Finding Relevant SNP Sets and Predicting Disease Risk Using Simulated Annealing

DongHoi Kim, Saangyong Uhmn and Jin Kim*

Dept. of Computer Engineering, Hallym University
Chuncheon, Gangwondo 200-702 Republic of Korea
{kdh,suhmn,jinkim}@hallym.ac.kr

Abstract

We applied simulated annealing algorithm and decision tree to find set of single nucleotide polymorphisms relevant to a disease and build a risk prediction model. For time complexity problem of simulated annealing caused by initial set and candidate generation, we constructed an initial set of the variants by fast heuristic algorithm and proposed a transition rules based on contribution of available variants. The experiment results show that we can obtain new set of variants with the reduced number of variants and the improved prediction performance compared to others by traditional feature selection algorithms.

Keywords: *NP, risk prediction, simulated annealing, decision tree, feature selection, machine learning*

1. Introduction

After the completion of Human Genome Project in 2003, research efforts are devoted to find genetic variants relevant to specific diseases and their effects. As a part of them, many researches are being carried out for single nucleotide polymorphism (SNP) [1] and copy number variation (CNV) to identify the differences in human genomes between individuals or groups of individuals and their relation to diseases for disease prediction and personalized medicine. Several machine learning techniques such as support vector machine (SVM), decision tree[2-5], or neural net are applied for analysis. However, the advances in high throughput technologies for sequencing and genotyping have led to the accumulation of a huge amount of SNP data and the quantity is now approaching the limit of today's ordinary computers. As the number of SNPs increases, the time and space complexity increase so drastically that feature selection algorithm is required to select set of relevant SNPs such as forward selection, backward elimination[6, 7] and so on. However the optimum prediction accuracy is not attainable because those techniques easily fall into and hardly come out of local minima because of their greedy nature.

In this study, we applied simulated annealing (SA) [8] to select relevant SNPs and improve the performance of disease prediction accuracy. Coming from annealing in metallurgy involving heating and controlled cooling, simulated annealing may come out of local minima and find better solution. But it also has a weak point that the whole performance heavily depends on the initial state. To avoid this situation, we constructed initial state by forward selection and applied SA with efficient transition rules. The experiment results showed that our implementation of SA attained improved performance in disease prediction and new set of relevant SNPs compared to the

previous feature selection algorithms. We believed that our results may contribute to the field of research in life science.

This remainder of this paper consists as follows. In the following section, we give background information. In Section 3, we give a formal definition of the problem. In Section 4, we review some of existing methods for the problem. In Section 5, we present our speedup strategy and proposed transition rule. We describe data sets and give experimental results in Section 6. Finally, we conclude the paper in Section 7 and discuss future research direction.

2. Background

2.1 SNP

Human genome is a set of 23 chromosomes each of which is a sequence of four nucleotides, A, T, G, and C. A single nucleotide polymorphism is a variation of a nucleotide in those sequences between individuals or groups of individuals which results in difference in disease susceptibility or pharmacokinetics. Fig. 1 depicts a SNP in two sequences.

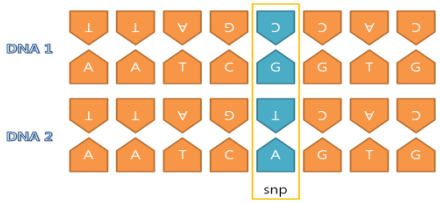

Figure. 1 SNP

We may build up the foundation of disease prediction or personalized medicine by analyzing and understanding SNPs. It may be easily observed through genetic experiments that single or small number of those variations in patients with genetic disease, called Mendelian disease. The other genetic diseases called complex diseases require more SNPs to consider for analysis as its name implies. In order to find several SNPs, we have to analyze several of them at the same time, not one by one, which means we have to analyze all possible subsets of SNPs to find an optimal set. The more SNPs it considers, the higher time and space complexity is which makes it impossible to evaluate all possible subset of SNPs. To overcome this problem, feature selection algorithm is applied prior to analysis to select smaller number of relevant SNPs. However, the more SNPs it considers, the more probable the method includes irrelevant SNPs which fall down the prediction accuracy. For this problem, we need a way to select subset of SNPs and change several of them based on the prediction performance.

