Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: 06 Dec, 2022
School Project

BIOINF593 Final Project

Prediction of Metastasis Event using Hierarchical Classification with Elastic Nets

Alec Chu [1,*], Benjamin Osafo Agyare [2,*] and Blessing Oloyede [3,*]

[1]Department of Cellular and Molecular Pathology, University of Michigan, Ann Arbor, 48109, USA
[1]Department of Bioinformatics, University of Michigan, Ann Arbor, 48109, USA
[2]Department of Statistics, University of Michigan, Ann Arbor, 48109.
[3]Department of Chemical Biology, University of Michigan, Ann Arbor, 48109, USA.

*To whom correspondence should be addressed.

Received on Dec 5; revised on Dec 5; accepted on Dec 5

Abstract

Motivation: Metastasis is major contributor towards cancer-related mortality and can be difficult to detect during early stages. The ability to identify cancers that may have already metastasized can help increase patient survival. In this study, we utilize publicly available expression profile datasets of cancers from primary sites with or without distal metastasis. We train an elastic net models to predict the origin of primary cancer tissue and whether the primary cancer has metastasized or not.
Results: Using the elastic-net for hierarchical classification, we were able to predict the origin tissue at an accuracy of 97% and whether the cancer has already metastasized at an accuracy of 90%. When examining the top influential genes in the model we find that many mitochondrial genes were negatively correlated with metastasis.
Availability: All results, tools, available upon request.
Supplementary information: Supplementary data are available upon request.

1 Introduction

Cancer metastasis is the spread of cancer cells from a primary tumor site to surrounding tissues or distant locations and contributes to around 90% of cancer moralities ((Seyfried, 2013)). In recent years, cancer incidences have been steadily increasing due to multiple factors such as longer average lifespans and better early detection. To address this, advancement in cancer therapeutics and surgeries has resulted in drastic improvement in prognosis for most localized cancers. However, patient survival continues to be significantly impacted after detection of metastasis, which is most commonly is detected during cancer recurrence and can be years after tumor resection. The appearance of metastasis after tumor resection implies that the majority of cancers have metastasized prior to surgery. This makes detection of cancers with metastatic potential important so that patients can begin conventional metastasis treatments such as surgery, chemotherapy, hormone therapy, immunotherapy, or radiation therapy earlier.

Different cancers also have different metastatic potential. An estimated 6% of breast cancer patients are presented with metastasis, with bones, brain, liver, and lungs being the most common metastasis location. On the other hand, approximately 60% of lung cancer patients will have metastasis commonly to the brain, bone, liver, adrenal glands, thoracic cavity, or distal lymph nodes ((Riihimäki, 2018)). Differences in metastatic potential and location of metastasis makes early detection and diagnosis paramount for effective treatment.

For a cancer to metastasize, it must overcome a series of obstacles including detachment from the primary tumor location, intravasate into the circulatory and lymphatic system, evade immune responses and attacks, extravasate at distal capillary locations, invasion of distal locations, and proliferation at distal locations ((Hunter, 2008)). In spite of the prevalence of metastasis in cancer patients, the mechanism of metastasis is extremely inefficient and risky for detached cells. Current efforts into using these circulating tumor cells (CTCs) as a biomarkers for metastasis have been inconclusive due to the difficulty of accurately predicting whether these tumor cells are capable of establishing metastasis at distal locations.

Therefore, new methods and ways to help predict the metastatic potential in patients will be an asset for clinicians. Early detection of metastatic events can drastically improve patient survival by introducing them to therapeutics earlier and treating distal metastasis while it is still small.

In this study, we use publicly available expression datasets to investigate if expression data from primary tissue can be a predictor

of metastasis events using a hierarchical classification mechanism as discussed in the next sections of this paper. We then applied the model to examine if there are shared underlying gene expressions that can help predict whether a cancer has metastasized or not.

2 Approach

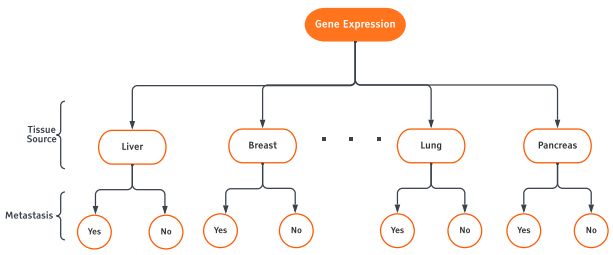

Fig. 1. Hierarchical structure of the classification task.

We implement a hierarchical classification model that follows the structure as seen in Figure 1.

