STUDIA INFORMATICA 2011
Volume 32 Number 3B (99)

Jerzy MARTYNA
Jagiellonian University, Institute of Computer Science

MACHINE LEARNING FOR THE IDENTIFICATION OF THE DNA VARIATIONS FOR DISEASES DIAGNOSIS

Summary. In this paper we give an overview of a basic computational haplotype analysis, including the pairwaise association with the use of clustering, and tagged prediction (using Bayesian networks). Moreover, we present several machine learning methods in order to explore the association between human genetic variations and diseases. These methods include the clustering of SNPs based on some similarity measures and selecting of one SNP per cluster, the support vector machines, etc. The presented machine learning methods can help to generate a plausible hypothesis for some classification systems.

Keywords: computational haplotype analysis, SNP selection

UCZENIE MASZYNOWE DLA IDENTYFIKACJI ZMIAN DNA DO DIAGNOZOWANIA CHOROBY

Streszczenie. W pracy przedstawiono podstawowe metody uczenia maszynowego dla wyboru haplotypów, m.in. asocjacji par z użyciem klastrowania i przewidywania, znaczonego SNP (*Single Nucleotide Polimorhisms*), maszyny wektorów wspierających (ang. *Support Vector Machines*, SVM) itp. Metody te znajdują zastosowanie w przewidywaniu chorób. Mogą być także pomocne do generowania prawdopodobnych hipotez dla systemów klasyfikacji chorób.

Słowa kluczowe: obliczeniowa analiza haplotypów, wybór SNP

1. Introduction

The human genome can be viewed as a sequence of three billion letters from the nucleotide alphabet $\{A,G,C,T\}$. More than 99% of the positions of the genome possesse the same nucleotide. However, in the 1% of the genome numerous genetic variations occur, such as the diletion/insertion of a nucleotide, multiple repetitions of the nucleotide, etc. It is obvious that many diseases are caused by variations in the human DNA.

More than one million of the common DNA variations have been identified and published in the public database [29]. These identified common variations are called *single nucleotide polymorphisms* (SNPs). The nucleotides which occur often most in the population are referred to as the *major alleles.* Analogously, the nucleotides which occur seldom are defined as the *minor alleles*. For instance, nucleotide A (a major allele) occurs in a certain position of the genome, whereas nucleotide T (a minor allele) can be found in the some position of the genome.

Several diseases are identified by means of one of the SNP variations. The identification of the mutation of the SNP variations at a statistically significant level allows one to postulate a disease diagnosis. It is more often implemented by means of the use of the machine learning method.

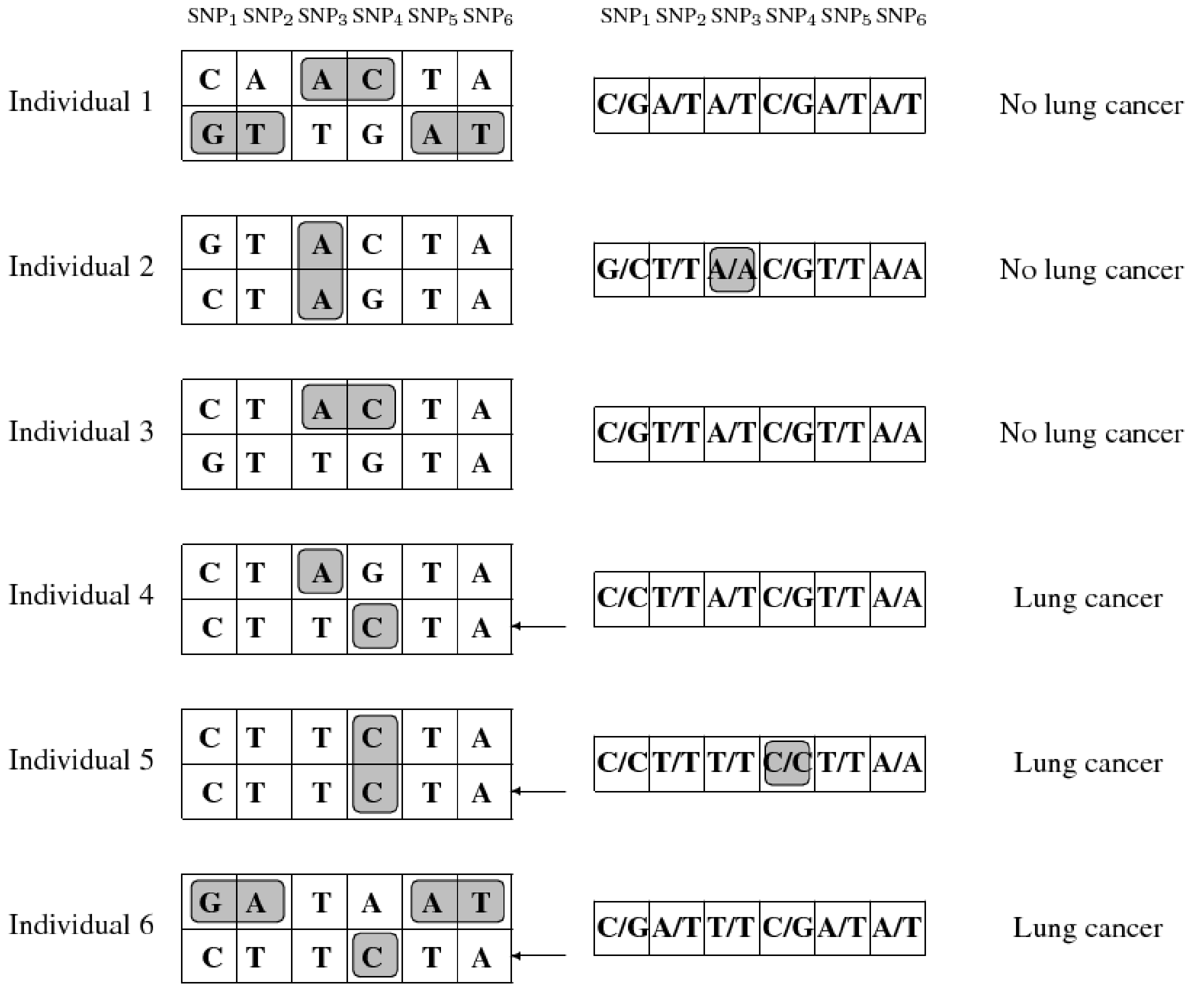
Currently, a haplotype analysis for the identification of the DNA variations relevant for the diagnosis of several diseases is used. We recall that the haplotype is a set of SNPs present in one chromosome. Thus, the machine learning methods for an effective haplotype analysis in order to identify several complex diseases are used.