2.2 Simulates Annealing (SA)

SA is a probabilistic approach that can be used to find a global minimum of a function in combinatorial optimization problems. To apply this algorithm to an optimization problem, state space $S=\{s_1, \ldots, s_n\}$ and a cost function $C:S \rightarrow R$, where R is the set of real number, should be defined. A real value $C(S)$ should be assigned to each state s. The goal of the optimization problem is to find the optimal state s_{opt} whose score is $MIN(MAX)$ $\{s_i | 1 \leq i \leq n\}$. SA continuously generates a new candidate state s_{new} from a current state $s_{current}$ by applying transition rules and acceptance rules[9].

The criteria of the acceptance rules are:

1) If $\Delta E \leq 0$, accept a new state s_{new}.

2) If $\Delta E > 0$, accept a new state s_{new} with probability $P(\Delta E) = e^{-\frac{\Delta E}{T}}$ where T is a temperature and $\Delta E = C(S_{new}) - C(S_{current})$ is a cost difference.

Probability $P(\Delta E)$ prevents the system from fixating at local minimum. A state $s_{current}$ is called local minimum if there is no new state s_{new} in S that is generated from the state $s_{current}$ by applying the transition rule and that has a lower cost than that of the $s_{current}$.

Temperature T controls a probability to accept a new state s_{new}. Initially, T starts from a high temperature and after each iteration, T decreases to become zero by applying an annealing schedule. The probability of accepting a new state with a higher cost than that of the current state also decreases as temperature T decreases. If a careful annealing schedule and number of iterations are given, SA converges to a global minimum state s_{opt}. The main disadvantage of SA is its requirement for a large amount of computation time. Because SA is based on Monte Carlo methods, which allows for a new state with a higher cost than that of a current state. To reduce this computation time, speedup strategies are used in this paper.

3. Problem Definition

We have a set of n SNPs, $\{s_1, s_2, \cdots, s_n\}$, of m patients and their disease existence. We also have machine learning methods such as SVM and decision tree which are able to estimate the accuracy of predicting disease existence for any given subset, $S \subseteq \{s_1, \cdots, s_n\}$. We define the accuracy of prediction by a certain machine learning method, M, as $p_M(S, n)$ for subset S. The number of possible subsets of $\{s_1, s_2, \cdots, s_n\}$ is $2n$-1. The goal is to find a subset S which gives maximal prediction accuracy, P_{max}.

$$\text{Select } S' \subseteq \{s_1, \cdots, s_n\} \text{ to maximize } p(S, n)$$

This is one of typical optimization problems. The simplest method to solve this problem is a brute force approach to evaluate all possible subsets and find one giving the best prediction accuracy. It guarantees the subset with maximum accuracy. However, if the number of SNPs is large, this approach is intractable. This is a class of problems for which it is believed that no efficient algorithm exists, called *NP-hard*. In other words, the algorithms that are guaranteed to find an optimal solution in reasonable time with the size of n may not exist. To overcome this problem, we can use the feature selection approach to select a small subset of SNPs as features for classification. Feature selection is a process commonly used in machine learning, wherein a subset of the features available from the data is selected for application of a learning algorithm. The best subset contains the least number of dimensions that most contribute to accuracy; we discard the remaining, unimportant dimensions. Feature selection approaches are not guaranteed to find optimal solution but very likely to find near-optimal solution in reasonable time.

