ISTITUTO DI ANALISI DEI SISTEMI ED INFORMATICA
"Antonio Ruberti"
CONSIGLIO NAZIONALE DELLE RICERCHE

P. Bertolazzi, G. Felici, P. Festa

LOGIC BASED METHODS FOR SNPS TAGGING AND RECONSTRUCTION

R. 667 2007

Paola Bertolazzi – Istituto di Analisi dei Sistemi ed Informatica del CNR, viale Manzoni 30 - 00185 Roma, Italy. Email: bertola@iasi.rm.cnr.it.

G. Felici – Istituto di Analisi dei Sistemi ed Informatica del CNR, viale Manzoni 30 - 00185 Roma, Italy. Email: felici@iasi.rm.cnr.it.

P. Festa – Dipartimento di Matematica e Applicazioni "R.M. Caccioppoli", Università degli Studi di Napoli FEDERICO II, Compl. MSA, Via Cintia, 80126 Napoli Italy, e-mail: paola.festa@unina.it.

ISSN: 1128–3378

Collana dei Rapporti dell'Istituto di Analisi dei Sistemi ed Informatica "Antonio Ruberti", CNR

viale Manzoni 30, 00185 ROMA, Italy

tel. ++39-06-77161
fax ++39-06-7716461
email: iasi@iasi.rm.cnr.it
URL: http://www.iasi.rm.cnr.it

Abstract

SNPs are positions of the DNA sequences where the differences among individuals are embedded. The knowledge of such SNPs is crucial for disease association studies, but even if the number of such positions is low (about 1% of the entire sequence), the cost to extract the complete information is actually very high. Recent studies have shown that DNA sequences are structured into blocks of positions, that are conserved during evolution, where there is strong correlation among values (alleles) of different loci. To reduce the cost of extracting SNPs information, the block structure of the DNA has suggested to limit the process to a subset of SNPs, the so called Tag SNPs, whose knowledge allows to derive all the others. In this paper we apply a technique for feature selection based on integer programming to the problem of Tag SNP selection. Moreover, to test the quality of our approach, we tackle the problem of SNPs reconstruction, i.e. the problem of deriving unknown SNPs from the value of Tag SNPs and propose two reconstruction methods, one based on a majority vote and the other on a machine learning approach. We test our algorithm on two public data sets of different nature, providing results that are, when comparable, in line with the related literature. One of the interesting aspects of the proposed method is to be found in its capability to deal simultaneously with very large SNPs sets, and, in addition, to provide highly informative reconstruction rules in the form of logic formulas.

1. Introduction

It is well known that genetic variation among different individuals is limited to a small percentage of positions in DNA sequences (99% of two DNA molecules being identical). These positions are called Single Nucleotide Polymorphisms (SNPs) and are characterized by the fact that two possible values (alleles) of the four bases (T, A, C, G) are observed across a population at such sites, and that the minor allele frequency is at least 5%. The knowledge of such polymorphisms is considered crucial in disease association studies over a population of individuals, and is the target of the HapMap project, that has already released a public data base of one million SNPs *genotyped* from four populations of three geographical areas (Africa, East Asia, and Europe). Genome-wide association studies aims to identify common genetic factors that influence health and disease. A genome-wide association study is defined as any study of genetic variation across the entire human genome that is designed to identify genetic associations with observable traits (such as blood pressure or weight), or the presence or absence of a disease or condition. Even if the number of SNPs identified in human genome is low (about seven millions of sites) w.r.t. the complete sequence, the costs for extracting this knowledge is prohibitive, depending both on the number of examined individuals and on the number of typed SNPs (see [16] for a background on SNPs methods and technologies). To reduce the costs of these studies the dimension of data set must be kept small both in the number of individuals and in the number of typed sites. However, if the population dimension is too small, the probability of obtaining meaningless results is high, hence we have to reduce the number of SNPs.

Given a population of individuals, one of the major research challenges in the last years has been to find a selected number of SNPs (*Tag SNPs*) that are representative of all the other SNPs. This approach is supported by the observation that DNA molecules have a *block structure* [19, 10, 12, 23]. Blocks are subsequences of DNA that have been transmitted during the evolution without splits in the sequence. A result of block transmission is *Linkage Disequilibrium* (LD), a parameter related to some combinations of alleles or genetic markers that in a population can occur more or less frequently than would be expected from a random formation of haplotypes from alleles based on their frequencies. In case of two SNPs a measure of LD is given by $\delta = p_1p_2 - h_{12}$, where p_1p_2 denote the marginal allele frequencies at the two loci and h_{12} denotes the haplotype frequency in the joint distribution of both alleles. A block is a region of the DNA where for each pair of SNPs $\delta \neq 0$; this means that the information contained in a block is redundant and suggests that it is possible to find a small set of SNPs able to predict all the others. Such SNPs are commonly called Tag SNPs and the problem is called *Tag SNP Selection* (TSS for short).

Most of the previous work on TSS [13, 14, 24] restricts to find a partition of SNPs sequences into blocks and to identify a set (eventually of minimum cardinality) of SNPs that are representative of the entire sequence. However, a crucial aspect in association studies is to find a set of Tag SNPs and to design a *reconstruction method* such that all the other SNPs can be predicted from the Tag SNPs through the proposed method. This topic has been recently addressed by a number of studies (see [18] for a complete bibliography). In this paper, we refer mainly to two of these works: the work of Shamir [15] and the work of Paschou [18].