Currently, a haplotype analysis for the identification of the DNA variations relevant for the diagnosis of several diseases is used. We recall that the haplotype is a set of SNPs present in one chromosome. Thus, the machine learning methods for an effective haplotype analysis in order to identify several complex diseases are used.

The main goal of this paper is to present some computational machine learning methods which are used in the haplotype analysis. This analysis includes the haplotype phasing, the tag SNP selection and identifying the association between the haplotype or a set of haplotypes and the target disease.

2. Basic Concepts in the Computational Analysis

Let us assume that all the species of chromosomes reproduced sexually have two sets: one inherited from the father and the other inherited from the mother. Every individual in this sample also has two alleles for each SNP, one of them in the paternal chromosome and the other in the maternal chromosome. Thus, for each SNP two alleles can be either the same or different. When they are identical, we refer to them as homozygous. Otherwise, when the alleles are different, the SNP is called heterozygous.



a) Haplotypes b) Genotypes c) Phenotypes

Fig. 1. Difference between haplotype, genotypes and phenotypes
Rys. 1. Różnica pomiędzy haplotypami, genotypami i fenotypami

Let our major allele of the SNP be colored gray and the minor colored black. Let us assume that the individual haplotype is composed of six SNPs constructed from his/her two chromosomes. Thus, a haplotype is a set of the SNPs present in one chromosome. Each of the haplotypes stems from the pair of the chromosomal samples and each pair is associated with one individual.

Genotypes are represented by two major alleles. When the combined allele is composed of the two major alleles, it is colored gray (see Fig. 1). In turn, when the SNPs have one minor allele and one minor allele, they are colored gray. In turn, when the SNPs have one minor allele and the other SNPs one major, then they are colored as white.

A phenotype is a typical observable manifestation of a genetic trait. In other words, a phenotype of an individual indicates a disease or lack of diseases (see Fig. 1c).

The haplotype analysis has more advantages than the single SNP analysis. The single SNP analysis cannot identify a combination of SNPs in one chromosome. For example, haplotype *CTTCTA* marked with arrow in Fig. 1a indicates the lung cancer phenotype, whereas the other individuals do not have lung cancer.

The haplotype analysis can be made in a traditional and a computational way. In the traditional analysis [22], [26] chromosome are separated, DNA clons, the hybrid constructed, and as a result haplotype – the disease indicated.

The traditional haplotype analysis is carried out biomolecular methods. However, this method is more costly than the computational analysis.

The computational haplotype analysis (which includes the haplotype phasing, the tag SNP selection) has been successfully applied to the study of diseases associated with haplotypes. This analysis can be considered by means of use the data mining methods.

3. Selected Methods of the Haplotype Phasing

3.1. The Pairwise Associated with the Use Clustering

The goal of the haplotype phasing is to find a set of haplotype pairs that can resolve all the genotypes from the genotype data. Formally, let the haplotype phasing problem be formulated as follows:

For a given $G = \{g_1, g_2, \dots, g_n\}$ set of n genotypes, where each genotype g_i consists of the allele information of m SNPs, $s_1, s_2, \dots, s_m$, namely

$$g_{ij} = \begin{cases} 0 & \text{when the two allele of SNP are major homozygous,} \\ 1 & \text{when the two allele of SNP are minor homozygous.} \\ 2 & \text{when the two allele of SNP are heterozygous.} \end{cases}$$

where $i = 1,2,\dots,n$, and $j = 1,2,\dots,m$.

