Chapter 4
Machine Learning Algorithms for Analysis of DNA Data Sets

John Yearwood
University of Ballarat, Australia

Adil Bagirov
University of Ballarat, Australia

Andrei Kelarev
University of Ballarat, Australia

ABSTRACT

The applications of machine learning algorithms to the analysis of data sets of DNA sequences are very important. The present chapter is devoted to the experimental investigation of applications of several machine learning algorithms for the analysis of a JLA data set consisting of DNA sequences derived from non-coding segments in the junction of the large single copy region and inverted repeat A of the chloroplast genome in Eucalyptus collected by Australian biologists. Data sets of this sort represent a new situation, where sophisticated alignment scores have to be used as a measure of similarity. The alignment scores do not satisfy properties of the Minkowski metric, and new machine learning approaches have to be investigated. The authors' experiments show that machine learning algorithms based on local alignment scores achieve very good agreement with known biological classes for this data set. A new machine learning algorithm based on graph partitioning performed best for clustering of the JLA data set. Our novel k-committees algorithm produced most accurate results for classification. Two new examples of synthetic data sets demonstrate that the authors' k-committees algorithm can outperform both the Nearest Neighbour and k-medoids algorithms simultaneously.

DOI: 10.4018/978-1-4666-1833-6.ch004

INTRODUCTION

Machine learning algorithms have useful applications in broad areas and are very important. Many valuable results on machine learning techniques have been obtained in the literature recently. To illustrate the broad character of associated applications let us just refer to a few articles by Bagirov & Yearwood (2006), Bagirov, Rubinov & Yearwood (2002), Haidar, Kulkarni & Pan (2008), Pan, Haidar & Kulkarni (2009), Verma & Kulkarni (2007), Witten & Frank (2005), Yearwood et al. (2009), Yearwood & Mammadov (2010).

On the other hand, the data sets of nucleotide and protein sequences have been rapidly growing, see Baldi & Brunak (2001) and Gusfield (1997). Enormous amounts of DNA, RNA and protein data are continuously being generated. This is why it is especially important to devise efficient machine learning algorithms in order to automate the analysis of nucleotide sequences.

Nucleotide and protein sequences stored in databases are very long. They cannot be accurately represented using short tuples of values of numerical or nominal feature attributes, and cannot be regarded as points in a finite dimensional space. In order to achieve agreement between classifications produced by machine learning algorithms and biological classifications, sophisticated local alignment scores have to be used as a measure of similarity between DNA sequences. These scores do not satisfy axioms of Minkowski metrics, which include as special cases the Euclidean distance, Manhattan distance, and max distance.

To verify the effectiveness of new machine learning methods for automated classification and clustering of DNA sequences, the researchers have to rely on classes and groupings of known data sets that have already been considered in the biological literature. A comparison of the results produced by new machine learning algorithms with known groupings is essential for automating further classifications and clusterings and the development of new advanced machine learning programs that may lead to discoveries of biological significance.

The present paper is devoted to experimental analysis of several algorithms for clustering and classification of a JLA data set derived from the non-coding segments in the junction of the large single copy region and inverted repeat A of the chloroplast genome in Eucalyptus collected by Australian biologists. We compare the effectiveness of several algorithms in their ability to achieve agreement with known biologically significant classes already obtained for this data set by Freeman, Jackson & Steane (2001).

Our experimental analysis shows that all algorithms based on local alignment scores achieve better results than straightforward alternatives using simple statistical measures. The experiments compare the results of k-medoids, Nearest Neighbour, k-committees algorithms, and a machine learning algorithm based on graph partitioning in their ability to achieve agreement with the results published in the biological literature before. All of these algorithms rely on local alignments. For unsupervised clustering of the JLA data set, the machine learning algorithm based on graph partitioning performed best. For supervised classification, the k-committees algorithm produced the most accurate results. Finally, we present two examples of synthetic data sets, where our novel k-committees algorithm outperforms both the classical Nearest Neighbour algorithm and the k-medoids algorithm.

The results demonstrate that machine learning algorithms based on local alignments achieve good agreement with classifications published in the biological literature. They can be used to obtain biologically significant machine learning results.

PRELIMINARIES AND BACKGROUND INFORMATION

We use standard machine learning terminology and notions and refer the reader to the monographs