This is the published version (version of record) of:


Available from Deakin Research Online:

http://hdl.handle.net/10536/DRO/DU:30018221

©2008 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.

Copyright : 2008, IEEE
A Hybrid Approach to Selecting Susceptible Single Nucleotide Polymorphisms for Complex Disease Analysis

Pengyi Yang1 and Zili Zhang1,2
1Intelligent Software and Software Engineering Laboratory
Faculty of Computer and Information Science
Southwest University, Chongqing 400715, China

2School of Engineering and Information Technology
Deakin University, Geelong, VIC 3217, Australia
zzhang@deakin.edu.au

Abstract

An increasingly popular and promising way for complex disease diagnosis is to employ artificial neural networks (ANN). Single Nucleotide Polymorphisms (SNP) data from individuals is used as the inputs of ANN to find out specific SNP patterns related to certain disease. Due to the large number of SNPs, it is crucial to select optimal SNP subset and their combinations so that the inputs of ANN can be reduced. With this observation in mind, a hybrid approach – a combination of genetic algorithms (GA) and ANN (called GANN) is used to automatically determine optimal SNP set and optimize the structure of ANN. The proposed GANN algorithm is evaluated by using both a synthetic dataset and a real SNP dataset of a complex disease.

1. Introduction

Complex diseases are generally influenced by complex interactions among multiple Single Nucleotide Polymorphisms (SNPs) residing in various genes. The advances of genome-wide association study facilitate the identification of a large number of SNPs that contributes to the changes of gene function and interactions among them. Therefore, analyzing SNP data can help us to find genes and SNP patterns that may be responsible for the development of certain complex disease [1].

To map the SNP patterns of a complex disease, strategies like retrospective likelihood method [2], Artificial Neural Network (ANN) method [3, 4], and statistical hybrid approach [5] have been explored. Among these strategies, ANN seems promising in disease diagnosis as it is suitable for the analysis of complex nonlinear patterns, which is the most important nature of SNP data [6]. If the SNP profile of certain disease is successfully obtained by ANN, the susceptibility of new individuals to this disease can be diagnosed by simply presenting the disease related SNP genotypes of those people as the inputs of the ANN. Although ANNs have been successfully applied to SNP data analysis, there are still some problems that remain to be addressed.

When solving problems using ANNs, the quality of solutions heavily rely on the quality of data to be processed. The less noise and redundancy in the data, the better the results. However, like many other biological data, SNP data are extremely noisy and redundant. Thus, it is evident that we need efficient ways to optimize ANN’s structure and remove redundant and noisy data in order to obtain the optimal SNP profile for complex disease analysis. With such an optimal SNP profile as input, the ANN can yield high classification accuracy. In this work, we address the SNP subset selection and ANN structure optimization problems in complex disease analysis through a hybrid approach.

The rest of the paper is structured as follows: Section 2 discusses the hybrid approach, which is a combination of genetic algorithms and artificial neural networks. Section 3 is the evaluation of the proposed hybrid approach using simulation data. The evaluation of the proposed algorithm using real data is provided in Section 4. Finally, Section 5 concludes the paper.

2. The Genetic Algorithm and Artificial Neural Network Hybrid Approach

To obtain a small yet optimized SNP subset for complex disease analysis, we need to identify those SNPs that can produce high data classification accuracy when combined. Of course, if we can test all the possible combinations of
SNPs, it is easy to find the best SNP combinations that can produce the highest data classification accuracy. Unfortunately, it is infeasible to do an exhaustive check of all possible SNP combinations with a relatively large SNP set. One effective way to select SNPs that meet our requirement (small yet optimal) is the hybrid approach – the combination of GA and ANN. GA has excellent ability to solve problems with large search domains and can be used to select the best combinations of the input attributes of SNPs in our case.

Genetic and neural network hybrid approach is not new. Actually it has been applied in different applications [7, 8]. Yang and Honavar provide a comprehensive evaluation of such a hybrid approach [9]. The hybrid algorithm (GANN) we employ in this study is a variation of neural-genetic approach proposed by Keedwell and Narayanan [7]. In Keedwell and Narayanan’s hybrid approach, the neural network can only take Boolean inputs. This has been extended to accept continuous data in our GANN approach. Moreover, the activation function used in Keedwell and Narayanan’s approach is a single layer step-function, which can not capture nonlinear pattern information such as gene-gene interaction in SNP data [6]. In the GANN, a hidden layer was employed to map the nonlinear structure of SNP data.

