Poster L-33

Meta-Learning of Clustering Algorithms: Analyzing Gene Expression Time Courses

Authors:
Francisco A. T. de Carvalho (*Centro de Informática, Universidade Federal de Pernambuco*)
Ricardo B. C. Prudencio (*Centro de Informática, Universidade Federal de Pernambuco*)
Marcílio C. P. de Souto (*Departamento de Informática e Matemática Aplicada, Universidade Federal do Rio Grande do Norte*)
Ivan G. Costa (*Dept. of Computational Molecular Biology, Max Planck Institute for Molecular Genetics*)
Benjamin Georgi (*Dept. of Computational Molecular Biology, Max Planck Institute for Molecular Genetics*)
Teresa Ludermir (*Centro de Informática, Universidade Federal de Pernambuco*)
Stefan Roepcke (*Dept. of Computational Molecular Biology, Max Planck Institute for Molecular Genetics*)
Alexander Schliep (*Dept. of Computational Molecular Biology, Max Planck Institute for Molecular Genetics*)

Short Abstract: Whether to cluster at all, which clustering method to use and how many clusters to choose are pressing questions in bioinformatics. Mostly, decisions are made based on experience guided by benchmarking or indicators for reliability of solutions or model-fit. We use meta-learning to answer these questions with machine learning techniques.

Long Abstract:
Meta-Learning of Clustering Algorithms: Analyzing Gene Expression Time Courses

Whether to cluster at all, which clustering method to use and how many clusters to choose are pressing questions in bioinformatics. Mostly, decisions are made by users of clustering software based on experience
guided by benchmarking or indicators for reliability of solutions or model-fit. However, as clustering algorithms always produce solutions, often inappropriate methods or parameters are used and invalid results
produced. Meta-learning refers to the application of machine learning techniques in choosing methods and guiding in setting parameters. We apply meta-learning to the problem of analyzing gene expression time-courses.

The mining of time-course data has attracted great attention in the data mining community in recent years. In this context, several models have already been developed such as, for instance, Box-Jenkins, exponential smoothing, different types of artificial neural networks.

Besides these classical models, clustering time courses and other sequences of data has become an important topic, motivated by several research challenges including gene expression data from molecular
biology, as well as the challenge of developing methods to recognize dynamic change in time

series.

In both cases, the selection of an adequate model for a given set of series can be a hard task depending in the candidate models and the characteristics of the series. For example, each different machine learning algorithm (e.g., k-means and model-based clustering methods) makes implicit or explicit assumptions about the structure of the data.

Specifically, in the context of the analysis of time series from gene expression, one of the greatest challenges is the high variability in the experiments. Contributing factors are the nature of the choices made by the specialist, or differences in the experimental setting. For these types of data, we can find distinct characteristics
such as: number of measured observations, number of replicated measurements, time between observations, technology used to quantify expression, presence (or absence) of missing data, among others.

Given the previous issues, it is very unlikely that a single clustering method will have a suitable performance in all gene expression time series. Despite this, it is a common practice in this field to choose the methods only based on the easiness of understanding and use, and not on the suitability of the method for the given data. In this context, the use of methods of meta-learning will not only achieve better results, as it will help the users on the choice of a method.

Meta-Learning approaches were developed in the field of supervised machine learning with the aim of selecting the best learning algorithm for a given problem. Meta-learning uses empirical examples to produce a machine learning model (meta-learner) responsible for associating the candidate machine learning algorithms (base-algorithm) with the
characteristics of the given problem. Each example for a meta-learning algorithm consists of a machine learning problem, described by one set of attributes associated with the performance obtained empirically by the base-algorithm in the given problem.

In this context, we will extend the meta-learning approach for the context of non-supervised learning. Then, we will apply it to gene expression time-course data, such as Yeast cell-cycle or Drosophila development.

Poster L-33

Computational prediction of cis-regulatory elements in abiotic stressed Arabidopsis.

Authors:
Juan Caballero-Perez (*CINVESTAV Campus Guanajuato*)
Octavio Martinez (*Langebio*)
Luis Herrera-Estrella (*Langebio*)

Short Abstract: Complex networks are present in the modulation of gene expression and in the determination of cis-regulatory elements that can be determined by using gene expression data from microarrays experiments with data mining studies. We used this

approach in public data for Arabidopsis and propose combinatorial models for abiotic stress.

Long Abstract:

Arabidopsis thaliana is a model plant, now with the complete genome sequenced, we can combine experimental data with computational analysis in the search of to understand how a plant respond to environment stress. Almost all the genes (verified and predicted) are represent in commercial microarray (as Affymetrix ATH1-25K) so many people are generating public experimental data. We had obtained experimental data for different abiotic stress conditions: phosphate starvation, sulfate limitation, cold shock, heat shock, osmotic stress, high salinity, drought, genotoxic conditions, and regulators treatments (ABA, IAA, ACC, zeatin, and GA3), all performed using the Affymetrix ATH1-25K which is an oligonucleotide-based chip representing approximately 23 000 genes of arabidopsis.

The data were filtered using a p-value threshold of < 0.05 and a signal threshold of >= 25. Other filter applied was the list of unspecific oligonucleotides obtained from TAIR. All data sets were normalized obtaining the logarithm base 2 of the ratio between conditions and adjusting with LOESS correction.

The complete genome for Arabidopsis thaliana was obtained from public repositories in NCBI, the version is the 5.0 released in november 2004. All sequences are mapped in the chromosomes, thus the coordinates of position and the coding chain for each gene were available. Promoters regions of 2 kbp of length was extracted upstream from the initial codon (+1) in annotated sequences.

For motif prediction, we used MotifSampler, a Gibbs sampler program, with a background made with all the promoter regions in the genome with a word length fixed of 8 bases, was used to find 50 most statistically represented common motifs of a fixed length of 8 bases, allowing overlapping and presence in both chains. Selection of size was found probing different lengths (5, 6, 7, 8, 9, 10, 11, 12) in a testing group formed by the purple phosphatase acid family of arabidopsis. The motifs predicted were compared with the reported in plant CAREs public databases (like PLACE and PlantCARE) to see if it has been previously reported.

For all the predicted CAREs a back feed method for pairs of motifs was searched in the test group and a control group (genes whit unaltered expression), to trace combinations of putative modules in different distances between motifs (5, 15, 25, 150 bases) using regular expressions in Perl-based scripts.