Machine Learning Methods for Microarray Data

Claire Laville
T00553911
lavillec16@mytru.ca

1 Introduction

This study evaluates the performance of machine learning (ML) models in the classification of celiac disease from gene expression data. It further investigates the genes selected by one of the models as most salient to the disease.

1.1 Gene Expression

Gene expression is the process whereby an organism's DNA is transcribed into RNA and subsequently translated into amino acids. Expression is regulated by proteins called transcription factors, which bind to non-coding DNA sequences called promoter or enhancer sequences. Activator proteins initiate transcription, whereas repressor proteins turn it off. Transcription is also repressed if the DNA molecule is methylated. Abnormal methylation may be triggered by environmental factors such as diet, chemical exposure, and extreme stress. The study of gene expression has led to targeted gene therapies along with a deeper understanding of biological processes.

Microarray experiments are the most precise way to measure gene expression. A microarray is a chip etched with microscopic wells, each of which is coded with oligonucleotide probes for a different gene. mRNA is collected from a population and then reverse-transcribed into complementary DNA (cDNA) or complementary RNA (cRNA). Probes on the chip bind to complementary sequences, and the expression level can be quantified by the exposing probe to a laser beam. Studies can be conducted on a single population (*e.g.*, to study the effects of a medical treatment over time) or on multiple populations (*e.g.*, to compare lab-grown mutant organisms to the wild type).

The NCBI Gene Expression Omnibus (GEO) [1] is the largest public repository of microarray data. A GSE (series) file consists of multiple gene expression samples (GSM accession numbers), each containing normalized microarray data from an individual, as well as details about each probe on the microarray (GPL accession number). The GEO also generates visual expression profiles for each gene in a series and heat maps showing genes with similar expression patterns.

1.2 Machine Learning

In the branch of supervised ML known as classification, an algorithm is trained on labelled data, iteratively learning to predict the label (e.g., disease state) of novel data. Previous applications of classification algorithms to microrarray data include the differentiation of leukemias [2], predicting the outcome of breast cancer treatment [4], and determining the cause of heart failure [6]. Among the most widely used algorithms are decision trees, random forests, support vector machines, and artificial neural networks.

1.3 Celiac Disease

Celiac disease [9] is an autoimmune disorder which causes a wide range of gastrointestinal and neuropsychological symptoms upon the consumption of gluten. It affects roughly 1% of humans and has also been detected in dogs and mice. The cause of celiac disease is multifactorial, with genetic and environmental components. Given this, a significant amount of noise and complexity in the gene expression data is to be expected.

In 2009, geneticists [5] extracted white blood cells from of 132 unrelated adults in the UK, 110 of whom had celiac disease and had successfully managed their symptoms with a gluten-free diet. mRNA from the subjects' blood was reverse-transcribed into cRNA. This cRNA was hybridized to an Illumina HumanRef-8 v.2 beadchip microarray, which probes for 18,981 genes. The resulting gene expression series was the dataset used for this experiment.

2 Methods

The gene expression series GSE11501, an 80MB Simple Omnibus Format in Text (SOFT) file, format, was downloaded from the GEO. The open-source Python library GEOparse [3] was used to extract expression data to a Pandas dataframe for analysis. A lookup table was constructed from the GPL file to link the Illumina probe identifiers with the associated gene symbol, description, and RefSeq accession number.

Classification was conducted in Jupyter Notebooks using the ML library SciKit-Learn [10]. The expression series was divided into training and testing sets with a 75%/25% test-train ratio. The data was stratified such that the training and testing sets each contained the same proportion of healthy controls (approximately 17%). Before fitting each model to the training data, the random seed was set to 0 for consistency.

The imbalance between samples from celiac patients and healthy controls posed a problem for training and cross-validation. If the classifier labelled all instances as positive, it would achieve a raw accuracy rate of about 83%. For that reason, the classifiers were evaluated based on f1 score, the harmonic mean of the true positive rate and sensitivity to false positives. The f1 scores given here are weighted to reflect the imbalance of classes in the training and test data.

