ELSEVIER

Knowledge-Based Systems 15 (2002) 235–242

Knowledge-Based SYSTEMS

www.elsevier.com/locate/knosys

A data mining approach to discover genetic and environmental factors involved in multifactorial diseases

L. Jourdan[a,*], C. Dhaenens[a], E.-G. Talbi[a], S. Gallina[b]

[a]*LIFL, Batiment M3, Cité Scientifique, 59655 Villeneuve d'Ascq Cedex, France*
[b]*Biological Institute, Multifactorial Disease Laboratory, 1 rue du Professeur Calmette, B.P. 245, 59019 Lille Cedex, France*

Received 16 March 2001; revised 22 April 2001; accepted 31 May 2001

Abstract

In this paper, we are interested in discovering genetic and environmental factors that are involved in multifactorial diseases. Experiments have been achieved by the Biological Institute of Lille and many data has been generated. To exploit these data, data mining tools are required and we propose a two-phase optimisation approach using a specific genetic algorithm. During the first step, we select significant features with a specific genetic algorithm. Then, during the second step, we cluster affected individuals according to the features selected by the first phase. The paper describes the specificities of the genetic problem that we are studying, and presents in detail the genetic algorithm that we have developed to deal with this very large size feature selection problem. Results on both artificial and real data are presented.

Keywords: Data mining; Clustering; Genetic algorithm; Feature selection; Multifactorial disease

1. Introduction

Common diseases such as Type 2 diabetes, obesity, asthma, hypertension and certain cancers represent a major public health concern (around 160 million people have Type 2 diabetes). These complex diseases have a multifactorial aetiology in which environmental factors (for example: Body Mass Index, which is a measure of obesity, or the age at onset, which is the age of the individual when diabetes was diagnosed), as well as genetic factors (genetic markers) contribute to the pathogenesis of the disease. In order to localise the involved factors, we must be able to detect complex interactions such as [(gene A and gene B) or (gene C and environmental factor D)] in one or more populations. Classical methods of genetic analysis test models with one genetic factor. Some methods have been adapted to test the interaction of a second factor, once a first major factor has been detected. These methods are not intended to test a huge amount of genetics models combining more than two genes. So, they have only a limited capacity to dissect the genetics of complex disease traits.

In order to elucidate the molecular bases of Type 2 diabetes and obesity, the Multifactorial Disease Laboratory at the Biological Institute of Lille had performed large analyses on collections of affected families from different populations. Since the precise molecular mechanisms leading to these diseases are largely unknown, a genome-wide scan strategy was used to localise the genetic factors. This strategy requires no presumptions about interesting loci (a locus is a specific portion of DNA, located without any ambiguity on one chromosome). Seeking for interaction patterns in the huge amount of data generated during this first step should increase the power to detect regions containing genes with no individual major effect, but whose interaction leads to the disease. The biological analysis already performed consists of four steps:

1. Collect families with at least two or three affected members.
2. Extract DNA of parents and offspring from a blood sample.
3. Characterise each DNA sample at 400 loci spread over the 23 chromosomes.
4. Compute the genetic similarities for each pair of relatives at each locus.

Loci are polymorphous, so that parents may have different variants, called alleles. Each individual has two alleles for each locus, one inherited from his father and the other from

[*] Corresponding author. Tel.: +33-320337186.
E-mail addresses: jourdan@lifl.fr (L. Jourdan), talbi@lifl.fr (E.-G. Talbi).

0950-7051/02/$ - see front matter © 2002 Elsevier Science B.V. All rights reserved.
PII: S0950-7051(01)00145-9

his mother. So we can calculate a genetic similarity for a pair of relatives (for example two brothers), based on the number of common alleles they have. For example, at a given locus, two brothers will have the probability: 25% to share no alleles, 50% to share one allele, 25% to share two alleles.