First, train an elastic-net multinomial regression model to classify from which tissue source the tumor grew. Specifically, we predict the tissue source as one of the seven (7) cites, i.e breast, colon, gastric, kidney, liver, lung, pancreas, and skin. We then fit a hierarchical elastic-net logistic regression model to predict whether the cancer has metastasized. Due to the hierarchical nature of our prediction task, we define accuracy as ability to first correctly predict the tissue source, and then conditionally predicting metastasis given the predicted tissue sources and the gene expressions.

3 Methods

3.1 Datasets

3.1.1 Data sources

Expression data from various primary tissues with and without metastasis were downloaded from the GDC. We chose primary tissues of the most common tissue types on the GDC and compiled a total of approximately 5000 samples. These samples represent data collected from various studies including The Cancer Genome Atlas Program (TCGA), Clinical Proteomic Tumor Analysis Consortium (CPTAC), Human Cancer Models Initiative (HCMI), Count Me In (CMI). Around 400 additional datasets were also gathered from various independent studies available on GEO ((Kim SK, 2014) (Siegel MB, 2018) (Rothwell DG, 2014) (McDonald OG, 2017a) (Badal B, 2017) (McDonald OG, 2017b) (Wang, 2021) (Menck K, 2022)) to further increase the number of metastasis samples and introduce different types of expression data. Overall, our samples are represented by about 23% Breast, 11% Colon, 8% Gastric, 14% Kidney, 8% Liver, 20% Lung, 2% Pancreas, and 15% Skin tissues. With respect to metastasis, approximate 80% of samples were primary tissues without metastasis while 20% were from primary tissues with metastasis.

3.1.2 Data-Prepossessing

To simplify the model, we filtered out non-protein-coding genes from the dataset such as lncRNAs and ncRNAs to remove sources of confusions. To ensure standardization of the 19,938 features, we utilized transcripts per millions (TPM), then transformed the features by Z scores to obtain unit variance across features. This ensures that differences between samples and methods can be normalized to sequencing depth and that no feature will dominate the predictive power of the model by their raw scale.

As a matter of key relevance, to reduce data leakage, it is noteworthy that we are standardizing the test feature sets based on the mean and standard deviation of the training features. The section below discusses in detail the strategies adopted in splitting the data.

3.1.3 Data-Splitting

To assess the performance of the models under consideration, we adopt training-testing splitting where the training set comprises 70% of the dataset, totalling 3,875 samples while the testing set takes up the remaining 30%, totalling 1,665 samples. To foster adequate representation of all tissue sources, especially as our data is imbalanced, the split was performed using stratified sampling.

The training set was further split into cross-validation sets. This step is highly imperative in model selection and parameter tuning. Specifically, randomly splitting and assigning each the 3,875 training samples into 10 folds, we obtain a training and validation sample of 3,488 and 387 respectively, yet ensuring proportionate representation of each tissue source across all splits.

3.2 Modeling

3.2.1 Multinomial Model

Given our multi-class prediction task for the tissue source prediction, we use the multinomial model which extends the binomial when the number of classes is more than two Hastie T (2021). Suppose the response variable has K levels $\mathcal{G} = \{1, 2, \ldots, K\}$ and features $X \in \mathcal{R}^{N \times p}$ for a dataset of sample size N with p predictors. Here we model

$$\Pr(G = k \mid X = x) = \frac{e^{\beta_{0k} + \beta_k^T x}}{\sum_{\ell=1}^{K} e^{\beta_{0\ell} + \beta_\ell^T x}} \tag{1}$$

Thus, there is a linear predictor for each class.

3.2.2 The Elastic-Net Model

The elastic net (Zou, 2005) is a regularized method that coalesces the L_1 and L_2 penalties of the lasso and ridge regression methods by learning from their shortcomings to improve the regularization of statistical models. It does so by linearly combining the variable selection feature of the lasso and parameter shrinkage property of the ridge model simultaneously. This effective regularization technique allows for controlling multicollinearity, perform regression in high dimensional data settings ($p >> n$), and reduce excessive noise in our data to allow for isolating the most influential variables while balancing prediction accuracy Boehmke and Greenwell (2019). Specifically, for a multi-class prediction task given by 1, we specify our model is as follows:

Let Y be the $N \times K$ indicator response matrix, with elements $y_{i\ell} = I(g_i = \ell)$. Then the elastic net penalized negative log-likelihood function becomes

$$\ell\left(\{\beta_{0k}, \beta_k\}_1^K\right) = -\frac{1}{N}\sum_{i=1}^{N}\left(\sum_{k=1}^{K} y_{il}\left(\beta_{0k} + x_i^T\beta_k\right)\right) + \frac{1}{N}\log\left(\sum_{\ell=1}^{K} e^{\beta_{0\ell} + x_i^T\beta_\ell}\right) + \lambda\left[(1-\alpha)\|\beta\|_F^2/2 + \alpha\sum_{j=1}^{p}\|\beta_j\|_1\right] \tag{2}$$