In [15] a dynamic programming algorithm is described that takes as input a given set of DNA sequences (training set) and identifies a set of Tag SNPs used to predict all the other SNPs of the given set. The measure of the goodness of the identified set of Tag SNPs is the prediction error, i.e. the number of wrong SNPs predicted for each sequence. The prediction function is based on the biological hypothesis that each SNP is strongly related only with the two neighboring Tag SNPs (hypothesis strictly related to the Linkage Disequilibrium structure of the DNA molecules). This results in an algorithm whose computational complexity is sufficiently low. The prediction function uses a majority vote in order to determine which value is more likely to appear in the unknown position. In the paper, a random algorithm to find the Tag SNPs is also proposed. The two algorithms are tested on different training sets with the leave-one-out testing strategy and compared with other state-of-the-art algorithms.

In [18] an algorithm based on linear algebra is proposed, that identifies a set of Tag SNPs and a set of linear formulas on these Tag SNPs able to predict all the other SNPs. The algorithm is tested on a large set of data from public data bases and from association studies. In this approach, since each predicted SNP is a function of all the Tag SNPs, it is possible to keep into account possible LD between distant

h_1	C	C	T	A	T	G	C
h_2	A	C	T	A	G	G	A
g_{12}	C/A	C	T	A	T/G	G	C/A

Figure 1: Genotypes.

g_1	C	C	G/T	A	G	A	C
g_2	C/A	C	T	A	G	T	A
g_3	C/A	A	T	A	-	C/ A	A
g_3	3	2	1	1	0	3	1

Figure 2: Numerical coding of genotypes.

SNPs.

In this paper, we propose a new algorithm that finds a minimum set of Tag SNPs using a feature selection method based on the solution of an integer program that is a particular type of set covering problem, and then test the quality of the selected Tags using both a majority voting rule and a more evolved learning strategy.

The paper is organized as follows. In Section 2, the problem of Tag SNPs Selection (TSS) is defined and related state-of-the-art is annotated, focusing in particular on the two papers [15, 18]. In Section 3, our method for feature selection is presented and specific aspects that arise when it is applied to TSS problems are discussed. Then, in Section 4 the reconstruction techniques used in the paper are described. In Section 5 data sets are described, evaluation criteria are introduced, and experimental results are presented. Conclusions and further work are discussed in the last section.

2. Problem Formulation

Let us introduce the concepts of genotype and haplotype. In our scenario a *genotype* is a combination of alleles located in homologous chromosomes; in case of diploid organisms, characterized by two chromosomes (the maternal one and the paternal one), the genotype is a sequence of pairs of alleles of certain SNPs of the DNA sequence. If the two alleles are identical, the SNP is called homozygote, while if the two alleles are different, it is called heterozygote. In case of heterozygote SNPs, *the phase* (i.e. the value of the SNP associated to maternal and paternal chromosomes) is not known. When the phase is given, then the genotype is split in two sequences, called *haplotypes* (see Figure 1). While typing genotypes is not too costly, the process to read the haplotypes is very expensive and mathematical and statistical methods have been proposed to derive haplotypes from the genotypes of a population or for a single individual. Tag SNPs selection and reconstruction methods use information about haplotypes and/or genotypes; without loss of generality we will illustrate our approach with reference to genotypes but we test it both on haplotype and genotype data sets.

Now we introduce some notation. Let $G = \{g_1, ..., g_n\}$ be the set of input genotypes, where each of the n elements is represented by an m-dimensional vector (m is the number of SNPs). Formally, for each $h \in \{1, ..., m\}$ and for each $p \in \{1, ..., n\}$, $g_p^h \in \{0, 1, 2, 3\}$, where 0 denotes a not-defined value, 1 denotes a homozygous site whose allele is the most frequent one, 2 denotes a homozygous site whose allele is the less frequent one, and 3 denotes a heterozygous site. In Figure 2 we show the numerical coding procedure, by adding as the last row of the table the numerical coding for genotype g_3. Given the genotype matrix G, the aim of Tag SNPs selection can be declared as follows: define a partition of the SNPs set in two sets, TAG and $non - TAG$, and a reconstruction function that is able to derive for each individual the value of the alleles in a SNP of the $non - TAG$ set from the value of the alleles in the SNPs of the TAG set. To better qualify the objective of Tag SNPs selection, we introduce the notion of *prediction error* associated with a pair of TAG and $non - TAG$ sets of a given set of individuals. The *prediction error* is the proportion of the number of alleles that are wrongly reconstructed on the total number of alleles in

SNPs of the $non-TAG$ set.

Clearly, once defined the prediction error, the objective of Tag SNPs selection is to find a partition of the SNPs set and a reconstruction function for which the prediction error is minimized. This point of view sets the stage for a typical learning problem, where the available data is split into a training set and a testing set. The training set is used to support the search of the partition of the SNPs set and of the reconstruction function, while the testing set is used to compute the *prediction error*, which is a measure of the generalization capabilities of the method.