The allele information of an SNP of a genotype is either major, minor or heterozygous. Each genotype represents the allele information of SNPs in two chromosomes. Like the genotype, each haplotype $h_i \in H$ consists of the same m SNPs $s_1, s_2, \dots, s_m$. Each haplotype represents the allele information of SNPs in one chromosome. We define haplotype h_i $(i = 1,2,\dots,2^m,\ j = 1,2,\dots,m$ as follows:

$$h_{ij} = \begin{cases} 0 & \text{when the allele of SNP is major,} \\ 1 & \text{when the allele of SNP is minor.} \end{cases}$$

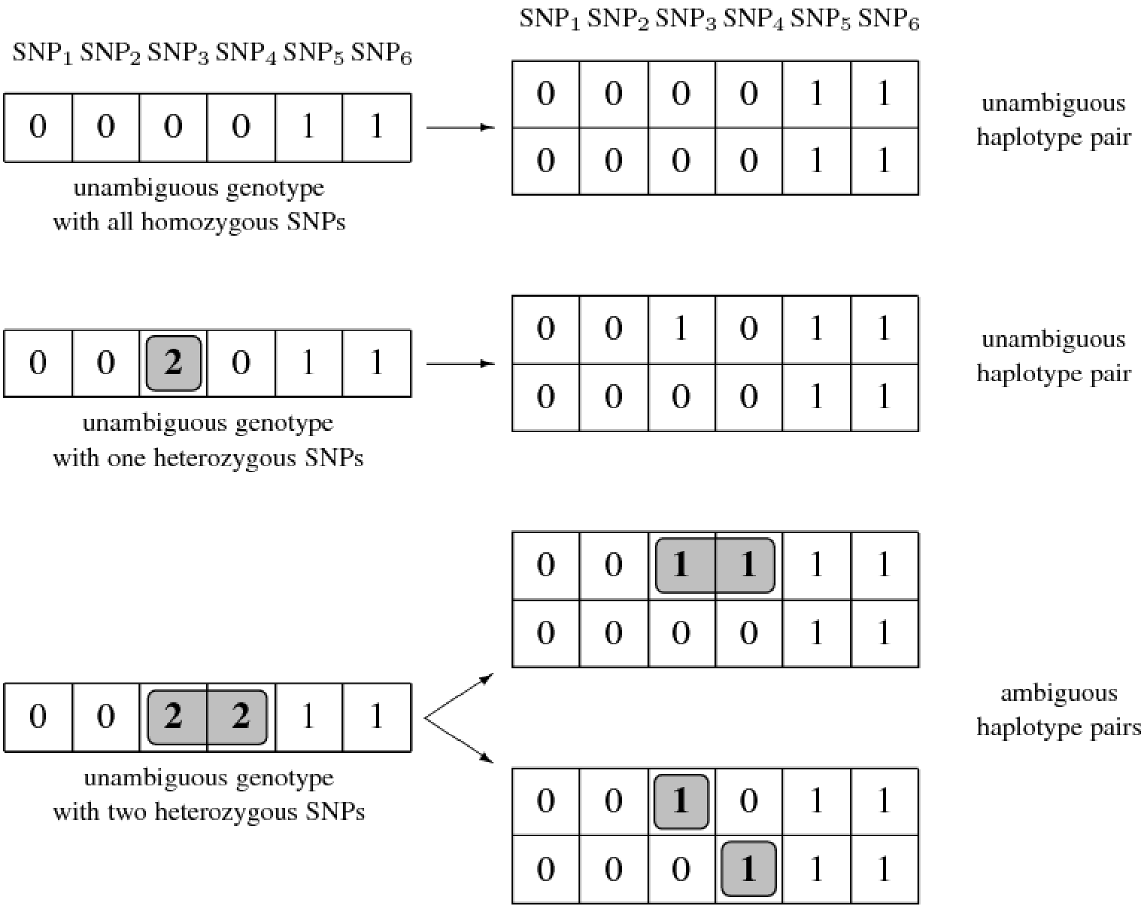

SNP1 SNP2 SNP3 SNP4 SNP5 SNP6
0 0 0 0 1 1
unambiguous genotype with all homozygous SNPs
SNP1 SNP2 SNP3 SNP4 SNP5 SNP6
0 0 0 0 1 1
0 0 0 0 1 1
unambiguous haplotype pair
0 0 2 0 1 1
unambiguous genotype with one heterozygous SNPs
0 0 1 0 1 1
0 0 0 0 1 1
unambiguous haplotype pair
0 0 2 2 1 1
unambiguous genotype with two heterozygous SNPs
0 0 1 1 1 1
0 0 0 0 1 1
0 0 1 0 1 1
0 0 0 1 1 1
ambiguous haplotype pairs

Fig. 2. Finding a set of haplotype pairs and ambiguous genotypes
Rys. 2. Znajdowanie par haplotypów i niejednoznaczne genotypy

Now we can formulate the haplotype phasing problem as follows:

Problem : Haplotype phasing

Input : A set of genotypes $G = \{g_1, g_2, \dots, g_n\}$

Output : A set of n haplotype pairs

$$O = \{\langle h_{i1}, h_{i2}\rangle \mid h_{i1} \oplus h_{i2} = g_i, h_{i1}, h_{i2} \in H, i = 1, 2, \dots, n\}$$

The haplotype phasing is shown in Fig. 2. Three genotype data are given on the left side. When the two alleles of SNPs are homozygous, the SNPs are with the same color. When the two alleles in the genotype are of an SNP, have one heterozygous the haplotype pairs are identified unequivocally. When the two alleles in the genotype have two heterozygous, the haplotype pairs cannot be identified unequivocally. Thus, the genotype is identified by means of an additional biological analysis method.