4. Related Works

Simulated annealing is applied in many optimization problems since its introduction and the above-defined problem is not an exception. In [10], Pérez-Enciso applied simulated annealing with the Bayesian information criterion to identify the causal mutations for a trait. In [11], Schwender and Ickstadt proposed a procedure called

logicFS which used logic regression as a predictor and simulated annealing as search algorithm in logic regression to identify SNPs and SNP interactions associated with a disease. In [12], Üstünkar et el. applied SA to select representative SNPs for GWAS. They selected SNPs based on multiple testing adjusted p-value and applied proposed algorithm on different chromosomes and merged the SNP subsets to find representative SNP set. A Naive Bayes classifier was used for objective function.

5. Methods

5.1 Speedup Strategy

SA is composed of roughly two phases which are high-temperature phase and a low-temperature phase. In the high temperature phase, SA gives a high probability to all the states with higher costs than that of a current state. This allows any state in the solution set to be a current state. At a lower-temperature phase, SA gives a high probability to states with a lower or not much higher temperature than that of a current state. This allows only the states near a current state to be a current state. The high-temperature phase is similar to a random search, and the low-temperature phase is similar to a greedy local search. By combining two different search mechanisms SA may provide better solution compared to traditional greedy algorithms in relatively short period of time to exhaustive search algorithm. However, random search nature at the early stage of the process may take huge amount of time to get near the solution. To overcome this problem, we adopted forward selection (FS) algorithm to get the initial state of SA. In FS, solution set is empty at the beginning. Each feature which is not in the set is evaluated with the solution and one with the greatest improvement is added at each iteration. This procedure is repeated until additional SNP provides little improvement. To evaluate a feature, we applied decision tree to the previous solution with it and calculated the difference caused by the inclusion. The solution by FS is much closer to the optimal solution compared to the randomly chosen so that the time from the initial set to the one with lower temperature becomes unnecessary and more time can be devoted at the final stage of the process.

5.2 Transition Rule

Two rules can be applied to a current SNP set to generate a new SNP set. The basic two rules are as follows.

1) Insertion(m) : This operation inserts m SNPs to current SNP set randomly from not in the current SNP set. Initially information gain (IG) [13] value of each SNP is calculated and the value is applied to rank SNP by ascending order. Then the probability P_i to be inserted of a SNP i is $r_i / \frac{n(n-1)}{2}$ where r_i is the rank of SNP i. The reason for P_i is to give more weight for insertion to the SNP with higher prediction accuracy. During the insertion operation, 0 to m number of SNP can be inserted and the number m is decided randomly. In the experiment m is set to 3.
2) Deletion(m) : This operation deletes m SNPs randomly in the current SNP set. Initially information gain(IG) value of each SNP is calculated and the value is applied to SNPs to rank by descending order. Then the probability P_i to be deleted of a SNP i is $\mathrm{r_i} / \frac{\mathrm{n(n-1)}}{2}$ where r_i is the rank of SNP i. The reason for P_i is to give more weight for deletion to the SNP with lower prediction accuracy.

During the deletion operation, *0* to *m* number of SNP can be deleted and the number *m* is decided randomly. In the experiment *m* is set to 3.

5.3 Temperature Scheduling

The schedule implemented in SA is $T = T_i \times e^i$ here e is a constant defining the rate of annealing, i is the iteration number, T_i the initial temperature and T the current temperature. The value of e can easily be calculated from the total number of iterations, k, the final temperature, T_f, and T_i

$$e = \left(\frac{T_f}{T_i}\right)^{\frac{1}{k}} \quad (2)$$

Fig.2 shows the description of our SA.

```
begin SA
    S_current ← SNP set generated from the fast heuristic algorithm
    T ← T_i
    E_current ← prediction accuracy (S_current)
    S_opt ← S_current
  E_opt ← E_current
    while(T>T_f){
        S_new ← S_current ± SNP set applying transition rule
        E_new ← prediction accuracy (S_new)
       if Metropolis conditions are satisfied {
           S_current ← S_new
           E_current ← E_new
           if(E_opt < E_new) then
               E_opt ← E_new
               S_opt ← S_new
         }
        T ← T·e
    }
end SA
```