3 Results

3.1 Predictive Accuracy

An f1 score of 1.0 indicates perfect prediction; this is impossible with naturally occurring data.

Parameters: Gini index and entropy are two similar methods for building decision trees. Entropy measures the information gain that would result from splitting the tree at a particular node (column value), whereas the Gini index compares the impurity (number of mislabelled items) resulting from a random partition of the data. Limited-memory Broyden–Fletcher–Goldfarb–Shanno (LBFGS) and Adam are optimization algorithms for neural network classifiers. LBFGS is noted for better performance on smaller data sets [10], as exhibited below.

Classifier	f1-score	Remarks
Decision Tree (Gini)	0.77	low score 0.68
Decision Tree (entropy)	0.84	
Random Forest	0.74	
Gaussian Naive Bayes	0.80	
Gradient Descent	0.89	loss: logistic
Support Vector Machines	0.91	
Perceptron (LBFGS)	0.74	all positive
Perceptron (Adam)	0.06	all negative

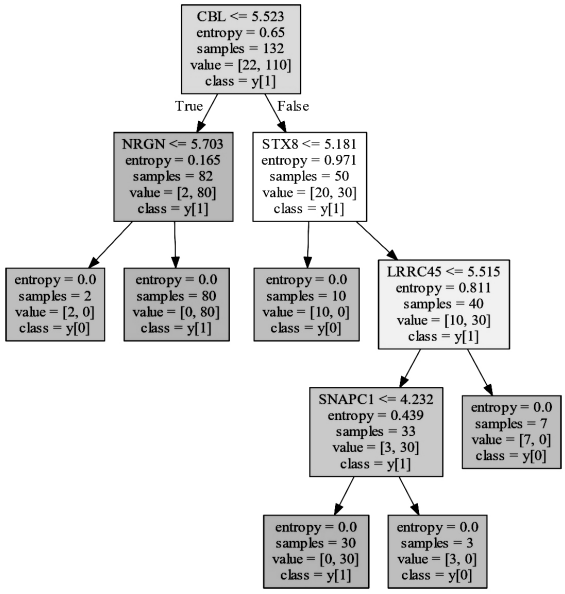

Figure 1. Decision tree (y[1] = celiac)

3.2 Identification of Salient Genes

The decision tree, while not the most accurate classifier of the models above, is valuable in that it builds a white-box (human-readable) model of the genes involved in predicting disease state. A second decision tree (Figure 1) using entropy, with no test data withheld, was compared with ground-truth knowledge of the genetics of celiac disease.

The two genes most strongly correlated with celiac disease belong to the human leukocyte antigen (HLA) complex on chromosome 6, although non-HLA genetic factors are believed to determine the clinical manifestation of the disease [5, 9]. However, the critical HLA-DQA1 and HLA-DQB1 genes are not probed for by the HumanRef-8 microarray. Ground-truth comparison was impossible in this case.

The gene that was selected by the decision tree as the most salient, CBL, is a proto-oncogene on chromosome 11. It is not cited as a locus of interest for celiac disease. Nevertheless, based on the gene expression profile generated by GEO (https://tinyurl.com/CBLprofile), it is apparent that the samples in the control group expressed this gene at higher levels. The sample size is too small to determine whether CBL is indeed involved in the disease.

4 Discussion

Naïve ML models are able to predict celiac disease with a fairly high level of accuracy, but more research is needed to determine whether it can correctly identify critical genes.

The primary challenge in applying ML to gene expression data is the large number of genes and small number of samples in an expression series. Strategies for addressing this problem fall into two categories: dimensionality reduction, for reducing the number of features inputted to the model; and the creation of novel samples through computation.

Some dimensionality reduction has already been undertaken. In particular, the use of recursive feature elimination with cross-validation (RFECV) boosted the performance of SciKit-Learn's linear classifiers (support vector machines and stochastic gradient descent) considerably. Both RFECV and the less-effective SelectPercentile are algorithms for feature selection: that is, they consider each column in isolation. Feature *extraction* methods, such as principal component analysis, create novel features from correlated columns and may be better at capturing interactions among genes [11]. Researchers have used autoencoders, a type of neural network that produces a lower-resolution representation of the input, to remove noise from expression data [13].