When applying GAs to solve problems, we need to encode the chromosome strings of the problems in a certain way. The encoding scheme we used is to mark every input SNP with an id, and the GA chromosome is encoded in a string of candidate SNP ids.

The fitness function can be defined as the evaluation accuracy of the ANN over certain combination of input attributes. The process of training, testing and selecting can be described as follows:

- First, the algorithm starts with a set of randomly initialized chromosomes.
- The SNPs contained in a chromosome are used as the inputs of the ANN. After the ANN tests a kind of combination, it returns the evaluation accuracy of this combination to GA.
- After the whole population is tested, GA selects the chromosomes by checking their evaluation accuracy.
- The favorite chromosomes are crossed over and mutated with certain crossover probability $p_c$, mutation probability $p_m$, and the next generation begins.
- Repeat the above steps until certain generation is reached. The final chromosomes are then output as the most suitable SNP combinations that determine the underlying patterns of the SNP data.

According to the above description, we can have the structure of the GANN model, which is shown in Figure 1.

3. Data Simulation and Analysis

3.1. Experiment Design

It is advisable to evaluate the usefulness of this GANN hybrid model before applying to real SNP data analysis. Hence, an artificially generated data pattern was used in this evaluation experiment. It seems arbitrary, but it does demonstrate the feasibility of the proposed GANN model. It also shows how the GANN works in a concise way.

Suppose the underlying patterns of the data are as follows (in regular expression form):

- if $input:0^+1^+$ then $output:0$
- if $input:1^+0^+$ then $output:1$

Table 1 shows the data used to train and test the GANN on learning the pattern described above. Apparently, the inputs with $id_3$ and $id_6$ are noisy data. In this case, if all data are fed into the ANN in the training process, the pattern may still be obtained, but with higher overall classification error compared with noise-free data. In real applications, the patterns of the data may be far more complex than the patterns described here. When noisy data are included in the training and testing process, the prediction ability of the ANN on new unseen data may decrease significantly. Our purpose is to use GA to optimize the ANN model. In this example, GA is used to filter out the noisy attributes ($id_3$ and $id_6$).

3.2. Evaluation

The ANN used in this study is a three layer standard back-propagation neural network. The learning
Table 1. The training and test data.

<table>
<thead>
<tr>
<th>patterns</th>
<th>id of candidate inputs</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>Test</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>1</td>
</tr>
</tbody>
</table>

strategy proposed by Brierley and Batty [10] was implemented. The single point crossover and mutation as well as the binary tournament selection [11] were adopted for the GA. Parameters setting of ANN are: learning rate = 0.3, epoch = 100 and hidden node = input node/2, and parameter setting of GA are: population size = 30, possibility of crossover = 0.7, possibility of mutation = 0.01 and termination condition = 20 generation.

The algorithm starts with 30 randomly initialized chromosomes with length of six (since only six inputs are pattern related attributes in this case). In real world, the number of pattern associated attributes is unknown and we need to try different number of inputs several times.

The overall error of ANN is estimated by RMS (root mean squared). The RMS errors in classify noise-free data and data with noise in ANN evaluation process are shown in Figure 2. The performance graph of GA (Figure 3) shows the average fitness scores of chromosomes from 1 to 11 generations.

As shown in Figure 3, GANN reached the 100% accuracy of data classification and is stable after 11 generations. At this stage, the candidate inputs with attributes id3 and id6 are filtered out by this algorithm.

This experiment illustrates that GANN can identify optimal dataset and optimize ANN model. The noise-free data were obtained by selecting the “right” attributes.

After demonstrating the feasibility of GANN in attributes selection, we are now ready to apply this algorithm to SNP selection and complex disease analysis.

4. SNP Data Selection and Complex Disease Analysis

4.1. SNP Data

The SNP dataset published by Klein et al. [12] was used in this study. It includes a genome-wide screen of 96 cases and 50 controls for polymorphisms associated with Age-Related Macular Degeneration (AMD). Within these 96 cases, 50 are unioocular choroidal neovascularization (Neov) and 46 are geographic atrophy either central or non-central to the macula (CGA).