Figure 2. SA Algorithm

6. Experiments and Results

6.1 Data

We created case-control dataset using PLINK [14] with option `--simulate` and selected SNPs with p-value $\leq$ 0.001 in association study with `--assoc`. The first dataset consists of 50 SNPs of 45 cases and 45 controls. The second dataset consists of 200 SNPs of 100 cases and 100 controls. And in the third, there are 200 SNPs of 200 cases and 200 controls. Though the genotypes of each SNP can be used for decision tree, we use 1, 2 and 3 for minor homozygote, heterozygote, and major homozygote, respectively, to compare performance of other machine learning techniques such as SVM.

6.2 Evaluation Function

In this study, we used C4.5, an implementation of decision tree, as machine learning technique for disease prediction and confusion matrix [15] (see Fig. 3) to calculate

prediction accuracy. We selected a set of SNPs with best prediction accuracy as one with optimum disease relevant SNPs.

		Actual	
		+	-
Predict	+	True Positive (*TP*)	False Positive (*FP*)
	-	False Negative (*FN*)	True Negative (*TN*)

Figure 3. confusion Matrix

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

True positive (TP) means cases are correctly classified as cases and true negative (TN) contains controls classified as controls correctly. False negative (FN) means that cases are incorrectly classified as controls and false positive (FP) means that a control is misclassified as a case. Prediction accuracy by confusion matrix is used as cost for SA and the difference between those values of two consecutive states is to select the optimal set of SNPs.

First, we constructed an initial state by decision tree with forward selection algorithm and applied Leave-One-Out Cross Validation (LOOCV) [16] because the number of samples in datasets is not large enough to split samples into two sets, training and test. We selected a set of SNPs with the best prediction accuracy as the initial state.

The proposed SA began execution with the initial state at temperature T=20 and applied transition rule 100,000 times repeatedly until T=0.1. At each iteration, prediction accuracy of newly constructed set of SNPs was evaluated and if improved, it was selected as current state and otherwise it is determined by probability function. To evaluate prediction accuracy, decision tree and LOOCV were applied to the set.

6.3 Experimental Results

The experimental results were presented in Figure 4.

Dataset	Initial set		Final set	
	Number of SNPs	Prediction Accuracy	Number of SNPs	Prediction Accuracy
1	5	89.00%	19	90.50%
2	91	89.00%	86	90.00%
3	103	88.75%	86	91.75%

Figure 4. Experimental Results

As stated above, we used the result of DT with forward selection as the initial set for SA because it was already a near-optimal solution for the problem which made it possible to get around the problem of a random initial set.

For the first dataset, decision tree with forward selection provided a set of 5 SNPs out of 50 with 89.00 % of prediction accuracy and the proposed method improved it to

90.50 % with 19 SNPs. For the second, the initial state consisted of 91 with 89.00 % of accuracy and completed its execution with 86 with 90.00 %. In addition, it achieved higher prediction accuracy with less number of SNPs. For the third, we have 86 in the results set compared to 103 in the initial one, 91.75 % to 88.75 % of accuracy. We have observed that for all three datasets, the prediction accuracy was improved by the proposed method, which means that SA could improve prediction accuracy over that by the existing heuristic methods. In addition, the improved prediction accuracy was achieved by smaller number of SNPs, which means that SA can also be applied to minimize the number of SNPs in the solution. It also should be noted that the number of SNPs is large compared to the number of samples in the dataset: 86 SNPs for 200 samples and 86 for 400. With 86 SNPs, we can have 386 different combinations much larger than the number of samples. We can imagine that there is additional possibility to reduce number of SNPs if we analyze the classification tree.