In this framework, we bring to evidence the strong connection between Tag SNPs selection and state-of-the-art methods developed in the Data Mining context to project data from high dimensional spaces into smaller ones with maximal information retention. Those methods usually fall under the definition of Feature Selection, and typically use integer programming formulations of the Feature Selection problem.

Our believe is that such class of methods can be effective to determine Tag SNPs also from very large initial sets, and that the problem of Tag SNPs reconstruction can be considered as a standard supervised learning problem. The use of Feature Selection methods and the possible reconstruction functions to it related are the topic of the next two sections, respectively.

3. Feature Selection for Biological Data Analysis

In the analysis of large multidimensional data, it is required to synthesize the available information in order to discover the relevant phenomena hidden in the data. Such a task has always been one of the main topic of statistics and data analysis. The complexity of such task increases quickly with the number of dimensions considered. In the case of Biological Data Analysis, or, more specifically, genetic data, the relevance of the dimensionality issue cannot be underestimated: typically, the information collected is in the form of a limited number of observations (individuals), for each of which an extremely large number of variables have been measured, associated with the expression of the genes or with the state of haplotypes.

In such cases, the task of identifying a reduced number of variables (e.g., genes, haplotypes) that provide the relevant information to bring to evidence some fact about the data has a twofold function: on the one hand, the dimension needs to be reduced in order to be analyzed by some method (be it visual inspection or automatic model building in any level of detail); on the other hand, the phenomena that are of interest for research purposes are by nature based on a limited amount of information (possibly combined in a complex way).

Several scientific contributions have been proposed to extract the relevant features from a possibly very large set, and many methods are available. Most of them are either *filter methods* or *wrapper methods* [4].

The formers are based on the definition of a function that evaluates the features independently from the use that will be done of them in the following steps, e.g. if they will be used in a regression model, or a decision tree, for a supervised or a unsupervised learning algorithm. The *wrapper methods*, on the other hand, encapsulate the feature selection step in the general data discovery process, combining it with the final analysis method that uses the selected features. In this framework, each candidate subset is tested using the final analysis algorithm and then evaluated on the basis of its performances.

It is widely acknowledged that wrapper methods can provide better results in term of final accuracy; nevertheless, they are extremely demanding from the computational point of view and must be considered unpractical for the analysis of large data sets, as it is the case for the problems considered in this paper. Our attention is thus pointed toward methods that fall in the filter category, where the available features are evaluated on the basis of some measure of their relevance. Here, an important issue to settle is why a certain variable/feature has to be considered relevant.

In the most general setting, the relevance of a variable is directly connected with its variability, i.e. with a measure of its variation among all the observed individuals. In such direction, many methods consider the *entropy* [17, 21], the standard deviation, or other similar measures to assess the importance of a variable. Such general standpoint is appropriate for non supervised learning, when the objective is to synthesize and analyze the data without a particular class to be learned. Instead, when supervised learning is the objective of the analysis, one wants to identify those features that contribute to distinguish individuals of one class from individuals of another classes. In this framework, we often consider class-dependent variability measures, that evaluate a feature only on the amount of its variation between individuals of different classes (e.g., the so-called *class entropy*). This would lead to a measure of relevance that is somehow correlated with the class variable.

Nevertheless, the relevance of a feature is to be assessed also in relation with the other features that compose the selected set. In particular, if two features exhibit a high degree of correlation, then their common presence in the final set can be considered redundant. Such consideration is used in many constructive feature selection methods that determine the final feature set by the iterative addition of a feature, considering its relevance given the features already selected (similarly, the same iterative approach can be adopted to remove a feature from an initially complete set; such an approach would although be impractical for very large feature sets).

3.1. Optimization Model for Feature Selection

As already pointed out, the use of a synthetic measure of the relevance of a feature is a key ingredient in many feature selection methods. Such synthetic approach has the clear merit of being simple, intuitive, and computationally efficient, but it may fail to catch some peculiarity in the structure of the data that could be of some importance in the learning process. To overcome this drawback, some authors have proposed feature selection methods that model explicitly the contribution of a variable over all data and try to optimize the overall contribution of a set of features ([2, 3]). The method adopted in this paper falls into this category (similar applications can be found also in [1]). It is based on the very simple observation that a good set of features must provide information that is able to differentiate a large number of individuals in the data set; in other words, we want the features to maintain the largest amount of variability with respect to all the pairs of individuals that must be differentiated in the specific application. For non supervised learning, the objective is to maintain the features that, together, differentiate all pairs of individuals; for supervised learning, it is sufficient to require that all pairs of individuals belonging to different classes have different measures for one or more features.