The genetic similarity for each pair is calculated with multi-point methods, in order to have extrapolated values at regular positions between loci. That leads to 3652 values corresponding to 3652 points of comparison on the 23 chromosomes. Then, a comparison is made between the genetic similarity observed and what would be expected regarding the statistical sharing probabilities. This leads to a binary matrix, where a 1 indicates that the observed similarity is greater than the expected probability.

The method must detect a combination of factors, for which a subset of lines share a common pattern of values. However, the method should not converge to a unique solution, but must produce several solutions in order to take into account the heterogeneity of the populations. Moreover, only very few factors (less than 5%) should be relevant.

This is an unsupervised clustering problem, where 3654 features[1] (3652 comparison points and two environmental factors) have to be considered, but where biologists have in mind that only very few features are relevant.

We would like here to point some specifications of the problem regarding usual data mining problems:

- A lot of features have to be considered (up to 3654).
- Very few significant features (less than 5%).
- Only few data (number of pairs of individuals), compared to the large number of features, are available.
- We can only work with affected people because those who are not affected, when extracting their DNA, may become affected.
- The objective is to discover not only a single association, but several associations of genes and environmental factors, and to group affected people according to these associations.

To deal with an unsupervised clustering problem on such a large number of features, it is not possible to apply directly classical clustering algorithms. Algorithms, such as k-means algorithm, dedicated for clustering, are not able to deal with so many features. Their executions would be time consuming and results obtained would not be exploitable. Therefore, we adapt a two phase approach:

- A feature selection phase using a genetic algorithm.
- A clustering phase.

For the first phase, a feature selection, we developed a genetic algorithm to extract most influential features and in particular influential associations of features. This heuristic approach has been chosen as the number of features is large. For this specific problem, some advanced mechanisms have been introduced in the genetic algorithm such as some dedicated genetic operators, the sharing, the random immigrant, and a particular distance operator has been defined. Then, the second phase is a clustering based on the features selected during the previous phase. The clustering algorithm used is k-means.

This paper is organised as follows. First, we give some mathematical background. Then, our approach, an adaptation of a genetic algorithm for this particular feature selection problem, is detailed. Section 4 presents the clustering phase. In Sections 5 and 6, we report results on experiments realised with data from GAW11, a workshop on genetic analysis and with real datasets.

2. Mathematical background

In this section, we give definitions of concepts and functions used in the rest of the paper.

Unsupervised classification (or clustering or learning from observations). This is a data mining task in which the system has to classify a set of objects without any information on the characteristics of classes. It has to find its own classes (clusters).

Feature (or attribute). It is a quantity describing an instance. Here, a feature may be an environmental factor or a genetic comparison point.

Support. The notion of support in data mining denotes the number of times (the number of rows) a set of features is met over the number of times at least one member of the set is met.

Here is a formal definition of the support: let R be a set, r be a binary database over R and $X \subseteq R$ be a set of items. The item set X matches a row $t \in R$ if $X \subseteq t$. The set of rows in r matches by X is denoted by $|\{t \in R/X \subseteq t\}|$ and the $support = |\{t \in r/X \subseteq t\}|/|\{t \in r/(\exists x_i \in X/x_i \subseteq t)\}|$.

Scattering criterion. To measure the quality of clustering we use the classical scattering criterion based on statistical concepts. It is defined as follows:

Given c clusters, n features,
Let χ_j be the j-th cluster with $j = 1, \ldots, c$,
Let m be the mean vector $m = (m_1 \quad m_2 \quad \cdots \quad m_n)$ and m_j the vector of means for the j-th cluster $m_j = (m_{j_1} \quad m_{j_2} \quad \cdots \quad m_{j_n})$.
X_i is an element of the cluster χ_j and $(X_i - m_j)^t$ is the transposition of the vector $(X_i - m_j)$.
The scatter matrix for the j-th cluster is then: $P_j = \sum_{X_i \in \chi_j} (X_i - m_j)(X_i - m_j)^t$. We can define the intra

[1] The term feature will now be used according to the data mining terminology. It is defined in Section 2.