Engineering Applications of Artificial Intelligence 51 (2016) 58–70

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

Complex diseases SNP selection and classification by hybrid Association Rule Mining and Artificial Neural Network—based Evolutionary Algorithms

Aicha Boutorh *, Ahmed Guessoum

Laboratory for Research in Artificial Intelligence, University of Science and Technology Houari Boumediene, Algiers, Algeria

ARTICLE INFO

Available online 15 February 2016

Keywords:
Complex diseases
SNP selection
Association Rule Mining
Artificial Neural Networks
Grammatical Evolution
Genetic Algorithm

ABSTRACT

Recently, various techniques have been applied to classify Single Nucleotide Polymorphisms (SNP) data as they have been shown to be implicated in various human diseases. One of the major problems related to SNP sets is the large p, small n problem which refers to the high number of features and the small number of samples, which makes the classification task complex. In this paper, a new hybrid intelligent technique based on Association Rule Mining (ARM) and Neural Networks (NN) which uses Evolutionary Algorithms (EA) is proposed to deal with the dimensionality problem. On the one hand, ARM optimized by Grammatical Evolution (GE) is used to select the most informative features and to reduce the dimensionality by parallel extraction of associations between SNPs in two separate datasets of case and control samples. On the other hand, and to complement the previous task, a NN is used for efficient classification. The Genetic Algorithm (GA) is used for setting up the parameters of the two combined techniques. The proposed GA-NN-GEARM approach has been applied on four different SNP datasets obtained from the NCBI Gene Expression Omnibus (GEO) website. The created model has reached a high classification accuracy, reaching in some cases 100%, and has outperformed several feature selection techniques when combined with different classifiers.

1. Introduction

In the last two decades, extensive efforts have been made to computationally study the functional and structural consequences of Single Nucleotide Polymorphisms (SNP) (Mooney, 2005). SNP is a DNA sequence variation resulting from an alteration of a single nucleotide in the genome and is an important source of the human genome variability (Collins et al., 1998). Numerous studies have shown that SNPs may have important biological effects, and have been implicated in several human diseases. It is well established in the Genome Wide Association Studies (GWAS) field that SNP profiles characterize a variety of diseases (Sachidanandam et al., 2001). In this context, machine learning and data mining techniques have widely been used to analyse SNP data (Chen et al., 2008; He et al., 2010; Schwender and Ickstadt, 2008).

SNP data is currently used in the development of efficient algorithms for the classification of complex diseases. However, SNP datasets are characterised by their high dimensionality. They all tend to contain high levels of noise and be small in size. These factors make it difficult to develop an efficient classifier (Knudsen, 2011).

Despite the success achieved by standard artificial intelligence techniques in analysing gene expression data, it is becoming clearer that analysing large-scale data with only single standard intelligent approaches is an intractable problem. In this sense, selecting highly discriminative SNPs, a problem called feature selection (FS) in the computational intelligence field (Mohamad et al., 2009; Zheng et al., 2009), has become increasingly appealing.

If a traditional classifier is used to classify a sample based on all the measured variables while using gene expression for disease diagnosis, a low accuracy is expected. The high number of features and the relatively small number of observations (samples), as microarray data, is a common phenomenon known in machine learning as the curse of dimensionality problem. The objectives of FS thus involve both minimizing the number of features that get selected and maximizing the classification performance. Dimensionality reduction methods are able to transform the initial dataset into one with a smaller number of features but with the same genetic information, which leads to better analyses. Its underlying idea is that some of the features (looked at as different dimensions of the problem) can be expressed in terms of others and are thus redundant. The combination of several intelligent

* Corresponding author.
E-mail addresses: aboutorh@usthb.dz (A. Boutorh), aguessoum@usthb.dz (A. Guessoum).

http://dx.doi.org/10.1016/j.engappai.2016.01.004
0952-1976/

approaches has been proven useful in analysing large, complex biological data, and is hence becoming more and more popular.

Evolutionary learning methods have already been successfully used in different microarray studies (Deutsch, 2003; Jirapech-Umpai and Aitken, 2005). In addition to gene selection, hybrid methods involving Evolutionary Algorithms have been successfully used to identify SNPs associated with various diseases.

On the other hand, it is commonly accepted that many complex diseases such as cancer arise from complex interactions among multiple SNPs (Zhang et al., 2008). This is known as multi-locus interactions (Cordell, 2002). Different ensemble methods have been proposed to identify SNP–SNP interaction (Upstill-Goddard et al., 2013). Using SNPs as genetic markers, our goal is to identify the complex interactions and the relationship among SNPs which may increase the classification performance of the disease of interest. The classification task is described as a pattern recognition problem. Feature extraction is the key to good pattern recognition since even the best classifier will perform poorly if the features are not well selected.