For the sake of ease of handling, we restrict the description of the method only to the case of unsupervised learning. Given a set of n individuals described by a set of m features, the simplest approach would require that, for each pair of different individuals, at least one feature among the selected ones has a different value in the two individuals of the pair. We thus define a binary variable $x_h = \{0, 1\}$ associated with each feature ($h = 1, ..., m$) whose value is 1 if the feature is chosen, and 0 otherwise. Moreover, we define the coefficients a_{ijh} equal to 1 when individuals i and j differ on feature h, and 0 otherwise. The problem is then described by means of the following integer programming formulation:

$$\begin{aligned} \min \quad & \sum_{h=1}^{m} x_h \\ & \sum_{h=1}^{m} a_{ijh} x_h \geq 1 \quad && i = 1 \dots n, \; j = 1 \dots n, \; i \neq j \\ & x_h \in \{0, 1\} && h = 1 \dots m, \end{aligned} \tag{1}$$

which is referred to as *Combinatorial Feature Selection* or *Minimal Test Collection.* Clearly, the problem could be solved by using any general purpose integer programming software, but only in case of small size instances. To solve large scale instances, since the mathematical formulation describes a straight forward Set Covering Problem (see [11]), devoted algorithms and heuristics may be used. In formulation (1), the objective function amounts to the cardinality of the selected feature set; the model thus finds the smallest set that enables to satisfy the pairwise differences at a minimal level. The differentiating information is thus considered a constraint, while the number of features selected has to be minimized. We extend the above model by reversing the roles of the constraints and of the objective function, based on the observation that in many learning problems the number of features that can be managed in the analysis step is somehow fixed in advance by the method used. Therefore, we question what are the k features that are able to provide the maximal amount of difference between all pairs of individuals. The natural formulation of this problem would be the following (see also [1]):

$$
\begin{array}{lll}
\max & \alpha & \\
 & \sum_{h=1}^{m} x_h \leq k & \\
 & \sum_{h=1}^{m} a_{ijh} x_h \geq \alpha & i = 1 \dots n,\ j = 1 \dots n,\ i \neq j \\
 & x_h \in \{0, 1\} & h = 1 \dots m,
\end{array}
\tag{2}
$$

where the optimal value of α represents the number of selected features on which all pairs, at least, are different. The value of α can thus be put in direct relation with the power of the information retained by the selected features, as it tries to distribute the discriminatory effort of the k features among all pairs of individuals. The advantage of such model is clear. In fact, while the simple minimization of the number of features for a fixed rhs of the constrains as in (1) may select, among two equivalent solutions, one that retains a smaller amount of discriminatory information, model (2) would always provide, among solutions of fixed dimension k, one with maximal discriminatory power.

The adaptation of models (1) and (2) to the case of supervised learning is obtained by a simple removal from the model of all the constraints that are associated with a pair of individuals of the same class.

Among the various *filter methods* available in the literature, those based on combinatorial feature selection are very difficult to solve. Moreover, while (1) amounts to a standard set covering model, model (2) has a more unusual structure and its formulation can be considered quite weak in the polyhedral sense. In the specific application dealt with in this paper, the large number of columns pushes up the number of variables, while the contained number of individuals limits the number of rows, that grows quadratically with n. In any case, since for realistic size problem instances the search for an optimal solution is impractical, we have considered appropriate to design an *ad hoc* heuristic method that is able to find solutions of good quality.

3.2. Solving the Feature Selection problem with a GRASP Heuristic

To solve the problem we propose a Greedy Randomized Adaptive Search Procedure (GRASP). GRASP is a randomized multistart iterative metaheuristic originally proposed by Feo and Resende [7, 8]. For a comprehensive study of GRASP strategies and variants, the reader is referred to the survey chapter by Resende and Ribeiro [20], and to the annotated bibliography by Festa and Resende [9] for a survey of applications.

GRASP consists basically of two phases: a construction phase and a local search phase. The construction phase iteratively builds a feasible solution. Once a feasible solution is obtained, the local search procedure attempts to improve it by producing a locally optimal solution with respect to some neighborhood structure. The construction and the local search phases are repeatedly applied and the best solution found is returned as an approximation of the optimal one. Figure 3 depicts the pseudo-code of a generic GRASP heuristic for a minimization problem. The construction phase makes use of an adaptive greedy function, a construction mechanism for the restricted candidate list, and a probabilistic selection criterion. Starting from an empty solution, at each iteration, an element is randomly selected from a *restricted candidate list* (RCL), whose elements are among the best ordered, according to some greedy function that measures the (myopic) benefit of selecting each element. In any GRASP heuristic, the construction procedure is *adaptive*, because the benefits associated with every element are updated at each iteration of the construction phase to reflect the changes brought on by the selection of the previous element.

There are several different mechanisms to build the RCL. Typically, it can be limited by the number of elements (cardinality-based criterion) or by their quality (value-based criterion). If the cardinality-based criterion is chosen, then the cardinality of RCL is a priori fixed to some p and the RCL is made up of those elements having the p best greedy function values, while in the value-based case, the cardinality of RCL depends on a threshold parameter $0 \leq \gamma \leq 1$.

```
procedure GRASP(MaxIterations)
    for $i = 1, \ldots,$ MaxIterations do
        Build a greedy randomized solution $x$;
        $x \leftarrow$ LocalSearch($x$);
        if $i = 1$ then $x^* \leftarrow x$;
        else if $w(x) > w(x^*)$ then $x^* \leftarrow x$;
    end;
    return $(x^*)$;
end GRASP;
```

Figure 3: Pseudo-code of a generic GRASP for a minimization problem.

For the problem we are willing to solve, the selection involves candidate columns and it is intuitive to relate the greedy function to the number of rows still to be fully covered that a column not yet chosen would cover if selected. More formally, at a generic iteration of the GRASP construction phase let $\overline{C} \subset C$ be the subset of columns already selected as partial solution and let R be the set of constraints. Moreover, the following quantities are computed:

a. for each row $r \in R$, $coverage(r)$ denotes the number of 1's in r covered by $\overline{C}$;

b. for each column $c \in C$, $maxmin\alpha(c)$ denotes the number of rows covered by c that are covered by the minimum number of columns (lowest coverage);

c. $M = \max\limits_{c \in C \setminus \overline{C}} maxmin\alpha(c)$;

d. $min\alpha$ denotes the lowest coverage;

e. $max\alpha$ denotes the highest coverage.

