ATLANTIS PRESS
Advances in Intelligent Systems Research (AISR), volume 141
International Conference on Applied Mathematics, Modelling and Statistics Application (AMMSA 2017)

Identification of Disease-Associated Combination of SNPs Using a Hybrid Algorithm with Multiple Encoding Approaches

Jing Zhao[1] and Bin Wei[2,*]
[1]College of Science Xijing University, Xi'an, China
[2]Key Laboratory of Network & Information Security of APF, Engineering College of APF, Xi'an, China
[*]Corresponding author

***Abstract*—Background: Individual SNP often only exhibit a small effect, but combinations of SNPs are assumed to be strongly influence the risk of disease. Obviously, selecting an optimal subset of SNPs, which most associated with disease, is a NP-hard problem. Results: To obtain a higher performance of predicting power for disease status and a higher computing efficiency, we proposed a double-filter-wrapper (DFW) algorithm to identify the optimal subset of SNPs. Moreover, few studies have been carried out to solve the SNPs encoding issues. On the basis of the differences of statistical properties between case and control, three types of encoding methods were proposed to generate the input for the DFW. Conclusion: We used five complex disease datasets to verify the effectiveness of our algorithm. The experimental results showed that our method appears more promising than other current methods for identifying the associated SNPs. In addition, the results also indicate that the encoding method proposed in this paper can much more accurately reflect the real situation.**

Keywords-GWAS; SNPs; feature selection; UMDA; SVM

I. INTRODUCTION

Due to the rapid development of genotyping technology, genome-wide association studies (GWAS) have been increasingly used to decipher DNA variations that are responsible for complex diseases[1]. Common DNA variations, in the form of single nucleotide polymorphisms (SNPs), hold much promise as a basis of disease association studies.

In the past few decades, thousands of SNPs were genotyped, and studies showed that only a portion of SNPs are strongly indicative of a targeted disease[2]. Thus, it is vital to select an optimal subset of SNPs that are more influential than the others, thereby allowing researchers to focus on the most promising SNPs for diagnostics and therapeutics.

Actually, finding association between SNPs and a disease can be viewed as a feature selection (FS) problem, that is, use a relevance criterion that decides whether a set of SNPs is significantly representative of the disease. Some algorithms have been proposed to identify disease associated SNPs using FS algorithm[3]. These algorithms can be split in two basic aspects: filter and wrapper. The advantages of filter methods are that they can easily scale up to high-dimensional datasets and computationally fast[4]. On the other hand, the wrapper methods have better classification performance than filter ones. Whereas, the major disadvantage of the wrapper methods is that the computation requirement is huge[5]. Combining the two algorithms seems to be a feasible way to overcome the disadvantages of the both of them. However, different filters may be yield different subsets that may leave out some potential relevant SNPs in the filter stage, which consequently do not have the chance to be considered in the wrapper evaluation. Therefore, this paper proposes a double-filter-wrapper (DFW) algorithm, in which two filter algorithms are used in the pre-select stage. Two filters ensure that the potential disease associated SNPs have a low probability to be filtered out in the initial stage. Recently, many wrapper algorithms have been proposed, which can be divided into two classes: the sequential search (SS) and the evolutionary algorithm (EA)[6]. Kudo and Sklansky found that the SS is suitable for the small and medium-sized problems, while the EA is better at large-sized problems. Thus, univariate marginal distribution algorithm (UMDA) is used in this paper and the support vector machine (SVM) is adopted as the evaluator. However, in UMDA, lack of diversity is the dominant factor converging to local optimum solutions[7]. Therefore, a novel dynamic elite selection strategy is proposed to overcome this problem.

In addition, in the most of previous studies, the encoding of SNP was usually limited to three types (0 and 1 stand for homozygous sites with major and minor allele, respectively, and 2 stands for heterozygous sites). The number of types is too small to express the abundant information. However, few of studies were devoted to solving the issues of SNP encoding. In this paper, three types of encoding methods, mono-SNP (MS) encoding, MS with frequency difference between the cases and controls (FDCC) encoding and MS_FDCC with distribution information, are used to generate the input for the DFW. The proposed method was tested on five complex disease datasets, and the results showed that our algorithm can identify the most disease associated SNPs.

II. METHOD

A two steps feature selection method is proposed in this paper. In the first step, two filter methods are used to remove irrelevant SNPs and reduce computational complexity. Then,

in the second step, a wrapper method is used to search the most associated SNPs.