We can use following methods in the haplotype phasing:

1) parsimony,
2) phylogeny,
3) the maximum likelihood (ML),

4) the Bayesian inference.

The first two methods are treated as a combinatorial problem [14]. The last two methods are based on the data mining approach and therefore are presented here.

3.2. The maximum likelihood (ML) method for the haplotype phasing

The maximum likelihood method can be based on the expectation-maximization (EM) method. This method, among others described in [14], works as follows:

Let D be the genotype data of n individuals. Each of their genotypes consists of SNPs. Let n' be the number of distinct genotypes. We denote the ith distinct genotype by g_i, the frequency of g_i in the data set D by f_i, the number of the haplotype pairs resolving g_i $(i = 1,2,...,n') = 1)$ by c_i. When H is a set of all haplotypes consisting of the same m SNPs, the number of haplotypes in H is equal to 2^m. Although the haplotype population frequencies $\Theta = \{p_1, p_2, ..., p_{2^m}\}$ are unknown, we can estimate them by the probability of the genotypes comprising the genotype data D, namely

$$L(D) = \Pr(D \mid \Theta) \approx \prod_{i=1}^{n'} \Pr(g_i \mid \Theta)^{f_i} = \prod_{i=1}^{n'} \left(\sum_{j=1}^{c_i} \Pr(h1_{ij}, h2_{ij} \mid \Theta \right)^{f_i} \qquad (1)$$

where $h1_{ij}$, $h2_{ij}$ are the haplotype pairs resolving the genotype g_i.

The EM method depends on the initial assignment of values and does not guarantee a global optimum of the likelihood function. Therefore, this method should be run multiple times with several initial values.

3.3. The Bayesian Inference Markov Chain Monte Carlo with the Use of the Haplotype Phasing Problem

The Bayesian inference methods are based on the computational statistical approach. In comparison with the EM method, the Bayesian inference method aims to find the posterior distribution of the model parameters given in the genotype. In other words, with the use of the EM method the haplotype population frequencies, θ, give a set of unknown frequencies in a population, and the Bayesian inference method provides the a posteriori probability $\Pr(H \mid D)$. The Markov Chain Monte Carlo metod approximates samples from $\Pr(H \mid D)$.

Some of the basic MCMC algorithms are:

a) the Metropolis-Hastings algorithm,

b) the Gibbs sampling.

Ad a) The Metropolis-Hastings algorithm was introduced in the papers [15], [25]. The method starts at $t = 0$ with the selection of $X^{(0)} = x^{(0)}$ drawn at random from some starting distribution g, with the requirement that $f(x^{(0)})>0$. Given $X^{(t)}=x^{(t)}$, the algorithm generates $X^{(t+1)}$ as follows:

1) Sample a candidate value $X^{\bullet}$ from the proposed distribution $g(\cdot\,|\,x^{(t)})$
2) Compute the Metropolis-Hastings ratio $R(x^{(t)},X^{\bullet})$, where

$$R(u,v)=\frac{f(v)g(u\,|\,v)}{f(u)g(v\,|\,u)} \tag{2}$$

$R(x^{(t)},X^{\bullet})$ is always defined, because the proposal $X^{\bullet}=x^{\bullet}$ can only occur if $f(x^{(t)})>0$

and $g(x^{\bullet}\,|\,g^{(t)})>0$.

3) Sample a value for $X^{(t+1)}$ according to the following

$$X^{(t+1)}=\begin{cases} X^{\bullet} & \text{with probability } \min\{R(x^{(t)},X^{\bullet}),1\} \\ x^{(t)} & \text{otherwise} \end{cases}$$

4) Increment t and return to step 1.

A chain constructed by the Metropolis-Hastings algorithm is Markov, since $X^{(t+1)}$ is only dependent on $X^{(t)}$. Note that depending on the choice of the proposed distribution we obtain an irreducible and aperiodic chain. If this check confirms irreducibility and aperiodicity, then the chain generated by the Metropolois-Hastings algorithm has a unique limiting stationary distribution.

Ad b) The Gibbs sampling method is specifically adapted for a multidimensional target distribution. The goal is to construct a Markov chain whose stationary distribution equals the target distribution f.

Let $X=(x_1,...,x_p)^T$ and $X_{-i}=(X_1,...,X_{i-1},X_{i+1},...,X_p)^T$. We assume that the univariate conditional density of $X_i\,|\,X_{-i}=x_{-i}$ denoted by $f(x_i\,|\,x_{-i})$ is sampled for $i=1,2,...,p$. Then from a starting value $x^{(0)}$, the Gibbs sampling mthod can be described as follows:

1) Choose an ordering of the components of $x^{(t)}$
2) For $\forall i$ sample $X_i^{\bullet}\,|\,x_{-i}^{(t)}\approx f(x_i\,|\,x_{-i}^{(t)})$
3) Once step 2 has been completed for each component of X in the selected order, set $X^{(t+1)}=X^{\bullet}$.