Then, for each $c \in C \setminus \overline{C}$ we compute

$$score(c) = \sum_{r \in R_c} [max\alpha - coverage(r)],$$

where R_c is the set of rows covered by c. It is clear that the greedy function $score(\cdot)$ measures how much additional cover will result from the selection of an unselected column c and a pure greedy choice would consist in selecting the column $c \in C \setminus \overline{C}$ with the highest greedy function value. To define the construction mechanism for the restricted candidate list RCL, let

$$\sigma_{min} \quad = \quad \min\{score(c) \mid c \in C \setminus \overline{C}\}$$

and

$$\sigma^{max} \quad = \quad \max\{score(c) \mid c \in C \setminus \overline{C}\}.$$

Denoting by $\mu = \sigma_{min} + \gamma \cdot (\sigma^{max} - \sigma_{min})$ a cut-off value, where γ is a parameter such that $0 \leq \gamma \leq 1$, the restricted candidate list RCL is made up by all columns whose greedy function value is greater than or equal to μ and all columns that cover the maximum number of rows with the lowest coverage, i.e.

$$RCL = \{c \in C \setminus \overline{C} \mid score(c) \geq \mu \text{ or } maxmin\alpha(c) = M\}.$$

Once randomly selected a column c from the RCL and added to the partial solution $\overline{C}$, quantities a.–e. are updated accordingly to the just made selection (adaptive component). Note that the extreme case $\gamma = 0$ corresponds to a pure random strategy, while the extreme case $\gamma = 1$ is equivalent to a completely greedy strategy.

The construction phase iteratively adds one element at a time to the set $\overline{C}$ that ends up with a representation of a complete feasible solution. Once a feasible solution $\overline{C}$ is obtained, a local search procedure attempts to improve it by producing a locally optimal solution with respect to some suitably defined neighborhood structure.

Let $\overline{C}$ be the feasible solution output of the construction procedure and let $\alpha_{\overline{C}}$ and $\beta_{\overline{C}}$ be respectively the value of the variable α corresponding to solution $\overline{C}$ and the number of features that in $\overline{C}$ are able to provide the maximal amount of difference between all pairs of individuals. Then, to realize the local search phase, we have designed a procedure that starting from $\overline{C}$ performs the following steps trying to find a better quality solution, i.e. a new set of columns $\hat{C}$ with lower cardinality (removal of redundant columns) and/or corresponding to a higher coverage:

1. Compute
$$\hat{C} = \overline{C} \setminus \{c \in \overline{C} \mid \mathit{score}(c) = \sigma^{max} \text{ and } |R_c^{\alpha_{\overline{C}}}| \neq \emptyset\},$$
where $R_c^{\alpha_{\overline{C}}} \subseteq R_c$ is the set of rows covered by c such that $\mathit{coverage}(r) = \alpha_{\overline{C}}$.

2. If $\beta_{\hat{C}} < k$, compute
$$\mathcal{A} = \max\{|R_c^{\alpha_{\hat{C}}}| \text{ such that } c \notin \hat{C},\ \mathit{score}(c) = \sigma_{min}\}.$$
and let be $j = \text{argmax } \mathcal{A}$.

   If $\max\{|R_c^{\alpha_{\hat{C}}}| \text{ s.t. } c \notin \hat{C},\ \mathit{score}(c) = \sigma_{min}\} \geq 1$, then set $\hat{C} = \hat{C} \cup \{j\}$.

3. Swap operation:

   if there exist

   ◇ $c \in \hat{C}$ such that $\mathit{score}(c) = \sigma^{max}$ and $|R_c^{\alpha_{\hat{C}}}| = 0$;

   ◇ $j \in C \setminus \hat{C}$ such that $\mathit{score}(c) = \sigma^{min}$ and $|R_c^{\alpha_{\hat{C}}}| > 0$

   then, set $\hat{C} = \hat{C} \setminus \{c\} \cup \{j\}$.

The local search procedure stops when no improving solution can be found and returns as output the current local optimum set of columns.

4. Reconstructing the SNPs from the Tag set

We have used two different reconstruction methods for testing the validity of our approach to Tag SNPs selection problem. The first one is simply based on a majority vote, while the second one is a more sophisticated method based on learning algorithms for logic data.

4.1. The Majority Vote Method

The majority vote procedure adopted here is inspired to that proposed by Shamir in [15]. It is based on the assumption that, for a given individual (or genotype) g_i, the allelic value of a non-Tag SNPs can be inferred from the allelic values for that SNP of the individuals that are similar to g_i over the Tag SNPs. We provide below a sketch of the method.

Given a training set T of genotypes and a set of Tag SNPs, let g_i be a genotype from the test set for which we know only the values of the Tag SNPs, we reconstruct the values of the remaining SNPs in the following way:

- in set T we identify the subset T' of individuals that are equal to the genotype g_i over the Tag SNPs. If such set is empty, we enlarge it with the individuals that are at minimum hamming distance from g_i over the Tag SNPs.

