LUNG CANCER SUB-CLASSIFICATION BASED ON GENE EXPRESSION PROFILES

Nathan LeRoy & Bingjie Xue
Department of Biomedical Engineering
University of Virginia
Charlottesville, VA 22903, USA
{xjk5hp,bx2ur}@virginia.edu

Kshitij Bhatta
Department of Mechanical and Aerospace Engineering
University of Virginia
Charlottesville, VA 22903, USA
qpy8hh@virginia.edu

ABSTRACT

Cancer is a complex disease related to abnormal cell growth with the potential to invade or spread to other parts of the body. Many cancer types/sub-types exist and are each associated with unique phenotypes, treatment plans, and prognoses. A current biomedical challenge is to predict the course of a cancer using as little information as possible and minimal clinical intervention. Using the ever-increasing landscape of annotated biological datasets available to researchers, we wish to create a machine learning model that can differentiate between two common subtypes of human lung carcinoma. Using publicly available data from the National Cancer Institute we wish to frame a binary classification task to differentiate squamous cell neoplasms cancer's from adenomas and adenocarcinomas cancers. We plan to utilize a support vector machine with a radial basis function kernel for this classification task.

1 PROJECT DEFINITION

Early detection and diagnosis of cancer is of paramount importance when it comes to clinical treatment plans and patient outcomes. Different cancer sub-types can be more aggressive than others and also may respond better to different kinds of therapies. Quick and accurate classification of a cancer's sub-type or lineage can help to guide a physicians decision making at point of care through minimally invasive techniques. We wish to utilize data that may be available post-biopsy to help inform this decision making process. Specifically, we wish to predict one of two lung cancer sub-type based on the unique gene expression profile of a patients tumor. We plan to utilize support vector machines (SVM) with a non-linear kernel to solve this binary classification task - specifically the radial bias function (RBF) kernel. The hope is that the learned model can accurately predict lung cancer sub-type from data outside the training set. The two sub-types considered in this problem are "Squamous cell Neoplasms" and "Adenomas and Adenocarcinomas".

1.1 INPUT OF THE PROBLEM

The input of the problem consists of an individual tissue sample. More specifically, the **gene expression profile** of the tissue sample. This tissue sample most likely comes from a primary tumor biopsy. This is a multidimensional input where each dimension represents a specific gene's level of expression in a normalized expression unit, FPKM (Fragments Per Kilo base of transcript per Million mapped reads, Eq. 1). This represents the abundance of certain genes in the sampled tissues. The dimension of the input is in the order of tens of thousands and depending on patterns found in gene expression across all genes, it is possible to find a pattern for different kind of cancers.

$$\text{FPKM}_i = \frac{X_i}{\tilde{l}_i N} \cdot 10^9 \tag{1}$$

where, $\tilde{l}_i$ is the effective length of a gene fragment, X_i is the read counts, and N is the total number of reads.

1.2 Output of the problem

The output of the problem is a binary classification: 1 if the sample is predicted to contain "squamous cell neoplasms" and -1 if the sample is predicted to contain "adenomas and adenocarcinomas".

1.3 Motivation for the problem

Different cancer sub-types can be more aggressive than others and also may respond better to different kinds of therapies. Early detection and diagnosis is paramount to effective clinical treatment and improving outcomes. This quick and accurate classification of a cancer's sub-type or lineage can help to guide a physicians decision making at the point of care through minimally invasive techniques. We hypothesize that a support vector machine will be sufficient to classify lung cancer sub-types based on a specific genomic profile after tissue biopsy.

2 Proposed Idea

2.1 Data Pre-process

The data input to this model is RNA sequencing data (See section 4 for more detailed information). For each of the transcriptome profiling files, we will extract: 1) gene names or identifiers; and 2) the per million mapped fragments (FPKM) scores of the genes. We will map the gene names or identifiers to the UniProt accessions to create a standardized input space. The complete set of training data will an m by n matrix with m training samples and an n-dimensional input.

2.2 Dimensionality Reduction

The large number of input features most likely will require the use of dimensionality reduction techniques such as Principle Component Analysis (PCA), Uniform Manifold Approximation and Projection (UMAP) or t-Distributed Stochastic Neighbor Embedding (t-SNE) to reduce the size of our input space. Applying one of the aforementioned methods will result in data with a lower number of new features that still retains most of the information and variability of the original feature-space.

2.3 Identify Gene Signatures with Statistical Regression

In addition to dimensionality reduction using the above methods, we will also apply statistical regression algorithms including 1) the least absolute shrinkage and selection operator (LASSO), 2) ridge regression, and 3) elastic-net regression to relate the expression of the genes to the cancer subtypes and generate classifiers and gene signatures of the cancer. These are more canonical examples of dimensionality reduction techniques in the bioinformatics space.