Simulated gene expression samples must accurately capture the variations in the natural data. Artificial microarray data has been created been synthesized using statistical inference [4, 7, 8] and generative adversarial networks [12].

5 Repository

Jupyter notebooks for this project can be viewed at https://github.com/clearthecity/microarray.

References

[1] T. Barrett and R. Edgar. 2006. Gene expression omnibus: microarray data storage, submission, retrieval, and analysis. *Methods Enzymol* 411 (Jan. 2006), 352–369. https://doi.org/10.1016/S0076-6879(06)11019-8

[2] H. Cualing, R. Kothari, and T. Balachander. 1999. Immunophenotypic diagnosis of acute leukemia by using decision tree induction. *Lab. Invest.* 79, 2 (Feb. 1999), 205–212.

[3] Rafal Gumienny. 2020. guma44/GEOparse. https://github.com/guma44/GEOparse original-date: 2015-08-16T08:26:27Z.

[4] Blaise Hanczar and Edward Dougherty. 2010. On the Comparison of Classifiers for Microarray Data. *CBIO* 5, 1 (March 2010), 29–39. https://doi.org/10.2174/157489310790596376

[5] Graham A. Heap, Gosia Trynka, Ritsert C. Jansen, Marcel Bruinenberg, Morris A. Swertz, Lotte C. Dinesen, Karen A. Hunt, Cisca Wijmenga, David A. Vanheel, and Lude Franke. 2009. Complex nature of SNP genotype effects on gene expression in primary human leucocytes. *BMC Med Genomics* 2 (Jan. 2009), 1. https://doi.org/10.1186/1755-8794-2-1

[6] Xiaohong Huang, Wei Pan, Suzanne Grindle, Xinqiang Han, Yingjie Chen, Soon J. Park, Leslie W. Miller, and Jennifer Hall. 2005. A comparative study of discriminating human heart failure etiology using gene expression profiles. *BMC Bioinformatics* 6, 1 (Aug. 2005), 205. https://doi.org/10.1186/1471-2105-6-205

[7] Hye Young Kim, Seo Eun Lee, Min Jung Kim, Jin Il Han, Bo Kyung Kim, Yong Sung Lee, Young Seek Lee, and Jin Hyuk Kim. 2007. Characterization and simulation of cDNA microarray spots using a novel mathematical model. *BMC Bioinformatics* 8, 1 (Dec. 2007), 485. https://doi.org/10.1186/1471-2105-8-485

[8] Matti Nykter, Tommi Aho, Miika Ahdesmäki, Pekka Ruusuvuori, Antti Lehmussola, and Olli Yli-Harja. 2006. Simulation of microarray data with realistic characteristics. *BMC Bioinformatics* 7 (July 2006), 349. https://doi.org/10.1186/1471-2105-7-349

[9] Online Mendelian Inheritance in Man (OMIM). 2019. CELIAC1 [#212750]. https://omim.org/entry/212750

[10] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830. https://scikit-learn.org/stable/index.html

[11] Markus Ringnér. 2008. What is principal component analysis? *Nat Biotechnol* 26, 3 (March 2008), 303–304. https://doi.org/10.1038/nbt0308-303 Number: 3 Publisher: Nature Publishing Group.

[12] Xiaoqian Wang, Kamran Ghasedi Dizaji, and Heng Huang. 2018. Conditional generative adversarial network for gene expression inference. *Bioinformatics* 34, 17 (Sept. 2018), i603–i611. https://doi.org/10.1093/bioinformatics/bty563 Publisher: Oxford Academic.

[13] James Zou, Mikael Huss, Abubakar Abid, Pejman Mohammadi, Ali Torkamani, and Amalio Telenti. 2019. A primer on deep learning in genomics. *Nat Genet* 51, 1 (Jan. 2019), 12–18. https://doi.org/10.1038/s41588-018-0295-5