Evolutionary Data Mining for Genomics

Laetitia Jourdan
LIFL, University of Lille 1, France

Clarisse Dhaenens
LIFL, University of Lille 1, France

El-Ghazali Talbi
LIFL, University of Lille 1, France

INTRODUCTION

Knowledge discovery from genomic data has become an important research area for biologists. Nowadays, a lot of data is available on the Web, but it is wrong to say that corresponding knowledge is also available. For example, the first draft of the human genome, which contains 3,000,000,000 letters, was achieved in June 2000, but, up to now, only a small part of the hidden knowledge has been discovered. This is the aim of bioinformatics, which brings together biology, computer science, mathematics, statistics, and information theory to analyze biological data for interpretation and prediction. Hence, many problems encountered while studying genomic data may be modeled as data mining tasks, such as feature selection, classification, clustering, or association rule discovery.

An important characteristic of genomic applications is the large amount of data to analyze, and, most of the time, it is not possible to enumerate all the possibilities. Therefore, we propose to model these knowledge discovery tasks as combinatorial optimization tasks in order to apply efficient optimization algorithms to extract knowledge from large datasets. To design an efficient optimization algorithm, several aspects have to be considered. The main one is the choice of the type of resolution method according to the characteristics of the problem. Is it an easy problem, for which a polynomial algorithm may be found? If the answer is *yes*, then let us design such an algorithm. Unfortunately, most of the time, the response to the question is *no*, and only heuristics that may find good but not necessarily optimal solutions can be used. In our approach, we focus on evolutionary computation, which has already shown an interesting ability to solve highly complex combinatorial problems.

In this article, we will show the efficacy of such an approach while describing the main steps required to solve data mining problems from genomics with evolutionary algorithms. We will illustrate these steps with a real problem.

BACKGROUND

Evolutionary data mining for genomics groups three important fields: evolutionary computation, knowledge discovery, and genomics.

It is now well known that evolutionary algorithms are well suited for some data mining tasks (Freitas, 2002). Here, we want to show the interest of dealing with genomic data, thanks to evolutionary approaches. A first proof of this interest may be the recent book by Gary Fogel and David Corne, *Evolutionary Computation in Bioinformatics*, which groups several applications of evolutionary computation to problems in the biological sciences and, in particular, in bioinformatics (Fogel & Corne, 2002). In this article, several data mining tasks are addressed, such as feature selection or clustering, and solved, thanks to evolutionary approaches.

Another proof of the interest of such approaches is the number of sessions around evolutionary computation in bioinformatics and computational biology that have been organized during the last Congress on Evolutionary Computation (CEC) in Portland, Oregon in 2004.

The aim of genomic studies is to understand the function of genes, to determine which genes are involved in a given process, and how genes are related. Hence, experiments are conducted, for example, to localize coding regions in DNA sequences and/or to evaluate the expression level of genes in certain conditions. Resulting from this, data available for the bioinformatics researcher may deal with DNA sequence information that are related to other types of data. The example used to illustrate this article may be classified in this category.

Another type of data deals with the recent technology called *microarray*, which allows the simultaneous measurement of the expression level of thousands of genes under different conditions (i.e., various time points of a process, absorption of different drugs, etc.). This new type of data requires specific data mining tasks, as the number of genes to study is very large and the number of conditions may be limited. Classical questions are the classification or the clustering of genes based on their expression pattern, and commonly used approaches may vary from statistical approaches (Yeung & Ruzzo, 2001) to evolutionary approaches (Merz, 2002) and may use additional biological information, such as gene ontology (GO) (Speer, Spieth & Zell, 2004). Recently, the biclustering that allows the grouping of instances having similar characteristic for a subset of attributes (here, genes having the same expression patterns for a subset of conditions) has been applied to this type of data and evolutionary approaches proposed (Bleuler, Prelié & Ziztler, 2004). In this context of microarray data analysis, association rule discovery also has been realized using evolutionary algorithms (Khabzaoui, Dhaenens & Talbi, 2004).

MAIN THRUST

In order to extract knowledge from genomic data using evolutionary algorithms, several steps have to be considered:

1. Identification of the knowledge discovery task from the biological problem under study;
2. Design of this task as an optimization problem;
3. Resolution using an evolutionary approach.

Hence, in this section, we will focus on each of these steps. First, we will present the genomic application that we will use to illustrate the rest of the article and indicate the knowledge discovery tasks that have been extracted. Then, we will show the challenges and some proposed solutions for the two other steps.

Genomics Application

The genomic problem under study is to formulate hypotheses on predisposition factors of different multifactorial diseases, such as diabetes and obesity. In such diseases, one of the difficulties is that sane people can become affected during their life, so only the affected status is relevant. This work has been done in collaboration with the Biology Institute of Lille (IBL, France).

One approach aims to discover the contribution of environmental factors and genetic factors in the pathogenesis of the disease under study by discovering complex interactions, such as ([gene A and gene B] or [gene C and environmental factor D]) in one or more population. The rest of the article will use this problem as an illustration.

To solve such a problem, the first thing is to formulate it into a classical data mining task. The difficulty of such a formulation is to identify the task. This work must be done through discussions and cooperation with biologists in order to agree on the objective of the problems. For example, in our data, identifying groups of people can be modeled as a clustering task, as we cannot take into account non-affected people. Moreover, a lot of loci have to be studied (3,652 points of comparison on the 23 chromosomes and two environmental factors) and classical clustering algorithms are not able to cope with so many points. So, we decided first to execute a feature selection in order to reduce the number of loci in consideration and to extract the most influential features that will be used for the clustering. Hence, the model of this problem is decomposed into two phases: feature selection and clustering.

From a Data Mining Task to an Optimization Problem

The most difficult aspect of turning a data mining task into an optimization problem is to define the criterion to optimize. The choice of the optimization criterion, which measures the quality of candidate knowledge to be extracted, is very important, and the quality of the results of the approach depends on it. Indeed, developing a very efficient method that does not use the right criterion will lead to obtaining the right answer to the wrong question. The optimization criterion either can be specific to the data mining task or dependent of the biological application. Several different choices exist. For example, considering the gene clustering, the optimization criterion can be the minimization of the minimum sum-of-squares (MSS) (Merz, 2002), while for the determination of the members of a predictive gene group, the criterion can be the maximization of the classification success using a maximum likelihood (MLHD) classification method (Ooi & Tan, 2003).

Once the optimization criterion is defined, the second step of the design of the data mining task into an optimization problem is to define the encoding of a solution, which may be independent of the resolution method. For example, for clustering problems in gene expression mining with evolutionary algorithm, Faulkenauer and Marchand (2001) use the specific CGA encoding that is dedicated to grouping problems and is well suited to clustering.

Regarding the genomic application used to illustrate this article, two phases have been isolated. For the feature selection, an optimization approach has been adopted, using an evolutionary algorithm (see next paragraph), whereas a classical approach (k-means) has been chosen for the clustering phase. Determining the optimization criterion for the feature selection was not an easy task, as it was difficult not to favor small sets of features.