Semi-supervised learning for phenotypic profiling of high-content screens (DRAFT)

Roger Bermudez-Chacon
Supervisor: Peter Horvath

ETH Zurich

July 29, 2013

Abstract

Semi-supervised machine learning techniques are particularly useful in experiments where data annotation and classification is time- and resource-consuming or error-prone. In biological experiments this is often the case. Here, we apply a graph-based machine learning method to classify cells in different stages of infection with the Semliki Forest Virus (SFV), which features have been extracted from image analysis of fluorescence microscopy results, obtained in turn from a genome-wide high-content screening experiment. The aim of this project is to investigate whether and to which extent intelligent control experiment design combined with semi-supervised learning can reach the accuracy of a human annotator and/or in certain cases substitute it.

Introduction

Recent advancements in high-throughput microscopy and data analysis made possible to perform large scale biological experiments and automatically evaluate them. For the detection of sub-cellular changes caused by different perturbations in the cell (RNAi or drugs), often supervised machine learning (SML) is used. Reliable training of an SML method, however, requires significant effort from a field expert.

As an alternative, semi-supervised machine learning (SSL) methods make use of information intrinsically found in the entire data, both annotated and unannotated, thus allowing to make use of a larger amount of information by exploiting, alongside with the annotated data, the relative distribution of unannotated data on the feature space[1]. This paradigm, under a few assumptions[1], has proven valuable in exploring and classifying biological data in fields as diverse as drug-protein interactions[2], gene expression[3], and medical diagnosis[4].

use footnote with assumptions or just reference to publication?

expand this introduction?

Materials and Methods

High-content screening

A human genome-wide siRNA library was used to produce human cell cultures with knocked-out genes, stored in a collection of 55 16x24-well plates. These cell cultures were exposed to a genetically engineered fluorescent SFV strand, and the corresponding green fluorescent protein production on all of the cultures was tracked over time.

confirm this is correct/well-written, explain this in more detail?

The protein expression was stopped at 4, 5, 6, and 7 hours after culture infection with SFV and microscopic pictures of the sample were obtained under a light microscope. Samples with no exposure to SFV were also analyzed as a control experiment, in the exact same manner as for the infected samples.

Image adquisition and analysis

For every sample at each infection stage, 9 tiled images were captured via a light microscope, by composing the green fluorescent signal of the produced protein, and a blue-colored image of the nuclei. All images were subsequently processed with an automatic random forest-based segmentation tool to identify individual cells from the images.

according to G. Balistreri. Confirm. Expand on this?

With the mapping between microscopic pictures and individual segments representing cells, features for each cell were extracted with CellProfiler[5]. A total of 93 features were retrieved and used in this experiment, corresponding to color intensity, area, shape, and texture descriptors. (For a complete list of the features used, see Appendix A)

Annotated data

From the genome-wide information, a small subset of the data was manually annotated by an expert on SFV infection, by visually identifying cell phenotypes directly from the segmented microscopic images and cross-checking with the time annotation on the respective source plate, and classifying them into the different stages of infection. This manual process yielded 3098 annotated cells.[2]

Semi-supervised learning implementation

A graph-based label propagation (label spreading[6]) approach was followed. In this kind of approach, an undirected graph is built using the data points (cells) as vertices, and edges are created for all pairs of vertices that satisify a neighboring condition, with weights proportional to the degree of relatedness or association between the pair of vertices.

In the original formulation, labels are associated to the vertices corresponding to annotated data, and neutral labels to the unannotated data; then, in an iterative fashion, the labeled vertices propagate along the edges to their neighbors' labels, with a strength proportional to their relatedness (edge weight).

In the present implementation, prior knowledge of the nature of the data was incorporated as an additional level of *soft labeling*, to exploit the fact that, for a group of data points used as experimental control[3], the cells (vertices) can be tracked back to their experimental conditions, which have a direct influence on what specific phenotypes (labels) are more likely to occur.

Graph construction

Phasellus id magna. Duis malesuada interdum arcu. Integer metus. Morbi pulvinar pellentesque mi. Suspendisse sed est eu magna molestie egestas. Quisque mi lorem, pulvinar eget, egestas quis, luctus at, ante. Proin auctor vehicula purus. Fusce ac nisl aliquam ante hendrerit pellentesque. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Morbi wisi. Etiam arcu mauris, facilisis sed, eleifend non, nonummy ut, pede. Cras ut lacus tempor metus mollis placerat. Vivamus eu tortor vel metus interdum malesuada.

Feature selection

Ut congue malesuada justo. Curabitur congue, felis at hendrerit faucibus, mauris lacus porttitor pede, nec aliquam turpis diam feugiat arcu. Nullam rhoncus ipsum at risus. Vestibulum a dolor sed dolor fermentum vulputate. Sed nec ipsum dapibus urna bibendum lobortis. Vestibulum elit. Nam ligula arcu, volutpat eget, lacinia eu, lobortis ac, urna. Nam mollis ultrices nulla. Cras vulputate. Suspendisse at risus at metus pulvinar malesuada. Nullam lacus. Aliquam tempus magna. Aliquam ut purus. Proin tellus.