7. Conclusion and Future Work

The curses of dimensionality of SNP data makes it infeasible to consider all possible combinations to find a set with the maximum prediction accuracy. Instead, several feature selection algorithms were adopted for that purpose such as forward selection, backward elimination, and so on. However, it provided fast in performance at the cost of optimality. Because of the greedy nature, the order of SNPs added is unchangeable, which resulted in the difference of accuracy. In this work, we adopt simulated annealing for this environment in which the SNPs already chosen can be eliminated in the following iteration steps, which makes it possible to get out of local minima and find better solution. And To cope with the problems caused by the random initial set and transition rule of simulated annealing, we used forward selection algorithm to select an initial set and proposed efficient rule. The experimental results showed that the proposed method could select a set of SNPs with improvement compared to the existing heuristic methods. In addition, it provides a set with smaller number of SNPs, which may lead cost reduction in medical diagnostics or personalized medicine. Based on the observation that the number of selected SNPs is large, we may reduce it by analyzing the classification tree.

Even though we used datasets generated by PLINK in this study, we think that the proposed method is applicable to real SNP dataset as a useful tool.

Acknow;dgements

This research was supported by Hallym University Research Fund, 2010(HRF-2010-028) and Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology (No. 2009-0077545).

References

[1] A. J. Brookes, GENE, vol. 8, no. 234, **(1999)**.

[2] J. R. Quinlan, Machine Learning, vol. 1, no. 1, **(1986).**

[3] J. Shavlik and T. Dietterich, Editors, "Readings in Machine Learning", Morgan Kaufmann, San Mateo, **(1986)**.

[4] J. R. Quinlan, "C4.5: Programs for Machine Learning", Morgan Kaufmann, San Mateo, **(1993)**.

[5] T.-S. Lim, W.-Y. Loh and Y.-S. Shih, "An empirical comparison of decision trees and other classification methods", TR-979, Dept. of Stat., Univ. of Wisconsin, Madison, **(1997)** June.

[6] R. Kohavi and G. H. John, Artificial Intelligence, vol. 97, no. 1-2, **(1996)**.
[7] A. J. Miller, "Subset Selection in Regression", Chapman and Hall, London, **(1990)**.
[8] S. Kirkpatrick, C. D. Gelatt and M. P. Vecchi, Science, vol. 220, no. 4598, **(1983)**.
[9] N. Metropolice, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller and E. Teller, J. Chemical Physics, vol. 21, no. 6, **(1953)**.
[10] M. Pérez-Enciso, Bioinformatics, vol. 22, no. 5, **(2006)**.
[11] H. Schwender and K. Ickstadt, Biostatistics, vol. 9, no. 1, **(2007)**.
[12] G. Üstünkar, S. Özöğür-Akyüz, G. Weber, C. Friedrich, Y. AydınSon and M. F. Christoph, Optimization Letters, **(2011)**, pp. 1–12.
[13] S. Kullback, "Information theory and statistics", John Wiley and Sons, NY, **(1959)**.
[14] http://pngu.mgh.harvard.edu/~purcell/plink/
[15] S. Geisser, "Predictive Inference: An Introduction", Chapman and Hall, New York, **(1993)**.
[16] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection", Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, vol. 2, no. 12, **(1995)**, December 10-14, pp.1137–1143, Adelaide, Australia.

Authors

Dong-Hoi Kim

Dong-Hoi Kim received MS and Ph.D degree in Computer engineering from Hallym University in 2002 and 2006, respectively. Since then he has been working as a lecturer on computer engineering at Hallym University. His research includes bioinformatics and data mining techniques.

Saangyong Uhmn

He received an MS degree in Computer Engineering at Hallym University in 1997 and worked as a lecturer at Hallym University. His research interests include distributed computing, algorithms and bioinformatics.

Jin Kim

Jin Kim received an MS degree in computer science from the college of Engineering at the Michigan State University in 1990, and in 1996 a PhD degree from the Michigan State University. Since then he has been working as a professor on computer engineering at the Hallym University. His research includes Bioinformatics and data mining.