- for each non-Tag SNPs of g_i, we predict its allelic value as the most frequent allelic value for that SNPs in the genotypes of T'.

The general principle of the method is that genotypes that are similar on the Tag SNPs are also likely to be similar on the remaining SNPs. Additional details on the method can be found in [15]; in that reference the author resticts the identification of set T' by the matching of only the two Tags that surround the SNPs whose allelic value must be predicted. In addition, the method is applied to haplotypes data and thus it combines two majority votes over the set of haplotypes associated to the genotypes of the training set in order to derive the values of the unknown SNPs.

4.2. The Extraction of Separating Logic Formulas

Lsquare is a learning method that operates on data represented by logic variables and produces rules in propositional logic that classify the items in one of two classes. The suitability of *Lsquare* for the reconstruction task is motivated by the observation that it uses a logic representation of the description variables, that are to all extents logic variables, and of the classification rules, that are logic formulas in Disjunctive Normal From (DNF). Such property enables to analyze and interpret the classification results also from the semantic point of view, as the reconstruction rules determined by the method express combination of the features that can be interpreted by domain experts and bring to light new knowledge in an easily understandable format. Moreover, the numerical coding of the allelic values of the SNPs is correctly represented with the *true* and *false* values of logic variables.

The classification rules are determined using a particular problem formulation that amounts to be a well know and hard combinatorial optimization problem, the *minimum cost satisfiability problem*, or MINSAT, that is solved using a very sophisticated solver based on decomposition and learning techniques [22]. The DNF formulas identified have the property of being created by conjunctive clauses that are searched for in order of coverage of the training set. Therefore, they usually are formed by few clauses with large coverage (the interpretation of the trends present in the data) and, if needed, additional clauses with smaller coverage (the interpretation of the outliers in the training set).

The system and its additional components have been presented and described in related papers ([5, 6]) and its detailed description is out of the scope of this paper.

5. Data sets and computational results

The approach proposed in this paper has been tested on different settings. Several data sets on Tag SNPs are available in the scientific community, and some of them have been considered appropriate to verify the validity of the methods here described. In particular, we have considered two types of experiments: the selection of Tag SNPs from a very large dataset of haplotypes from the HAPMAP database (http://www.hapmap.org/), and the selection of Tag SNPs from smaller datasets of genotypes available in the Yale Database ALFRED (http://alfred.med.yale.edu). In the first set of experiments, we tested the capability of the Feature Selection approach to select the Tag SNPs combined with the *majority vote* reconstruction strategy discussed in Section 4.1; in the second set, we consider the Tag SNPs selection and reconstruction problem as a sequence of supervised learning problems, and apply the Logic Miner *Lsquare* to predict the value of the alleles that are to be reconstructed.

5.1. The HAPMAP Data sets on Chromosome 21

From the data collected within the International HapMap Project we have selected samples from 4 different populations, for which the haplotype sequences of the Chromosome 21 are made available. Each population is described by a large number of SNPs and a comparably small number of individuals, as reported in Table 1, where CEU stands for Utah Residents with Northern and Western European Ancestry, YRI for Yoruba in Ibadan, Nigeria and CHB+JPT for Japanese in Tokyo, Japan and Han Chinese in Beijing, China.

For the three datasets we have formulated the Feature Selection problem described in Section 3.1, and have then solved it with the heuristics described in Section 3.2, as the dimensions of the problem were

Table 1: Description of HAPMAP Populations for Chromosome 21.

Code	Individuals	SNPs haplotypes	SNPs
CEU	60	120	34.103
YRI	60	120	38.852
CHB+ JPT	80	180	33.878

Table 2: Tag Identification and Reconstruction for the 3 HAPMAP Populations.

Population	Total SNPs	Tag SNPs	α value	Error rate
CEU	34.103	20	4	26.88
YRI	38.852	20	2	24.92
CHB+JPT	33.878	20	4	26.16
CEU	1500	20	4	20.24
YRI	1500	20	2	18.75
CHB+JPT	1500	20	4	16.18

prohibitive for the exact solution even with state of the art commercial software. CPLEX ILOG has been also used to implement a column-generation based fixed heuristics that was nevertheless dominated by the GRASP method when solution qualities and computing times were considered. The number of SNPs selected ranged from 10 to 20, for which the optimal α value of problem (2) ranged between 2 and 4. The precision of the method clearly improves with the number of selected SNPs but after the value of 20 we do not record a significant increase.

The selected SNPs were then used for the reconstruction of the remaining SNPs using the majority vote adapted from the procedure described in Section 4.1. In addition to the case where the Tag SNPs are selected from the complete set of SNPs, we have also run some experiments where the initial SNPs set was composed on only the first 1500 SNPs. Such exercise was done in order to compare our method with other results presented in the literature on the same data, that applied the same type of reduction.

Table 2 summarized some of the results obtained. We note a decrease in the recognition performances when increasing the number of SNPs from 1.500 to its full dimensions (ranging from 33.878 to 38.852 for the three populations), but the results for 1.500 compete with the previous results; on the other hand, other methods do not seem to have the capability of dealing simultaneously with such a large number of SNPs as the one proposed here.