The chain produced by the Gibbs sampler is a Markov chain. As with the Metropolis-Hastings algorithm, we can use the realization from the chain to estimate the expectation of any function of X.

Finally, the Bayesian inference method using the MCMC can be applied to samples consisting of a large number of SNPs or to samples in which a substantial portion of haplotypes occur only once. Furthermore, the Gibbs sampler is a popular genetic model that denotes a tree describing the evolutionary history of a set of DNA sequences [16].

4. Machine Learning Methods for Selecting Tagging SNPs

4.1. The Problem Formulation

The tag SNP selection problem can be formulated as follows: Let $S = \{s_1,...,s_n\}$ be a set of n SNPs in a studied region, $D = \{h_1,...,h_m\}$ be a data set of m haplotypes that consist of the n SNPs. According to definition 1, we assume that $h_i \in D$ is a vector of size n whose vector is a vector of size n whose vector element is 0 when the allele of a SNP is major and 1 when it is minor. Let the maximum number of the haplotypes consisting SNPs (htSNPs) be k.

We assume that function $f(T',D)$ provides a measure as to how well subset $T' \subset S$ represents the original data D. Thus, the tag SNP selection is given by

problem the tag SNP selection
input 1) a set of SNPs,
2) a set of haplotypes D,
3) a maximum number of htSNPs,
output a set of htSNPs T which is $T = \arg\max_{T' \subset S \;\; and \;\; |T'| \leq k} f(T',D)$.

In other words, the tag SNP selection consists on finding an optimal subset of SNPs of size k at most based on the given evaluation function f among all possibile subsets of the original SNPs.

Among the tag SNP selection methods based on the machine learning methods most often included are [22]:

1) the pairwise association with the use of clustering
2) the tagged SNP prediction with the use of Bayesian networks.

Now, we present these machine learning methods used for the tag SNP selection.

4.2. The Pairwise Association with the Use of Clustering

The cluster analysis for the paiwise association for the tag SNP selection was at first used by Byng et al. [4]. This method works as follows: The original set of SNPs is divided into hierarchical clusters. Within the cluster all SNPs are with a predefined level σ (typically

$\sigma > 0.6$) [4]. In other works, a.o. [1, 5] within each cluster the pairwise linkage equilibrium (LD).

In the papers [1, 5] is used so-called the pairwise linkage equilibrium (LD), given the joint probability of two alleles s_{1_i} and s_{2_j} equal to the product of the allele individual probabilities. Thus, under the assumption that these probabilities are independent, we have the LD [19], [12] given by

$$\forall_{ij} \Pr(s_{1_i}, s_{2_j}) = \Pr(s_{1_i}) \cdot \Pr(s_{2_j}) \tag{3}$$

For the two SNPs within the discrete region called a block here the LD is high, while for the two SNPs belonging to different regions it is small. Unfortunately, there is no agreement on the definition of the region [28, 13].

According to the clustering methods based on the LD pairwise, the LD parameter between htSNP and all the other SNPs is greater than the threshold level. These methods include:

1) the minimax clustering,
2) the greedy binning algorithm.

Ad 1) The former, *the minimax clustering* [1] is defined as $D_{\min\max}(C_i, C_j) = \min_{\forall s \in (C_i \vee C_j)}(D_{\max}(s))$, where $D_{\max}(s)$ is the maximum distance between the SNPs and all other SNPs in the two clusters. According to this method every SNP formulates its own cluster. Further, the two closest clusters are merged. The SNP defining the minimax distance is treated as a representative SNP for the cluster. The algorithm stops when the smallest distance between the two clusters is larger than level $1-\sigma$. Thus, the representative SNPs are selected as a set of htSNPs.

Ad 2) The latter, the *greedy binning* algorithm, initially examines all the pairwise LD between SNPs, and for each SNP counts the number of other SNPs whose pairwise LD with the SNP is greater than the prespecified level, σ. The SNP with the largest count is then clustered with its associated SNPs. Thus, this SNP becomes the htSNP for this cluster. This procedure is iterated until all the SNPs are clustered.

The pairwise association-based method for the tag SNP selection can be used for a disease diagnosis. The complexity of this method lies between $O(mn^2 \log n)$ and $O(cmn^2)$ [32, 5], where the number of clusters is equal to c, the number of haplotypes is equal to $m$, the number of SNPs is equal to n.

4.3. The Tag SNP Selection Based on Bayesian Networks (BN)

The tagged SNP prediction with the use of on Bayesian networks was first used by Bafna [2]. Recently, Lee at al. [23] proposed a new prediction-based tag SNP selection method, called the BNTagger, which improves the accuracy of the study.

