with a new Append function for Boolean network inference, which is novel in its application to GRN topology inference. To infer associated Boolean functions, a Pearson correlation coefficient (PCC) based Boolean modelling approach which was introduced in our previous research [8] has been used.

II. Methods

An overall view of the proposed system is shown in Fig. 1. In this approach we introduce a MRMR and ReliefF based combined model for the selection of the best set of regulatory genes for a target gene in a Boolean network inference problem. The approach has three main steps. Before applying the inference strategies, time series gene expression data is converted to binary format with the use of the *k-means* discretization method.

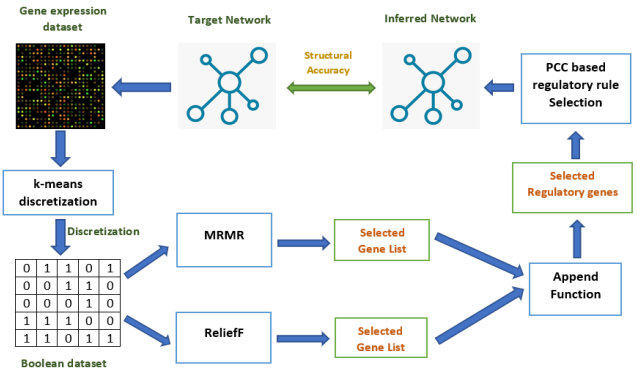

Fig. 1. *Overall view of the proposed system. This Boolean network inference method has 3 basic components: MRMR, ReliefF and Append function.*

A. Min-redundancy max-relavance criterion

In the first step, to select the most informative variables, $|S|$, of a target gene, h, from the set of all candidate Boolean variables, E, the multivariate mutual information between these genes and the target gene, is estimated using the MRMR. Using this criterion, the pairwise MI between each selected gene and the target gene, $I(h;x_j)$, $x_j \in (E - S)$, is calculated and used to identify the relevance that should be maximized. Then, based on this calculated MI score, irrelevant features are eliminated using a user-defined threshold. Then the MI between the selected gene and the rest of the other genes, $I(x_j;x_i)$ are calculated separately. This metric represents the redundancy between this candidate feature, x_j and the rest of the already selected features, $x_i \in S$, and it should be minimized. The formulation for the selection of the m-th feature is:

$$\max_{x_j \in (E-S)} \left[I(h\,;\, x_j) \; - \frac{1}{m-1} \sum_{x_i \in S} I(x_j\,;\, x_i) \right] \qquad (1)$$

An evaluation method-based stopping criterion is used to find the best set of regulatory genes for each target gene, and the regulatory gene search process is stopped at the point where no further improvement in gene-wise precision and dynamic accuracy is observed. The selected number of genes are considered as the maximum indegree (k) of the target gene.

B. ReliefF method

As the second step, a further optimal set, k, of genes is selected based on ReliefF [4], a simple and efficient FS model. ReliefF is an estimator of the quality of genes, based on their ability to differentiate between instances (time series gene expression data samples) that are near to each other. For a randomly selected instance R, ReliefF searches the N nearest neighbours from the same class (0 or 1) of R (nearest hits H) and N nearest neighbours from each of the different classes (nearest misses M), found in terms of the Euclidean distance between the two samples. Then, the quality estimation weight value, Wx, for gene x is updated according to Equation (2). If instance R and those in H have different values for gene x, then the weight, Wx is decreased. However, if instance R and those in M have different values for the gene x, then Wx is increased.

$$W_x = W_x - \frac{\sum_{n=1}^{N} DI_H}{l.N} + \sum_{c=1}^{C-1} P_c \cdot \frac{\sum_{n=1}^{N} DI_{M_c}}{l.N} \qquad (2)$$

In Equation (2), l is the number of instances (time samples) in class c (target gene). DI_H (or DI_{Mc}) is the sum of distances between the selected instance and each H (or M_c). P_c is the prior probability of possible values (0 or 1) of target gene.

C. Append function

In the third step, the Append function is used for the selection of common genes from the output gene subsets (k number of regulatory genes in each set) of the MRMR and ReliefF, and the selection of other suitable genes from the rest of the genes which belong to only one subset, and appends them to the selected list based on the performance evaluation criteria of gene-wise precision and dynamic accuracy.

III. Experiments

Evaluations of the proposed method are ongoing. The experiments are based on artificial and real gene expression datasets, adopted from the existing literature [6] and the performance of the proposed method is compared with the popular FS based GRN inference methods (TABLE I), GENEI3, ARCANE, and BTNET, which were implemented using learning algorithms, and require high computational resources. The proposed approach is a filter-based method, limited to only three iterations (highest maximum indegree of a target gene, $k = 3$) for the selection of the best regulatory gene set for a target gene. This method provides a significant improvement in computational efficiency. Comparison of the experimental results of existing methods [6] and the proposed approach (TABLE I) with five real gene networks indicates that the proposed method outperformed existing methods by obtaining high structural accuracy.

TABLE I. Structural Accuraccies Of Real Gene Networks

Method	CDC-15	CDC-28	Silico-1	Silico-2	S.cerevisiae
ARACNE	0.66	0.55	0.67	0.60	0.68
GENIE3	0.61	0.54	0.55	0.62	0.67
BTNET	0.59	0.57	0.68	0.76	0.69
Proposed Method	**0.80**	**0.78**	**0.83**	**0.85**	**0.81**

IV. Conclusions

In this study, we proposed an ensemble FS method-based efficient Boolean modelling approach for GRN inference. Among the various FS methods, MRMR and ReliefF were selected for this proposed ensemble approach, since they are filter methods, require low computational time, and lead to improved efficiency. With the use of an Append function, the selected subsets of genes from MRMR and ReliefF are further filtered and appended, to identify an optimal set of genes with improved structural accuracy. Finally, a Pearson correlation coefficient-based Boolean network regulation modelling approach is used for efficient regulatory rule selection. Our approach outperformed existing methods in terms of efficiency and structural accuracy, and further experiments are in progress using artificial and other real gene networks.

References

[1] S.A.Kauffman, "Metabolic stability and epigenesis in randomly constructed genetic nets," *Theoretical Biology,* vol. 22, no. 3, pp. 437-467, 1969.

[2] J. C. Ang, A. Mirzal, H. Haron and H. N. A. Hamed, "Supervised, Unsupervised, and Semi-Supervised Feature Selection: A Review on Gene Selection," *IEEE TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS,* vol. 13, no. 5, 2015.

[3] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *Bioinformatics and Computational Biology,* vol. 3, no. 2, pp. 185-205, 2005.

[4] M. Robnik-Šikonja and I. Kononenko, "Theoretical and Empirical Analysis of ReliefF and RReliefF," *Machine Learning,* vol. 53, p. 23–69, 2003.

[5] V. A. Huynh-Thu, A. Irrthum, L. Wehenkel and P. Geurts, "Inferring Regulatory Networks from Expression Data Using Tree-Based Methods," *PLoS ONE,* vol. 5, no. 9, 2010.

[6] S. Barman and Y.-K. Kwon, "A neuro-evolution approach to infer a Boolean network from time-series gene expressions," *Bioinformatics,* vol. 36, no. 26, pp. 762-769, 2020.

[7] A. A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. D. Favera and A. Califano, "ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context," *BMC Bioinformatics,* vol. 7, no. S7, 2006.

[8] H. N. Gamage, M. Chetty, A. Shatte and J. Hallinan, "An Efficient Boolean Modelling Approach for Genetic Network Inference," Submitted in *cibcb2021*, Melbourne, 2021.