Motivated by the success of the combination of intelligent techniques for the FS and classification tasks on biological data, and by the high performance of Evolutionary Algorithms, and knowing that traditional feature selection techniques tend to ignore the interaction between features (Zhao and Liu, 2009), while their combination may have a strong correlation with the target (Shen et al., 2014), we propose in this study GA-NN-GE-ARM, a new hybrid intelligent technique based on Association Rule Mining (ARM) and Neural Networks (NN) which uses Evolutionary Algorithms, Grammatical Evolution (GE) to optimize the rule extraction, and Genetic Algorithms (GA) to find the best parameters for the two techniques combined.

The Hybridization of AR and NN for SNP data was presented first in Boutorh and Guessoum (2015) for breast cancer; its authors reported reaching a high accuracy. In this work, we propose an improvement of the work presented in Boutorh and Guessoum (2015) and show that the accuracy we reach is also improved.

The rest of the paper is organised as follows. The Background of the combined techniques in the current study (FS, GA, GE, ARM and ANN) are presented in Section 2 and their applications for genetic data are reviewed. In Section 3, we present the SNP datasets that are used in this work. The steps of our proposed approach (GA-NN-GEARM) are detailed in Section 4, and the experimental results are given in Section 5. Finally, Section 6 concludes the paper.

2. Background

Four related background topics, Feature Selection, Genetic Algorithms and Grammatical Evolution (two types of Evolutionary Algorithms), Association Rule Mining and Artificial Neural Networks are briefly presented in this section.

2.1. Feature selection

Dimensionality reduction (DR), or feature selection, techniques play a crucial role in DNA microarray analysis (Saeys et al., 2007; Bolón-Canedo et al., 2014). The importance of these techniques for SNP datasets is undeniable both in biology and machine learning. Their aims from the computation viewpoint is to tackle the curse of dimensionality by improving the prediction performance of the predictors, providing faster and more cost-effective predictors and facilitating data visualization and data understanding (Guyon and Elisseeff, 2003). FS techniques have been adopted from the data mining and signal processing literature and used to simplify the process of locating SNP loci associated with the studied diseases (Saeys et al., 2007).

By definition, dimensionality reduction refers to the process of removing irrelevant features and identifying a lower dimension representation of a set of variables so that the learning algorithm focuses only on those selected features of the training data that are useful for the analysis and predictions (Christin et al., 2013; Guyon and Elisseeff, 2003).

In other words, given an $\boldsymbol{n}$ dimensional dataset, we try to find a subset of $\boldsymbol{k}$ dimensions, where $\boldsymbol{k} < \boldsymbol{n}$, which still captures the content and all the information available in the original data.

The analysis of SNP data is a key to disease-gene association studies. Various dimensionality reduction approaches have been used to perform the selection of the most informative SNPs. In Batnyam et al. (2013), the authors combined various existing techniques to find the most effective SNP data classification. The analysis was conducted in three stages: first, the selection of informative SNPs; second, the generation of an artificial feature from the selected SNPs; third, the classification task. There are usually three types of feature selection methods: filters, wrappers and embedded methods (Mohamad et al., 2009; Seo and Oh, 2012).

Filtering approaches evaluate the attributes based on general characteristics of the training data to select features that are independent of any predictor (i.e. statistical measures). They use an easy-to-calculate metric which allows a quick ranking of the features, where the top-ranking features are selected. Classical filtering methods are usually applied to microarray data, such as Correlation Feature Selection (CFS), Fast Correlation-Based Filters (FCBF), and ReliefF (Saeys et al., 2007). The main advantage of filters is that they are much faster than wrapper methods, whereas their disadvantage is that they do not interact with the classifier, usually leading to worse performance results.

Wrapper approaches use a machine learning algorithm to check the effect of various subsets of features (Ian and Eibe, 2005; Ruiz et al., 2006). They involve optimizing a predictor as part of the selection process. The main advantage of wrappers is that they commonly offer better classification accuracy which is the most important aim. However, the known disadvantage of these approaches is that the algorithm builds a model many times in order to evaluate different subsets. This is too computationally expensive most often, especially for SNP datasets which tend to be very large.

Embedded approaches can be seen as an intermediate solution to overcome the drawbacks of filters and wrappers. They generally use machine learning models for classification which eliminate features as part of the training process.

FS methods may also be divided into univariate techniques and multivariate techniques. The former are fast and scalable but ignore feature dependencies (Bolón-Canedo et al., 2014; Saeys et al., 2007); the latter overcome the drawback of the univariate type and incorporate feature dependencies but at the cost of being slower and less scalable (Bolón-Canedo et al., 2014).

There is a large suite of feature selection methods that deal with microarray gene data. The reader can find some review work in Ma and Huang (2008); Lazar et al. (2012); Saeys et al. (2007).

Nowadays, the trend is to focus on new combinations such as hybrid methods which combine two or more algorithms of conceptually different origins for FS and classification. In Halakou et al. (2015) the authors proposed three hybrid selection methods CNNFS, Ck-NNFS, and CRRFS. In these techniques, all Neural Network, k-Nearest Neighbours, and Ridge Regression were respectively injected in the wrapper phase as induction algorithms. The obtained results showed the performance of the proposed hybrid methods in addition to their dimensionality reduction ability in SNP selection.