The BNTagger method of the tag SNP selection uses the formalism of BN. The BN is a graphical model of joint probability distributions that comprises conditional independence and dependence relations between its variables [18]. There are two components of the BN: a *directed acyclic graph*, G and a set of conditional probability distributions, $\theta = \{\theta_1, ..., \theta_p\}$. With each node in graph G a random variable X_j is associated. An edge between the two nodes gives the dependence between the two random variables. The lack of an edge represents their conditional independence. This graph can be automatically learned from the data. With the use of the learned BN it is easy to compute the posterior probability of any random variable.

5. Machine Learning Methods for the Tag SNP Selection for the Sake of Disease Diagnosis

5.1. The Feature Selection with the Use of the Similarity Method

The feature selection with the use of the feature similarity (FSFS) method was introduced by Phuong [27]. This method works as follows:

We assume that N haploid sequences considering m SNPs are given. Each of them is represented by $N \times m$ matrix M with the sequences as rows and SNPs as columns. Each element of this matrix which represents the j-th alleles of the i-th sequence is equal to 0, 1, 2. 0 representing the missing data, 1 and 2 represent two alleles. The SNPs represents the attributes that are used to identify the class to which the sequence belongs.

The machine learning problem is formulated as follows: how to select a subset of SNPs chich can classify all haplotypes with the required accuracy. A measure of similarity between pairs of features in the FSFS method is given by

$$r^2 = \frac{(p_{AB} \cdot p_{ab} - p_{Ab} \cdot p_{aB})^2}{p_{AB} \cdot p_{ab} \cdot p_{Ab} \cdot p_{aB}}, \quad 0 \le r \le 1 \tag{4}$$

where A and a are the two alleles at a particular locus, p_{xy} is the frequency of observing alleles x and y in the same haplotype, p_x is the frequency of allele x alone.

The details of the algorithm used in the FSFS method [27] are given in the procedure presented in Fig. 3. As the input parameters are used S – the original set of SNP and K – the number of nearest neighbors of an SNP to consider. The algorithm initializes R to S. In each iteration the distance d_i^K between each SNP F_i in R and its K-th nearest neighbouring SNP is computed. Further, the FSFS algorithm removes its K nearest SNPs from R. In the next step is comparing the cardinality of R with K and adjusting K. Thus, the condition $d_0^K > 0$ is gradually decreased until d_0^K is less or equal to an error threshold θ.

The parameter K is chosen for as long as the desired prediction accuracy is achieved. In the experimental results given by Daly et al. [8] that the FSFS method can give a prediction accuracy of 88% with only 100 tag SNPs.

```
Input data:    $S$ – set of SNP, parameter K of the algorithm,
Output data:   $R$ – selected Tag SNPs,
1. select $R$ from $S$;
2. for $\forall F_i \in R$ do
         $d_i^K := D(F_i, F_i^K)$      /* $F_i^K$ is the $K$-th nearest SNP of $F_i$ in $R$
   endfor;
3. find $F_0$ such that $d_0^K := \arg\min F_{i \in R}(d_i^K)$;
   Let $F_0^1, F_0^2, \ldots, F_0^K$ be the nearest SNPs of $F_0$ and $R := R - \{F_0^1, \ldots, F_0^K\}$
   Initially $\theta = d_0$
4. if $K > |R| - 1$ then $K := R - 1$;
5. if $K = 1$ then goto 1;
6. while $d_0^K > \theta$ do
      begin
         $K := K - 1$;
         if $K = 1$ then goto 1;
         compute $d_0^K$;
      end;
7. goto 2;
8. if all $R$ are selected from $S$ then stop;
```

Fig. 3. FSFS algorithm for TAG SNP selection
Rys. 3. Algorytm FSFS dla wyboru znaczonego SNP

5.2. An Application of the SVM for the Tag SNP Selection for Disease Diagnosis

In this section, we describe an application the SVM method for the tag SNP selection with a simultaneous disease diagnosis.

The support vector machine (SVM) [30] is a machine learning method which was used to outperform other technologies, such as neural networks or k-nearest neighbor classifier. Moreover, the SVM has been succesfully applied for a binary prediction multiple of cancer types with excellent forecasting results [33, 20]. We recall that the SVM method finds an optimal maximal margin hyperplane separating two or more classes of data and at the same time minimizes classification error. The mentioned margin is the distance between the hyperplane and the closest data points from all the classes of data.

The solution of an optimization problem with the use of the SVM method requires a solution of a number of quadratic programming (QP) problems. It involves two parameters: the penalty parameter C and the kernel width σ. If $\sigma^2 \to \infty$ $C \to \infty$ is not fit for the problem under consideration because it has noise. If σ^2 and $C = C_1\sigma^2$ where C_1 is fixed then the SVM converges with the linear SVM classifier with the penalty parameter C_1. A well selected (C, σ^2) is crucial for unknown data prediction. In the paper [3] the procedure for finding good C and σ^2 was given.