AMD is the major cause of uncorrectable blindness of people in many countries. Like many other complex diseases, AMD is influenced by complex interactions among multiple genes and environment factors [12]. Therefore, the proposed GANN approach is well suited to such SNP dataset analysis.

4.2. Candidate SNPs Select and Data Encode

Recently, many papers discussed the association of many important SNPs with AMD development [12, 13, 14]. Within these AMD related SNPs, 10 SNPs claimed to have strong association with AMD development are selected in
this experiment. We call these SNPs candidate SNPs, which are listed in Table 2.

The SNP data need to be encoded into numeric format before being fed into the ANN. The encoding methods used in this study are consistent with those described in [3] with some minor changes. Input and output variables are normalized. Three genotypes of the SNP data are encoded as AA: 0, AB: 0.5 and BB: 1, respectively; and the diagnosis data are customized as 1 for cases, and 0 for controls.

4.3. Experiment Design

In SNP data analysis, parameters setting of ANN are: learning rate = 0.3, epoch = 1000, input neuron = 2 to 10 and hidden neuron = input neuron/2, and parameters setting of GA are: population size = 100, possibility of crossover = 0.7, possibility of mutation = 0.03 and termination condition = 50 generation.

For comparison purpose, a wrapper strategy with forward selection scheme was utilized to extract SNP subset for SNP data classification.

Forward selection wrappers are one of the most popular attributes selection strategies and they are optimal at each stage. However, they are unable to capture complex interactions between attributes in SNPs. Furthermore, while these approaches work reasonably well with linear classifiers, they may exhibit poor performance with nonlinear classifiers like ANN [15].

With GANN and forward wrapper selection method, we started from 2-SNP combination and compared 2 to 9-SNP selected combinations in SNP data classification.

Two of the common evaluation strategies, multiple random validation and cross-validation, were utilized to evaluate the performance of GANN. In SNP selection process, we used multiple random validation to re-divide the dataset into training and test sets in every generation. By randomly dividing data every generation, the selected SNPs can be considered as the most stable ones in accurate classification. While in ANN evaluation phase, 5-fold cross-validation was applied. The final classification accuracy over evaluation dataset was calculated by conducting 5-fold cross-validation 5 times and averaging the results.

4.4. Results

Firstly, the 10 SNPs listed in Table 2 were used as the inputs of ANN to classify AMD dataset. The averaged 5-fold cross-validation accuracy for this 10-SNP combination is 69.60%. Then, GANN and forward wrapper were utilized to select SNP combinations from 2 to 9, respectively. The selected combinations (from 2 to 9) were evaluated by using 5-fold cross-validation averaging 5 times. The evaluation accuracy of GANN and forward wrapper selected combinations as well as 10-SNP were visualized in Figure 4. The results in Figure 4 demonstrate that in most cases GANN have superior power on SNP sets optimization compared with forward wrapper. Also, it’s readily notice that, combinations with 4 and 5 SNPs have relatively higher classification power. Among all combinations, a GANN selected 4-SNP combination with rs1061170 (Y402H), rs2019727 (I62V), rs1049029 (A-1707207 rs1329428) and A-1641951 rs3753396 achieves the highest classification accuracy with 73.98%. This result is 4.48% higher than that using 10 SNPs as ANN inputs. It indicates that GANN is able to identify a small and accurate SNP combination for AMD analysis.

Detailed evaluation accuracy of GANN and forward wrapper selected SNP combinations are provided in Ta-