A. Filter Algorithm

Two of the most popular filter methods F-score and Relief are used in this paper.

1) F-score: F-score is a simple filter algorithm which measures the distinction between two classes with real values [8]. Given training vectors $x_k, k = 1,...,n$, if the number of positive and negative instances is n_+ and n_-, respectively, then the F-score of the ith feature is defined as follows:

$$F(i) = \frac{(\bar{x}_i^{(+)} - \bar{x}_i)^2 + (\bar{x}_i^{(-)} - \bar{x}_i)^2}{\frac{1}{n_+ - 1}\sum_{k=1}^{n_+}(x_{k,i}^{(+)} - \bar{x}_i^{(+)})^2 + \frac{1}{n_- - 1}\sum_{k=1}^{n_-}(x_{k,i}^{(-)} - \bar{x}_i^{(-)})^2} \tag{1}$$

where $\bar{x}_i$, $\bar{x}_i^{(+)}$, $\bar{x}_i^{(-)}$ are the average value of the *ith* feature for the whole, positive, and negative data sets, respectively; $x_{k,i}^{(+)}$ is the *ith* feature of the *kth* positive instance, and $x_{k,i}^{(-)}$ is the *ith* feature of the *kth* negative instance.

2) *Relief:* Relief is one of the most successful methods among the existing feature weighting algorithms. Its main advantage over others is that it takes into account the effect of interacting among attributes [9]. Given a sample x, the Relief searches for its two nearest neighbors including one from the same class, called nearest hit NH(x), and the other from the opposite class, called nearest miss NM(x), and the Euclidean distance measure is defined as follows:

$$d_{rs} = \|x_r - x_s\| = \sqrt{\|x_r\|^2 + \|x_s\|^2 - 2x_r x_s} \tag{2}$$

where x_r and x_s are two vectors representing samples *r* and *s* and d_{rs} is Euclidean distance between them.

The weight of *ith* feature W_i is updated according to the following equation:

$$W_i = W_i + \left|x^{(i)} - NM^{(i)}(x)\right| - \left|x^{(i)} - NH^{(i)}(x)\right|, i = 1,2,...,I \tag{3}$$

where $x^{(i)}$, $NM^{(i)}(x)$ and $NH^{(i)}(x)$ represent the *ith* feature of the selected sample *x*, *NM(x)* and *NH(x)*, respectively; *I* is the number of candidate input features.

B. Wrapper Algorithm

In UMDA, lack of diversity, particularly during the later stages of evolution, is the dominant factor converging to local optimum. Therefore, we propose a novel UMDA (NUMDA), in which the number of elite individuals selected at each iteration is not static but dynamical, to overcome the drawback that standard UMDA presents.

The main steps of NUMDA are given as follows: Firstly, *M* individuals are generated randomly. Secondly, individuals are ranked according to the fitness values from high to low, and the best N_t *(* N_t *<M)* of them are selected as the elites.

$$N_t = \left\lceil (N_{\max} - N_{\min}) \times \frac{iter_{\max} - iter}{iter_{\max}} + N_{\min} \right\rceil \tag{4}$$

where *iter* is the current iterations; $iter_{\max}$ is the maximum number of iterations; $N_{\max}$ and $N_{\min}$ represent the upper bound and the lower bound of the number of elites. The dynamical mechanism ensures that the algorithm not only has a fast convergence speed in the earlier iterations, but also keeps diversity in the later stages.

Then, the selected individuals are used to estimate the probability distribution $p_t(x)$. Afterwards, the $p_t(x)$ is used to generate the next population. The probability distribution of the *tth* iteration is defined as follows:

$$p_t(x) = \prod_{i=1}^{n} p\left(x_i \mid pop_{t-1}^{Sel}\right) = \prod_{i=1}^{n} \frac{\sum_{j=1}^{N} \delta_j (X_i = x_i \mid pop_{t-1}^{Sel})}{N_t} \tag{5}$$

where pop_{t-1}^{Sel} is the selected elites at *(t-1)th* iteration; $\delta_j\left(X_i = x_i | pop_{l-1}^{Sel}\right) = \begin{cases} 1 & X_i = x_i \\ 0 & others \end{cases}$.

C. Double-Filter-Wrapper Algorithm

F-score and Relief have their own characteristics, and there is a proportion of overlap among the lists of removed SNPs. In this study, the actual removed SNPs in the filter stage are formed by taking the intersection of the above two lists. The SNP(s) in the remove-list of F-score or Relief but not in the intersection set is (are) given a lower probability in the corresponding bit of NUMDA (see Fig.1).