The results obtained show that the proposed method is able to achieve correct recognition rates in line with what is considered a good result in similar literature. For example, in Shamir, larger proportions of Tag SNPs provide $80\% - 90\%$ correct recognition rates with a *leave-1-out* training-testing scheme.

5.2. The YALE Data set

In this second set of experiments we have extracted some datasets from a database of populations collected at the University of Yale. An extensive exploration on such data has been presented, among others, in [18]. The individuals of the considered populations are described by a smaller number of SNPs with respect to the the HapMap data described in the previous section, and we use them to test the combination of the combinatorial feature selection approach of Section 3.1 and the logic learning method *Lsquare* (Section 4.2).

The experiments presented consider three populations of the $17q25$ genotype, whose dimensions are summarized in Table 3.

The first operation performed on the data is the extraction of the Tag SNPs; using the feature selection model (2) we test the values of 10 and 20 Tag SNPs - k in formulation (2); such choice is also driven by the opportunity of comparing our results with those described in [18].

Once 10 or 20 SNPs have been extracted, we consider, for each non-tag SNPs, the following recognition task: first, we adopt a $70\% - 30\%$ split for training and testing data; second, using *Lsquare*, we extract from the training set a logic formula that predicts the ternary allelic value of the SNPs with the value of

Table 3: Dimensions of the 17q25 Yale Populations.

Population	Number of SNPs	Number of Individuals
Kar	63	49
Mxp	63	51
Ron	63	44
Sam	63	40

Table 4: Reconstruction error for Yale 17q25 for subset of SNPs on 30% test set.

Population	Tag SNPs	*lsq SNPs*	*maj SNP*	*lsq err*	*maj err*	*tot err*
Kar	10	3	50	31%	25%	26%
Mxp	10	4	49	33%	49%	39%
Ron	10	0	53	n.a.	16%	16%
Sam	10	5	48	25%	38%	36%
Kar	20	2	41	9%	15%	14%
Mxp	20	2	41	28%	32%	30%
Ron	20	1	42	6%	14%	13%
Sam	10	2	41	8%	36%	35%

the Tag SNPs. The formula is then used on the testing data and the correct recognition rate is considered to assess the quality of the method.

The only restriction that applies in this approach is related with the fact that some of the non-Tag SNPs may present an insufficient variability to state the related learning problem; e.g., if one column of the data matrix presents the same allelic value on a large portion of the rows, then we are not in possess of the appropriate separation of the individuals in *positive* and *negative* examples to define the supervised learning problem to be solved with *Lsquare*. For this reason, the reconstruction with a learned logic function is limited to those SNPs for which the allelic value is sufficiently differentiated; for the others, we use the same *majority vote* rule adapted to the case of genotype data.

The performances of the method are summarized in Table 4, where we report the number of Tag SNPs on which the learning problem can be formulated, the number of SNPs that are reconstructed with the logic rule learned by *Lsquare* (*lsq SNPs*), the number of remaining SNPs that are reconstructed with the majority vote (*maj SNPs*), and the related error rates (*lsq err* and *maj err*) followed by the total reconstruction error rate in the last column.

The results obtained on the Yale datasets are useful to draw some additional consideration on the proposed method. First of all, we note that the error rates obtained are for some populations similar and for other worst to those presented in [18] for the same number of Tag SNPs and for the same training/testing split of the data. On the other hand, it clearly emerges that, when the SNPs to be predicted are sufficiently varied and thus the logic learning method can be applied (as opposed to the majority rule), then the performances are indeed very good when restricted to this - unfortunately small - subset of the SNPs (the only exception being the `Kar` dataset with 10 tags. In this sense, we may conclude that the very intuitive majority rule adopted is not indicated for datasets with limited number of individuals as the Yale ones. In addition, we see that a full power learning approach can lead to better results, although its applicability is also threatened by the limited amount of individuals that do not express sufficient variability to formulate the learning problem.

The latter reconstruction method has the additional feature of providing a detailed specification of the relations among the reconstructed SNP and the Tag SNPs, in the form of logic formulas in Disjunctive Normal Form. For example, we report below one of the formulas determined by our method for the `Kar` population to predict the allelic value (coded as 1, 2 and 3) of SNP in position 24 from the allelic values of the 10 selected Tag SNPs.

Such information may result in important knowledge and lead to additional insight on the nature of the data analyzed.

Table 5: Logic Formulas for SNP 24 in Kar Population.

IF	TAG(25)=2	OR	TAG(25)=3	THEN	SNP(24)=1
IF	TAG(25)=1	AND	TAG(47)=3	THEN	SNP(24)=2
IF	TAG(25)=1	OR	TAG(11)=2 AND TAG(25)=2	THEN	SNP(24)=3

6. Conclusions

In this paper, a feature selection based technique has been proposed to solve the problem of Tag SNPs selection, i.e. the problem of finding, given the genotypes of a set of individuals, a subsets of SNPs that are strongly related to all the other SNPs. The feature selection problem is formulated as a variant of the set covering problem, where the target function is to find a set of k SNPs that cover at least α constraints and are able to distinguish each individual from the others. To test the approach the reconstruction problem is also tackled, i.e. the problem of deriving unknown SNPs from Tag SNPs. Two reconstruction functions are proposed, one based on a majority vote and the other based on a machine learning technique. The algorithm has been tested on two families of data sets of different nature, providing results that are, when comparable, in line with the related literature. One of the interesting aspects of the proposed method is to be found in its capability to deal simultaneously with very large SNPs sets, and, in addition, to provide highly informative reconstruction rules in the form of logic formulas.

