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

Roger Bermudez-Chacon
Supervisor: Peter Horvath

July 24, 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 allow to make use of a larger amount of information, by exploiting both annotated information and the relative distribution of unannotated data on the feature space.

expand this introduction?

Materials and Methods

High-content screening

A human genome-wide siRNA library was used to produce phenotypes of human cells with knocked-out genes. These cell cultures were infected with a genetically engineered strand Donec vel nibh ut felis consectetuer laoreet. Donec pede. Sed id quam id wisi laoreet suscipit. Nulla lectus dolor, aliquam ac, fringilla eget, mollis ut, orci. In pellentesque justo in ligula. Maecenas turpis. Donec eleifend leo at felis tincidunt consequat. Aenean turpis metus, malesuada sed, condimentum sit amet, auctor a, wisi. Pellentesque sapien elit, bibendum ac, posuere et, congue eu, felis. Vestibulum mattis libero quis metus scelerisque ultrices. Sed purus.

confirm this is correct/well-written

Data sources labeled and unlabeled datasources

Tools weka, numpy, scipy

methods label spreading

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.

write results

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.

write results

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.

fill in conclusions

Bibliography

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

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

[3] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[4] M. Estes, A. Kapikian, D. Knipe, and P. Howley, "Fields virology," *Philadelphia, PA: Lippencott, Williams and Wilkins*, 2007.

[5] I. Banerjee, Y. Yamauchi, A. Helenius, and P. Horvath, "High-content analysis of sequential events during the early phase of influenza a virus infection," *PLOS ONE*, vol. 8, no. 7, p. e68450, 2013.

[6] 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.