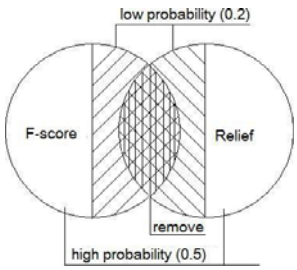

FIGURE I. SCHEMATIC DIAGRAM OF HYBRID FILTER ALGORITHM

The NUMDA-SVM contains two steps. The first step entails selection of a set of SNPs and the parameters of SVM by NUMDA. Then, in the second step, the selected SNPs and parameters are passed to the SVM to acquire a fitness value for each individual.

An individual of NUMDA comprises two parts: the feature mask and the SVM parameters, which is shown as follows.

$$\{x_1, x_2, ..., x_n, x_{n+1}, x_{n+2}, ..., x_j\} \tag{6}$$

In Eq.6, $x_1, x_2, ..., x_n$ correspond to the *n* SNPs (the bit with value '1' indicates the SNP is selected and '0' indicates not). $x_{n+1}, ..., x_j$ are used to encode the γ and penalty parameter *C*.

Based on the above analysis, the procedure of DFW is given as follows:

FILTER

1) Calculate the score of each SNP by F-score and Relief, respectively.

2) Generate D_R and D_F, where D_R and D_F are the list of pre-selection SNPs obtained by Relief and F-score, respectively.

3) Take the intersection of $D_{\overline{R}}$ and $D_{\overline{F}}$. $D_{\overline{FR}} = D_{\overline{F}} \text{ I } D_{\overline{R}}$, $D_{\overline{R}} = D - D_R$, $D_{\overline{F}} = D - D_F$, where D is the original dataset.

4) Remove $D_{\overline{FR}}$ from D, $D_p = D - D_{\overline{FR}}$.

WRAPPER

1) Initialize probability vector $P = [x_1, x_2, ...]$, $x_i \in \{0.2, 0.5\}$. 0.2 for the SNP(s) in $D_{\overline{F}} - D_{\overline{FR}}$ or $D_{\overline{R}} - D_{\overline{FR}}$ and 0.5 for the others (see Fig.1).

2) Generate *M* individuals by sampling from P.

3) Decode each individual to get the parameters of SVM and the SNPs subset.

4) Calculate the fitness value (*Acc*) of each individual.

5) Sort the individuals according to their fitness values from high to low.

6) Select N_t individuals with higher fitness value from population.

7) Estimate the probability distribution by (5) and generate the next population.

8) If the termination criterion is satisfied, stop. Otherwise, go to Step 3. The maximum number of iterations is the termination criterion.

III. ENCODING METHODS

A. MS Encoding

Each of SNP is given an integer number (0, 1, 2). 0 and 1 stand for homozygous sites with major and minor allele, respectively, and 2 stands for heterozygous sites.

B. MS_FDCC Encoding

Most previous studies were based on the differences of statistical properties between case and control. The results suggested there is a prominent difference between case and control. The FDCC method captures this character very well. A position weight matrix is derived from the case set by tabulating the frequency of each genotype occurs at each position.

$$M_{ij} = \frac{1}{n}\sum_{k=1}^{n} O_i(S_{kj}), i = 0,1,2;\ j = 1,2,\cdots,l \tag{7}$$

where $O_i(x) = \begin{cases} 1, i = x \\ 0, otherwise \end{cases}$, *n* is the number of samples in the case set, *l* is the number of SNPs in each sample and $S_{kj} \in \{0,1,2\}$. In the same way, a position weight matrix can be obtained for the control set. A FDCC encoding matrix is obtained by subtracting the case coding matrix from the control one.

C. MS_FDCC_DI Encoding

However, FDCC method has its limit. For example, the frequency of genotype 0 is 0.8 in case set and 0.7 in control set. The frequency of genotype 1 is 0.2 in case set and 0.1 in control set. When the MS_FDCC encoding method is used, these two features have the same value. However, the contribution of 0 and 1 to the disease may be different. Therefore, the distribution information (DI) is added to (7).

$$M_{ij}^{*} = \log(p_j(S_{kj} \mid c_t)) \times \frac{1}{n}\sum_{k=1}^{n} O_i(S_{kj}) \tag{8}$$

where $p_j(S_{kj} \mid c_t)$ is the frequency of S_{kj} in the class c_t; $c_t \in \{case, control\}$.