Additional experiments on more data sets and proper refinements of the reconstruction method are the object of further work in this research activity.

References

[1] Bertolazzi P., Felici G., Festa P., Lancia G., *Logic classification and feature selection for biomedical data*, Computer and Mathematics with Applications, vol 55(5) 2008, 889-899

[2] Boros E., Ibaraki T., and Makino K., Logical analysis of binary data with missing bits, *Artificial Intelligence* 1999; 107:219-263.

[3] Boros E., Ibaraki T., Kogan A., and Mayoraz E. adn Muchnik I., An implementation of logical analysis of data, *RUTCOR Research Report, 29-96, Rutgers University, NJ.* 1996.

[4] Felici G., de Angelis V., Mancinelli G., Feature Selection for Data Mining. In Data Mining and Knowledge Discovery Approaches Based on Rule Induction Techniques, G. Felici and E. Triantaphyllou eds.,Springer 2006; 227-252.

[5] Felici G. and Truemper K. A minsat approach for learning in logic domains. INFORMS Journal on Computing 2001; 13(3):1-17.

[6] Felici G. and Truemper K. The Lsquare System for Mining Logic Data. Encyclopedia of Data Warehousing and Mining, (J. Wang ed.), vol. 2, Idea Group Inc. 2006; 693-697.

[7] Feo T.A. and Resende M.G.C. A probabilistic heuristic for a computationally difficult set covering problem. Operations Research Letters 1989; 8:67-71.

[8] Feo T.A. and Resende M.G.C. Greedy randomized adaptive search procedures. J. of Global Optimization 1995; 6:109-133.

[9] Festa P. and Resende M.G.C. Grasp: An annotated bibliography. In Ribeiro C.C. and Hansen P., editors, *Essays and Surveys on Metaheuristics*, Kluwer Academic Publishers 2002; 325-367.

[10] Gabriel S.B., Schaffner S.F., Nguyen H., Moore J.M., Roy J., Blumenstiel B., Higgins J., DeFelice M., Lochner A., Faggart M., Liu-Cordero S.N., Rotimi C., Adeyemo A., Cooper R., Ward R., Lander E.S., Daly M.J., Altshuler D. The structure of haplotype blocks in the human genome.. Science 2002; 296:2225-2229.

[11] Garey M.R. and Johnson D.S. Computer and Intractability: a guide to the theory of NP-completeness. Freeman, San Francisco, 1979.

[12] Goldstein D.B., Weale M. E. Population genomics: Linkage disequilibrium holds the key. Curr. Biol. 2001; 11:R576-R579.

[13] Halldrsson B.V., Bafna V., Lippert R., Schwartz R., De La Vega F.M., Clark A.G. and Istrail S. Optimal Haplotype Block-Free Selection of Tagging SNPs for Genome-Wide Association Studies Genome Res. 2004; 14:1633-1640

[14] Halldrsson B.V., Istrail S., De La Vega F.M. Optimal Selection of SNP Markers for Disease Association Studies Hum. Hered 2004; 58:190-202.

[15] Halperin E., Kimme G., and Shamir R. Tag SNP selection in genotype data for maximizing SNP prediction accuracy Bioinformatics 2005; 21:195-203.

[16] Kwok Pui-Yan(ed.), Single Nucleotide Polymorphism: Methods and Protocols. Methods in Molecular Biology. Human Press Inc, Totowa, New Jersey, 2003.

[17] Liu H. and Motoda H. *Feature Selection for knowledge discovery and data mining.* Kluwer Academic Publishers, 2000.

[18] Paschou P., Mahoney M.W., Javed A., KiddJ.R., Pakstis A.J., Gu S., Kidd K.K., and Drineas P. Intra- and interpopulation genotype reconstruction from tagging SNPs Genome Res. 2007; 17:96-107

[19] Patil, N., Blocks of Limited Haplotype Diversity Revealed by High-Resolution Scanning of Human Chromosome 21. Science 2001; 294:1719-1723.

[20] Resende M.G.C. and Ribeiro C.C. Greedy randomized adaptive search procedures. In Glover F. and Kochenberger G., editors, *State-of-the-Art Handbook of Metaheuristics*. Kluwer, 2002, 219-249.

[21] Shannon C.E., A matematical theory of comunication, *Bell System Technical Journal* 1948; 27 379-423, 623-656.

[22] Truemper K. Design of Logic-Based Intelligent Systems. Wiley-Interscience, New York, 2004.

[23] Zhang K., Qin Z. S., Liu J. S., Chen T., Waterman M. S. and Sun F. Haplotype Block Partitioning and Tag SNP Selection Using Genotype Data and Their Applications to Association Studies. Genome Res. 2004;14:908-916.

[24] Zhang K., Qin Z. S., Liu J. S., Chen T., Waterman M. S. and Sun F. HapBlock: haplotype block partitioning and Tag SNP selection software using a set of dynamic programming algorithms. Bioinformatics 2005; 21:131-134.