2.4 Train the Classifier Model

We will apply support vector machines (SVM) with non-linear radial bias function (RBF) kernels to the expression of 1) the features identified from the dimension reduction, and 2) gene signatures identified from the statistical regression to train the classifier model.

2.5 Evaluation of the Performance

We will use a standard train/test split to evaluate the performance of the method.

3 Related Work

The size, scale, and diversity of biological datasets available is increasing at an alarming rate. As such, researchers have been incredibly quick to apply many machine learning techniques to said data. In the context of cancer, the core task at hand is to classify the cancer in such a way that we can make informed decisions about how it might respond to specific treatments, its propensity for metastasis, or the prognosis of the patient with the identified cancer. One such example is from a 2015 study by Li et al. (2015) who utilized epigenetic data from ChIP-seq experiments to predict the gene expression profiles in lung cancer. These profiles could then be correlated with disease outcomes and further clusterd to create ontological classifications of cancer sub-types. The researchers found greatest success leveraging two different non-linear classifiers: Gaussian SVM and Random Forest.

Similarly, gene expression data can itself be utilized to predict a certain known sub-type of a cancer. In 2021, Simsek et al. (2022) found success by training SVM's on gene expression data to predict the specifc sub-type of luekemia at a very early stage. The researchers noted that early diagnosis and detection of cancer sub-types can guide clinical treatment plans.

Classification of cancer sub-types is not limited to *just* transcriptomic or epigenetic data. Imaging data can also be directly used to predict cancer sub-types for early detection as well. Recently (2021) Bębas et al. (2021) showed that non-small-cell lung cancer sub-types can be accurately predicted using patient PET/MR scans alone. The researchers found the best results (75.48 %) were achieved using an SVM classifier and texture parameters histogram of oriented gradients (HOG) while obtaining the highest values of specificity and sensitivity.

While the above methods have shown great promise, none directly address the problem that we have laid out. That is: classification of lung cancer data from gene expression data alone. We seek to extend off these previous works by taking a similar approach utilized above: leveraging non-linear SVM classifiers to predict the specific sub-type of lung cell carcinoma based on gene expression data alone. Based on the current body of literature and large volume of available data, we feel that we will be successful in our classification task.

4 Datasets

For this project, we will download the transcriptome profiling data of human lung cancer that annotated as "squamous cell neoplasms" and "adenomas and adenocarcinomas" from The Cancer Genome Atlas (TCGA) (https://www.cancer.gov/tcga). This dataset includes 1,526 transcriptome profiling files from 1,202 patient cases: 609 cases (755 files) are squamous cell neoplasms; and 593 cases (771 files) are adenomas and adenocarcinomas. We will use a train/test split to evaluate the performance of the method.

5 Timeline

5.1 Week of April 4th

- Data for the chosen cancer sub-classes will be collected from The Cancer Genome Atlas.
- Data prepossessing step is carried out to extract required features and ensure same labels across different sample files.
- A dimensionality reduction technique will be chosen and implemented to reduce the number of input features.

5.2 Week of April 18th

- Majority of coding and model development will be finished.
- The team will iterate on the model and optimize the hyperparameters for training.
- Testing code will be generated for model characterization and performance.

5.3 Week of May 2nd

- The performance of the method will be evaluated using a train/test split.
- The team will finalize and optimize the code based on the performance evaluation.
- The team will prepare the presentation and write the final project report.

REFERENCES

Ewelina Bębas, Marta Borowska, Marcin Derlatka, Edward Oczeretko, Marcin Hładuński, Piotr Szumowski, and Małgorzata Mojsak. Machine-learning-based classification of the histological subtype of non-small-cell lung cancer using MRI texture analysis. *Biomedical Signal Processing and Control*, 66:102446, 2021. ISSN 1746-8094. doi: 10.1016/j.bspc.2021.102446. URL https://www.sciencedirect.com/science/article/pii/S1746809421000434.

Jeffery Li, Travers Ching, Sijia Huang, and Lana X. Garmire. Using epigenomics data to predict gene expression in lung cancer. *BMC Bioinformatics*, 16(5):S10, 2015. ISSN 1471-2105. doi: 10.1186/1471-2105-16-S5-S10. URL https://doi.org/10.1186/1471-2105-16-S5-S10.

Ebru Simsek, Hasan Badem, and Ibrahim Taner Okumus. Leukemia Sub-Type Classification by Using Machine Learning Techniques on Gene Expression. In Xin-She Yang, Simon Sherratt, Nilanjan Dey, and Amit Joshi, editors, *Proceedings of Sixth International Congress on Information and Communication Technology*, pages 629–637. Springer, 2022. doi: 10.1007/978-981-16-2102-4_56.