IV. EXPERIMENTAL RESULTS

A. Datasets

In this paper, five datasets (Autoimmune disorder (AD), Crohn's disease (CD), Lung cancer (LC), Rheumatoid arthritis (RA) and Tick-borne encephalitis (TE)) are used to evaluate the effectiveness of our algorithm. All the datasets were supplied by Brinza [10].The characteristics are presented in Table 1.

TABLE I. THE DATASET DESCRIPTION.

Data set	Number of SNPs	Number of cases	Number of controls
AD	108	384	652
CD	103	144	243
LC	141	322	273
RA	2300	460	460
TE	41	21	54

B. Results

Firstly, the effects of the three encoding methods are analyzed. The proposed algorithm with different encoding methods was used to predict the disease statuses and the results are showed in Table2. From the table we can see that the MS obtained the worst Acc for all of the five datasets. On the other hand, the MS_FDCC_DI reached the best results comparing

with other two encoding methods. Therefore, it can be concluded that the MS_FDCC_DI encoding method can more accurately reflect the difference between cases and controls.

TABLE II. THE PERFORMANCE OF DFW_SVM WITH DIFFERENT ENCODING METHODS.

Dataset	Encoding	Sn	Sp	Acc
AD	MS	0.6642	0.7462	0.7159
	MS_FDCC	0.6854	0.7926	0.7530
	MS_FDCC_DI	0.5449	0.9304	0.7881
CD	MS	0.8273	0.9835	0.9252
	MS_FDCC	0.8961	0.9717	0.9509
	MS_FDCC_DI	0.8963	0.9876	0.9535
LC	MS	0.8818	0.8240	0.8553
	MS_FDCC	0.9163	0.8826	0.9009
	MS_FDCC_DI	0.9317	0.8865	0.9110
RA	MS	0.7217	0.7583	0.7400
	MS_FDCC	0.7391	0.7648	0.7519
	MS_FDCC_DI	0.8022	0.8149	0.8085
TE	MS	0.7600	1.0000	0.9313
	MS_FDCC	0.9500	1.0000	0.9857
	MS_FDCC_DI	1.0000	1.0000	1.0000

Secondly, for further comparison, we executed IBPSO [11], PA [12] and HPG [13] on the five datasets. Table 3 shows the Sn, Sp and Acc obtained by these algorithms. From the table, we can see that the best classification result on the TE dataset was 1.00 using the DFW_SVM, whereas, only 0.9465, 0.9181 and 0.9599 were obtained by the three comparison methods. For the AD, CD, LC and RA datasets, the classification accuracy obtained by our method were also better than that of other algorithms. This means that the proposed algorithm has a higher ability to select the feature subsets to discriminate the cases and controls.

TABLE III. THE PERFORMANCE OF OUR ALGORITHM COMPARED WITH SOME PUBLISHED METHODS.

Dataset	Method	Sn	Sp	Acc
AD	IBPSO	0.3703	0.9180	0.7158
	PA	0.7039	0.7756	0.7491
	HPG	0.6853	0.7911	0.7520
	DFW	0.5449	0.9304	0.7881
CD	IBPSO	0.8478	0.9588	0.9174
	PA	0.8682	0.9713	0.9328
	HPG	0.8611	0.9548	0.9198
	DFW	0.8963	0.9876	0.9535
LC	IBPSO	0.8977	0.8426	0.8724
	PA	0.9068	0.8570	0.8840
	HPG	0.9192	0.8607	0.8924
	DFW	0.9317	0.8865	0.9110
RA	IBPSO	0.7848	0.7910	0.7878
	PA	0.7761	0.8127	0.7944
	HPG	0.7913	0.8018	0.7965
	DFW	0.8022	0.8149	0.8085
TE	IBPSO	0.9000	0.9636	0.9465
	PA	0.9000	0.9236	0.9181
	HPG	0.9600	0.9618	0.9599
	DFW	1.0000	1.0000	1.0000

Figure 2 shows the number of iterations of NUMDA versus Acc and the number of SNPs selected. For most of datasets, the numbers of SNPs selected converged at later stages; however, the classification accuracy kept improving. This can be explained by two facts: 1) the actual selected SNPs can be different even if the total number is the same; 2) optimizing the SVM's parameters can increase classification accuracy. Therefore, the feature subset and model parameters must be determined simultaneously.