Table 1

The prediction accuracy of existing metods

No.	Author(s)	Method	ALL/AML	Breast cancer	Colon	Multiple myeloma	SRBCT
1	Cho [6]	genetic algorithm	73.53% (1)	77.3% (3)			
2	Cho [7]	genetic algorithm	94.12% (17)	100% (21)			
3	Deb et al. [21]	evolutionary algorithm			97% (7)		
4	Deutsch [11]	evolutionary algorithms					100% (21)
5	Huang [17]	genetic algorithm and SVM					98.75% (6.2)
6	Lee [21]	Bayesian interference		100% (10)			
7	Lee [24]	SVM					100% (20)
8	Waddell [31]	SVM				71%	

Note: ALL/AML – acute lymphoblastic leukemia/acute myeloid leukemia,
SRBCT – small round blue cell tumor,
numbers in parentheses denote the number of selected genes.

According to the output results given by Waddell et al. [31 concerning the case of the multiple myeloma (about 0.035% people over 70 and 0.002% people between the age of 30 - 54 in the USA) it was possible to detect differences in the SNP patterns between the good human genome and the people diagnosed with this disease.

The obtained accuracy achieved 71% of the overall classification accuracy. Although the accuracy was not high, it was significant that only relatively sparse SNP data are used for this classification. The comparison of the SVM method with other existing methods is given in Table 1. It is noticeable that these methods are complementary. From Table 1 we see that the existing methods tend to select many genes with poor prediction accuracy. However, the SVM metod selects genes with relatively high prediction accuracy.

6. Conclusion

We have presented some machine learning methods concerning the tag SNP selection, additionally, some of which are used to diagnose diseases. These methods are applied to data sets with hundreds of SNPs. In general, they are inexpensive and with varying accuracy for the haplotype phasing, the tagged SNP prediction and, furthermore, diesease diagnosing. The missing alleles, genotyping errors, a low LD among SNPs, a small size of sample, lack of scalability with the increase of the number of markers are among basic weaknesses of the currently used machine learning methods used for computational haplotype analysis.

Nevertheless, the machine learning methods are more and more often used in the tag SNP selection and disease diagnosis.

BIBLIOGRAPHY

1. Ao S. I., Yip K., Ng M., Cheung D., Fong P., Melhado I., Sham P. C.: CLUSTAG: Hierarchical Clustering and Graph Methods for Selecting Tag SNPs. Bioinformatics, Vol. 21, 2005, p. 1735÷1736.
2. Bafna V., Halldörsson B. V., Schwartz R., Clark A. G., Istrail S.: Haplotypes and Informative SNP Selection Algorithms: Don’t Block out Information. [in:] Proc. of the Seventh Int. Conf. on Computational Molecular Biology, 2003, p. 19÷26.
3. Boser B. E., Guyon I. M., Vapnik V.: A Training Algorithm for Optimal Margin Classifiers. Fifth Annual Workshop on the Computational Learning Theory, ACM, 1992.
4. Byng M. C., Whittaker J. C., Cuthbert A. P., Mathew C. G., Lewis C. M.: SNP Subset Selection for Genetic Association Studies. Annals of Human Genetics, Vol. 67, 2003, p. 543÷556.
5. Carlson C. S., Eberle M. A., Rieder M. J., Yi Q., Kruglyak L., Nickerson D. A.: Selecting a Maximally Informative Set of Single-nucleotide Polymorphisms for Association Analyses Using Linkage Disequilibrium. American Journal of Human Genetics, Vol. 74, 2004, p. 106÷120.

6. Cho J. H., Lee D., Park J. H., Lee I. B.: New Gene Selection Method for Classification of Cancer Subtypes Considering Within-Class Variation. FEBS Letters, Vol. 551, 2003, p. 3÷7.

7. Cho J. H., Lee D., Park J. H., Lee I. B.: Gene Selection and Classification from Microarray Data Using Kernel Machine. FEBS Letters, Vol. 571, 2004, p. 93÷98.

8. Daly M., Rioux J., Schaffner S., Hudson T., Lander E.: High-Resolution Haplotype Structure in the Human Genome. Nature Genetics, Vol. 29, 2001, p. 229÷232.

9. Deb K., Reddy A. R.: Reliable Classification of Two-Class Cancer Using Evolutionary Algorithms. Biosystems, Vol. 72, 2003, p. 111÷129.

10. Dempster A. P., Laird N. M., Rubin D. B.: Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society, Vol. 39, No. 1, 1977, p. 1÷38.

11. Deutsch J.: Evolutionary Algorithms for Finding Optimal Gene Sets in Microarray Prediction. Bioinformatics, Vol. 19, No. 1, 2003, p. 45÷52.

12. Devlin B., Risch N.: A Comparison of Linkage Disequilibrium Measures for Fine Scale Mapping. Genomics, Vol. 29, 1995, p. 311÷322.

13. Ding K., Zhou K., Zhang J., Knight J., Zhang X., Shen Y.: The Effect of Haplotype-Block Definitions on Inference of Haplotype-block Structure and htSNPs Selection. Molecular Biology and Evolution, Vol. 22, No. 1, 2005, p. 48÷159.

14. Gusfield D., Orzack S. H.: Haplotype Inference. CRC Handbook in Bioinformatics, CRC Press, Boca Raton, 2005, p. 1÷25.