<table>
<thead>
<tr>
<th>Location</th>
<th>Effect</th>
<th>SNP ID</th>
<th>rs number</th>
<th>P-value</th>
<th>GANN</th>
<th>Forward Wrapper</th>
</tr>
</thead>
<tbody>
<tr>
<td>CFH (exon 2)</td>
<td>I62V</td>
<td>C2530382-10</td>
<td>rs800292</td>
<td>2.8 × 10^{-4}</td>
<td>5 5</td>
<td></td>
</tr>
<tr>
<td>CFH (exon 9)</td>
<td>Y402H</td>
<td>Y402H</td>
<td>rs1061170</td>
<td>1.3 × 10^{-5}</td>
<td>3 6</td>
<td></td>
</tr>
<tr>
<td>CFH (intron 9)</td>
<td>A/T</td>
<td>SNP_A-1713303</td>
<td>rs2019727</td>
<td>1.2 × 10^{-4}</td>
<td>8 7</td>
<td></td>
</tr>
<tr>
<td>CFH (exon 13)</td>
<td>A/G</td>
<td>SNP_A-1641951</td>
<td>rs3753396</td>
<td>2.2 × 10^{-1}</td>
<td>7 8</td>
<td></td>
</tr>
<tr>
<td>CFH (intron 15)</td>
<td>C/G</td>
<td>SNP_A-1660027</td>
<td>rs380390</td>
<td>4.1 × 10^{-8}</td>
<td>6 2</td>
<td></td>
</tr>
<tr>
<td>CFH (intron 15)</td>
<td>A/G</td>
<td>SNP_A-1707207</td>
<td>rs1329428</td>
<td>1.4 × 10^{-6}</td>
<td>2 1</td>
<td></td>
</tr>
<tr>
<td>CFH (exon 8)</td>
<td>D936E</td>
<td>C_2530274</td>
<td>rs1065489</td>
<td>2.7 × 10^{-1}</td>
<td>4 4</td>
<td></td>
</tr>
<tr>
<td>PLEKHA1</td>
<td>---</td>
<td>SNP_10q26</td>
<td>---</td>
<td>1.8 × 10^{-3}</td>
<td>9 9</td>
<td></td>
</tr>
<tr>
<td>LOC387715</td>
<td>A69S</td>
<td>SNP_A-1700234</td>
<td>rs1049024</td>
<td>1.1 × 10^{-2}</td>
<td>1 3</td>
<td></td>
</tr>
<tr>
<td>LOC387715</td>
<td>C/T</td>
<td>SNP_A-1700146</td>
<td>rs2736911</td>
<td>2.9 × 10^{-1}</td>
<td>10 10</td>
<td></td>
</tr>
</tbody>
</table>
SNP Selection for AMD Classification

Figure 4. Evaluation accuracy of SNP combinations from 2 to 10.

Table 3, and the ranking order of SNPs in GANN and forward wrapper selected results are provided in the last two columns of Table 2. From the GANN ranking results, SNP rs10490924 (A69S) has been identified as the top factor associated with AMD development while rs1061170 (Y402H) takes the third place. This result is basically consistent with current knowledge of AMD development in literature [12, 13, 14].

Table 3. Evaluation results of SNP combinations selected by GANN and forward wrapper.

<table>
<thead>
<tr>
<th></th>
<th>GANN</th>
<th>Forward Wrapper</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-SNP</td>
<td>68.69</td>
<td>69.98</td>
</tr>
<tr>
<td>3-SNP</td>
<td>70.56</td>
<td>70.13</td>
</tr>
<tr>
<td>4-SNP</td>
<td>73.98</td>
<td>72.34</td>
</tr>
<tr>
<td>5-SNP</td>
<td>73.59</td>
<td>73.59</td>
</tr>
<tr>
<td>6-SNP</td>
<td>73.17</td>
<td>72.89</td>
</tr>
<tr>
<td>7-SNP</td>
<td>70.84</td>
<td>70.53</td>
</tr>
<tr>
<td>8-SNP</td>
<td>70.76</td>
<td>70.37</td>
</tr>
<tr>
<td>9-SNP</td>
<td>69.97</td>
<td>69.97</td>
</tr>
</tbody>
</table>

5. Conclusions

In this study, we proposed a GANN hybrid algorithm for SNP data selection and ANN model optimization. A synthetic dataset with noise was used to evaluate the usefulness of this hybrid approach first. It then was applied to the analysis of a real-world SNP dataset generated from AMD complex disease. The forward wrapper selection scheme was compared with the proposed GANN. The classification accuracy using 5-fold cross-validation indicates that GANN has higher power of SNP subset optimization than that of forward wrapper. A GANN selected 4-SNP combination achieved the highest data classification accuracy with 73.98%, and the ranking order of GANN indicates that SNP rs10490924 (A69S) and rs1061170 (Y402H) are the main factors of AMD development.

The experimental results suggest that the GANN be able to select critical disease related SNPs and create small SNP subset for the analysis of complex diseases. This is an important step as it can provide information and guidance for future experiments.

References