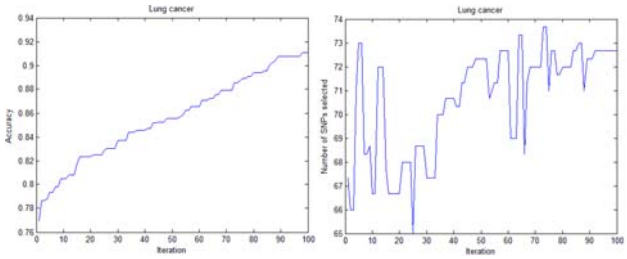

FIGURE II. THE NUMBER OF ITERATIONS VERSUS PREDICTION ACCURACY AND THE NUMBER OF SNPS SELECTED.

V. CONCLUSIONS

The SNP with small individual effect but jointly significant effects would be missed by single SNP analysis. In this paper, the DFW has been proposed to identify the subset of SNPs which are most associated with diseases. Two filters are used in DFW with the objective of including as many relevant SNPs as possible, so as not to leave out any potential relevant SNPs at the filter stage. In the wrapper stage, NUMDA with SVM is used to optimize the performance of selected feature subset. In addition, three different encoding methods have been applied and have shown the effectiveness of the DFW was benefited from the used of the MS_FDCC_DI. The experimental results on five disease datasets indicate that the MS_FDCC_DI can much more accurately reflect the real situation. Based on the results, it seems that the algorithm proposed in this paper is more promising to be used in the genome-wide association studies.

ACKNOWLEDGMENT

This study was supported by the Scientific research program funded by Shaanxi provincial education department (program NO. 15JK2187), Scientific research program funded by Xijing University (program NO. XJ160235), National social science foundation (program NO. 16BTJ033), Foundation of engineering college of APF (WJY201518).

REFERENCES

[1] An, L.A., Ehsan; Liu, Mingxia, A Hierarchical Feature and Sample Selection Framework and Its Application for Alzheimer's Disease Diagnosis. SCIENTIFIC REPORTS, 2017. 7 p. 45269

[2] Bin, W., A hybird algorithm for disease association stduy. Journal of biomedical science and engineering, 2016. 9(10): p. 129-136.

[3] Pes, B.D., Nicoletta; Angioni, Marta, Exploiting the ensemble paradigm for stable feature selection: A case study on high-dimensional genomic data. INFORMATION FUSION, 2017. 35 p. 132-147.

[4] Martinez-Gonzalez, B.P., Jose M.; Echeverry-Correa, Julian D, Spatial features selection for unsupervised speaker segmentation and clustering. EXPERT SYSTEMS WITH APPLICATIONS, 2017. 73(1): p. 27-42.

[5] 5. Li, Y.Y., Yuantao; Li, Guoyan, A fault diagnosis scheme for planetary gearboxes using modified multi-scale symbolic dynamic entropy and mRMR feature selection. MECHANICAL SYSTEMS AND SIGNAL PROCESSING 2017(91): p. 295-312.

[6] Gao, S.d.S., Clarence W, A modified estimation distribution algorithm based on extreme elitism. BIOSYSTEMS 2016. 150 p. 149-166.

[7] Alba, E.M., J.; Dorronsoro, B, Theory and practice of cellular UMDA for discrete optimization. PARALLEL PROBLEM SOLVING FROM NATURE - PPSN IX, PROCEEDINGS 2006. 4193 p. 242-251.

[8] Cheng-Lung, H., C. Mu-Chen, and W. Chieh-Jen, Credit scoring with a data mining approach based on support vector machines. Expert Systems with Applications, 2007. 33(4): p. 847-56.

[9] CARUANA R., F.D., Greedy attribute selection. Int. Conf. Machine Learning,, 1994 p. 28–36.

[10] Brinza, D. and A. Zelikovsky, Design and validation of methods searching for risk factors in genotype case-control studies. Journal of Computational Biology, 2008. 15(1): p. 81-90.

[11] Li-Yeh, C., et al., Improved binary PSO for feature selection using gene expression data. Computational Biology and Chemistry, 2008: p. 29-37.

[12] Peng, Y., Z. Wu, and J. Jiang, A novel feature selection approach for biomedical data classification. Journal of Biomedical Informatics, 2010. 43(1): p. 15-23.

[13] Shutao, L., W. Xixian, and T. Mingkui, Gene selection using hybrid particle swarm optimization and genetic algorithm. Soft Computing, 2008: p. 1039-48.