15. Hastings W. K.: Monte Carlo Sampling Methods Using Markov Chains and Their Applications. Biometrika, Vol. 57, 1970, p. 97÷109.

16. Hedrick P. W.: Genetics of Population. Jones and Bartlett Publishers, Sudbury 2004.

17. Huang H. L., Chang F. L.: ESVM: Evolutionary Support Vector Machine for Automatic Feature Selection and Classification of Microarray Data. Biosystems, Vol. 90, 2007, p. 516÷528.

18. Jensen F.: Bayesian Networks and Decision Graphs. Springer-Verlag, New York, Berlin Heidelberg 1997.

19. Jorde L. B.: Linkage Disequilibrium and the Search for Complex Disease Genes. Genome Research, Vol. 10, 2000, p. 1435÷1444.

20. Keerthi S. S., Lin C. J.: Asymptotic Behaviour of Support Vector Machines with Gaussian Kernel. Neural Computing, Vol. 15, No. 7, 2003, p. 1667.

21. Lee K. E., Sha N., Dougherty E. R., Vannucci M., Mallick B. K.: Gene Selection: A Bayesian Variable Selection Approach. Bioinformatics, Vol. 19, No. 1, 2003, p. 90÷97.

22. Lee P. H.: Computational Haplotype Analysis: An Overview of Computational Methods in Genetic Variation Study. Technical Report 2006-512, Queen’s University, 2006.

23. Lee P. H., Shatkay H.: BNTagger: Improved Tagging SNP Selection Using Bayesian Networks. The 14th Annual Int. Conf. on Intelligent Systems for Molecular Biology (ISMB), 2006.
24. Lee Y., Lee C. K.: Classification of Multiple Cancer Types by Multicategory Support Vector Machines Using Gene Expression Data. Bioinformatics, Vol. 19, No. 1, 2003, p. 1132÷1139.
25. Metropolis N., Rosenblum A. W., Rosenbluth M. N., Teller A. H., Teller E.: Equation of State Calculation by Fast Computing Machines. Journal of Chemical Physics, Vol. 21, 1953, p. 1087÷1091.
26. Nothnagel M.: The Definition of Multilocus Haplotype Blocks and Common Diseases. Ph.D. Thesis, University of Berlin, 2004.
27. Phuong T. M., Lin Z., Altman R. B.: Choosing SNPs Using Feature Selection. Proc. of the IEEE Computational Systems Bioinformatics Conference, 2005, p. 301÷309.
28. Schulze T. G., Zhang K., Chen Y., Akula N., Sun F., McMahonen F. J.: Defining Haplotype Blocks and Tag Single-nucleotide Polymorphisms in the Human Genome. Human Molecular Genetics, Vol. 13, No. 3, 2004, p. 335÷342.
29. Sherry S. T., Ward M. H., Kholodov M., Baker J., Phan L., Smigielski E. M., Sirotkin K.: dbSNP: the NCBI Database of Genetic Variation. Nucleic Acids Research, Vol. 29, 2001, p. 308÷311.
30. Vapnik V.: Statistical Learning Theory. John Wiley and Sons, New York 1998.
31. Waddell M., Page D., Zhan F., Barlogie B., Shaughnessy J. Jr.: Predicting Cancer Susceptibility from Single-nucleotide Polymorhism Data: a Case Study in Multiple Myeloma. Proc. of BIOKDD '05, Chicago, August 2005.
32. Wu X., Luke A., Rieder M., Lee K., Toth E. J., Nickerson D., Zhu X., Kan D., Cooper R. S.: An Association Study of Angiotensiongen Polymorphisms with Serum Level and Hypertension in an African-American Population. Journal of Hypertension, Vol. 21, No. 10, 2003, p. 1847÷1852.
33. Yoonkyung L., Cheol-Koo L.: Classification of Multiple Cancer Types by Multicategory Support Vector Machines Using Gene Expression Data. Bioinformatics, Vol. 19, No. 9, 2003, p. 1132.

Recenzent: Dr inż. Jerzy Respondek

Wpłynęło do Redakcji 11 marca 2011 r.

Omówienie

W pracy dokonano przeglądu podstawowych metod obliczeniowych stosowanych w eksploracji danych przy wyborze minimalnego podzbioru pojedynczego polimorfizmu nukleotydów (ang. *Single Nucleotide Polimorphisms*, SNP). Wybór ten jest oparty na haplotypach i pozwala on na znalezienie wszystkich SNP związanych z daną chorobą. W rezultacie, takie metody, jak asocjacja par z użyciem klastrowania, metoda maksymalnej wiarygodności (ang. *maximum likelihood metod*), algorytm Metropolis-Hastings, maszyna wektorów wspierających (ang. *suport vector machine*, SVM) itp., mają duże znaczenie w diagnozowaniu chorób onkologicznych. Metody te różnią się zarówno uzyskiwaną dokładnością, jak i liczbą genów branych pod uwagę.

Address

Jerzy MARTYNA: Jagiellonian University, Institute of Computer Science,
ul. Prof. S. Łojasiewicza 6, 30-348 Kraków, Poland, martyna@softlab.ii.uj.edu.pl.