Label propagation

Phasellus fringilla, metus id feugiat consectetuer, lacus wisi ultrices tellus, quis lobortis nibh lorem quis tortor. Donec egestas ornare nulla. Mauris mi tellus, porta faucibus, dictum vel, nonummy in, est. Aliquam erat volutpat. In tellus magna, porttitor lacinia, molestie vitae, pellentesque eu, justo. Class aptent taciti

[1] Smoothness, Cluster, and Manifold assumptions, see[1] p. 4-6

[2] A small number of imaging artifacts were also identified manually. However, accounting for such information was out of the scope of this work.

[3] Wild type cultures designated for plate effect monitoring, with no RNA interference applied.

sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Sed orci nibh, scelerisque sit amet, suscipit sed, placerat vel, diam. Vestibulum nonummy vulputate orci. Donec et velit ac arcu interdum semper. Morbi pede orci, cursus ac, elementum non, vehicula ut, lacus. Cras volutpat. Nam vel wisi quis libero venenatis placerat. Aenean sed odio. Quisque posuere purus ac orci. Vivamus odio. Vivamus varius, nulla sit amet semper viverra, odio mauris consequat lacus, at vestibulum neque arcu eu tortor. Donec iaculis tincidunt tellus. Aliquam erat volutpat. Curabitur magna lorem, dignissim volutpat, viverra et, adipiscing nec, dolor. Praesent lacus mauris, dapibus vitae, sollicitudin sit amet, nonummy eget, ligula.

Results

Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Donec odio elit, dictum in, hendrerit sit amet, egestas sed, leo. Praesent feugiat sapien aliquet odio. Integer vitae justo. Aliquam vestibulum fringilla lorem. Sed neque lectus, consectetuer at, consectetuer sed, eleifend ac, lectus. Nulla facilisi. Pellentesque eget lectus. Proin eu metus. Sed porttitor. In hac habitasse platea dictumst. Suspendisse eu lectus. Ut mi mi, lacinia sit amet, placerat et, mollis vitae, dui. Sed ante tellus, tristique ut, iaculis eu, malesuada ac, dui. Mauris nibh leo, facilisis non, adipiscing quis, ultrices a, dui.

Discussion

Morbi luctus, wisi viverra faucibus pretium, nibh est placerat odio, nec commodo wisi enim eget quam. Quisque libero justo, consectetuer a, feugiat vitae, porttitor eu, libero. Suspendisse sed mauris vitae elit sollicitudin malesuada. Maecenas ultricies eros sit amet ante. Ut venenatis velit. Maecenas sed mi eget dui varius euismod. Phasellus aliquet volutpat odio. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Pellentesque sit amet pede ac sem eleifend consectetuer. Nullam elementum, urna vel imperdiet sodales, elit ipsum pharetra ligula, ac pretium ante justo a nulla. Curabitur tristique arcu eu metus. Vestibulum lectus. Proin mauris. Proin eu nunc eu urna hendrerit faucibus. Aliquam auctor, pede consequat laoreet varius, eros tellus scelerisque quam, pellentesque hendrerit ipsum dolor sed augue. Nulla nec lacus.

Conclusions

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Bibliography

[1] O. Chapelle, B. Schölkopf, A. Zien, *et al.*, *Semi-supervised learning*, vol. 2. MIT press Cambridge, 2006.

[2] Z. Xia, L.-Y. Wu, X. Zhou, and S. Wong, "Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces," *BMC Systems Biology*, vol. 4, no. Suppl 2, pp. 1–16, 2010.

[3] I. Costa, R. Krause, L. Opitz, and A. Schliep, "Semi-supervised learning for the identification of syn-expressed genes from fused microarray and in situ image data," *BMC Bioinformatics*, vol. 8, no. Suppl 10, p. S3, 2007.

[4] E. Bair and R. Tibshirani, "Semi-supervised methods to predict patient survival from gene expression data," *PLoS biology*, vol. 2, no. 4, p. e108, 2004.

[5] A. E. Carpenter, T. R. Jones, M. R. Lamprecht, C. Clarke, I. H. Kang, O. Friman, D. A. Guertin, J. H. Chang, R. A. Lindquist, J. Moffat, *et al.*, "Cellprofiler: image analysis software for identifying and quantifying cell phenotypes," *Genome biology*, vol. 7, no. 10, p. R100, 2006.

[6] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," *Advances in neural information processing systems*, vol. 16, no. 16, pp. 321–328, 2004.

[7] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*. John Wiley & Sons, 2001.

[8] R. M. Haralick, K. Shanmugam, and I. H. Dinstein, "Textural features for image classification," *Systems, Man and Cybernetics, IEEE Transactions on*, no. 6, pp. 610–621, 1973.

Appendix A

Features analyzed

[Table with cell/nuclei intensity, shape and Haralick[7] texture features...]