INFOWISE BIOSCIENCE

Dynamic data quality monitoring system for ONCOLOGY

95% precision of outlier detection, minimizing the errors.

65% work efficiency increase, saving the workload for researchers.

Real-time data monitoring

The web application enables scientists to cross check data and make sure both accuracy and precision of predicted results by switching from different classification algorithms.

Customized digest report

Even without accessing the app, scientists will receive custom--ized digest reports containing information about outlier images.

Interactive visualizations

Also, scientists are able to interact with charts and graphics demonstrating historical data comparison and distribution.

Problem

Problem Definition

In order to help people get free from cancer and tumors, tons of drugs need to be tested to figure out their effectiveness, duration, as well as drawbacks. Unfortunately, due to techno-logy limitations, the whole process have to be accomplished manually. Even for experts with years of experience, it takes a lot of time to detect various kinds of cell phenotypes one by one.

Project Description

Our sponsor, Presage is a bioscience company dedicated to research of effective cancer drug combinations. As one of the core process, image outlier detection are adopted to profile cell images so that people can easily identify and quantify cell phenotypes. This is a awesome project that illustrates the idea of converting data to information. We have access to all the classified images, and we'd like to use our knowledge about machine learning to train historical data and develop a web application to detect cell phenotypes automatically.

Analysis

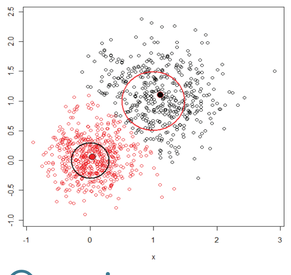
Model selection and model basics

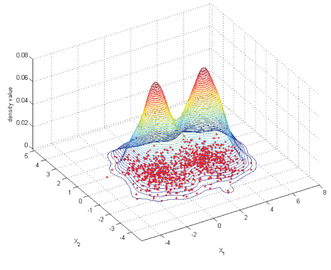
This is an unsupervised learning project and our goal is to classify images into two groups: good for further research, bad to eliminate. We selected two classic machine learning models:

K-means

The objective function of this algorithm is to minimize the within-cluster sum of square errors. This algorithm hard assi-gns each image to one label: 0 or 1 and stops till converge.

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in S_i} \|\mathbf{x}_j - \boldsymbol{\mu}_i\|^2$$





Mixture of Gaussians

Instead of assigning one image to only one label, this algorithm soft assigns each image to the probab-ilities that belong to each label. The objective function for this algorithm is to minimize a weighted sum of two component Gaussian densities:

$$p(\mathbf{x}|\lambda) = \sum_{i=1}^{M} w_i \, g(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i),$$

Implementation

Dataset

Our dataset consists of 1206 data points generated by real flu-orescence images of tissue sections. Each of them has 24 feat-ures thus we have planty of data to train and build our models.

Machine learning model

We used "scikit-learn" python package since it provides robust unsupervised learning programs.

Backend Implemetation

We choose python mainly because of its easy to read, implem-ent and has tons of great libraries.

Data Visualization

In order to better monitor and visualize the outcome, we ada-pted realtime updating spline chart from highcharts.js. Also we developed a heatmap with D3.js to demonstrate the historical distribution of outliers.

presage BIOSCIENCES

Information School
UNIVERSITY of WASHINGTON

Lin Li , Jun Qian | MS in Information Management 2012 cohort