Here β is a $p \times K$ matrix of coefficients. β_k refers to the k th column (for outcome category k), and β_j the j th row (vector of K coefficients for variable j) Hastie T (2021). The tuning parameters $\lambda \geq 0$ and $\alpha \in [0, 1]$ control the amount of regularization, and the mixing rate of the ridge and

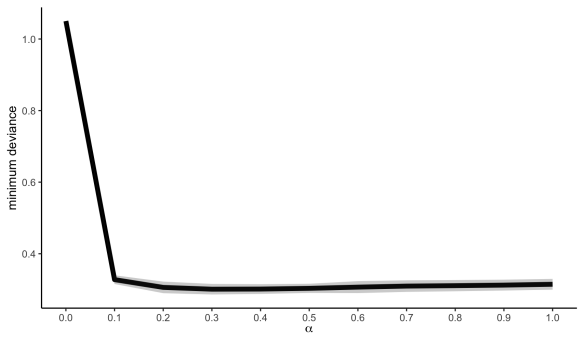

Fig. 2. Hyperparameter tuning for α. The best tuned α is 0.4, as it yields the smallest multinomial deviance. This is used in training our elastic net model for predicting tissue source.

Table 1. Confusion Matrix for tissue cite prediction.

		Ground Truth							
		Breast	Colon	Gastric	Kidney	Liver	Lung	Pancreas	Skin	Total
Prediction	Breast	373	0	0	0	0	1	0	0	374
	Colon	1	169	4	1	2	0	2	1	180
	Gastric	1	1	120	1	0	1	1	0	125
	Kidney	0	0	0	221	0	0	0	0	221
	Liver	0	1	0	0	137	1	0	0	139
	Lung	2	3	5	2	0	310	5	1	328
	Pancreas	0	1	2	0	0	2	48	0	53
	Skin	0	0	0	0	1	1	0	243	245
	Total	377	175	131	225	140	316	56	245	1665

The elastic-net at $= 0.4$ yields an impressive accuracy of 97.36%. The confusion matrix further details what classes are better predicted. Generally, the model easily misclassifies tissues as lung as seen from the table in terms of prediction error.

lasso penalties respectively. Thus, setting $\alpha = 0$ leaves a ridge model while an α value of 1 resets to a lasso model.

3.3 Tumor cite prediction

Using tumor source as our response and the genes as our features, we fit the elastic-net multinomial regression model. Our prediction, $\hat{G}_1 = \Pr(G = k|X)$ for $\mathcal{G} = \{1, 2, \ldots, 7\}$ classes, each for the tumor source. To tune the hyperparameters, we first create a common fold id, which allows us to apply the same cross-validation folds to each model. We subsequently create a tuning grid that searches across a range of alphas, given by $\alpha = \{0, 0.1, 0.2, \ldots, 1\}$. To obtain the optimal value of $\lambda \geq 0$, we train our models on the 10-fold cross-validation. The optimal λ is obtained as the value that minimizes the multinomial deviance from the model fits. By iterating this procedure for all values of α, we obtain the best training model as one that attains the lowest deviance. This is used to make prediction for subsequent analysis.

3.4 Metastasis prediction

Our metastasis prediction follows similar approach as the tumor cite prediction. However, given that we have two classes of outcomes, that is a yes or no response, we use the elastic-net logistic regression model. This is a special case of the multinomial model when $K = 2$, implying that there are $\mathcal{G} = \{1, 2\}$ classes. Inheriting the hierarchical structure, our prediction $\hat{G}_2 = \Pr(G = k|\hat{G}_1, X)$.

4 Results

4.1 Tumor cite prediction

After training several variations of the elastic-net model (i.e 10 values of α split evenly in the interval $[0, 1]$), and using cross-validation to tune and obtain the best λ, the regularization hyperparameter, we evaluate the models as the best α that minimizes the multinomial deviance. The results from this procedure is illustrated in 2.

Figure 2 shows that $\alpha = 0$ (the fully ridge model) performs worst as it generates the highest deviance. The smallest deviance is achieved at $\alpha = 0.4$. Hence, we select this as the best tuned parameter for predicting the tumor cite, with it's corresponding λ value of 0.007619 tuned from the 10-fold cross-validation.

Using the best tuned parameters, our elastic-net model yields a prediction accuracy of 97.36%. With our multi-class prediction task, coupled with class imbalance, we further provide the confusion matrix (Table 1) which illustrates in detail the performance of the elastic-net multinomial regression model. While the model yields a good prediction accuracy, we observe from table 1 that higher rates of missclassification are attributable to the Lung. Nonetheless, a prediction accuracy of 97.36%, is a rather an impressive score, especially given that with 7 classes, a random assignment would produce about 14% accuracy. This gives us a huge confidence in the fitting the metastasis model, which is a hierarchical model whose performance heavily relies on the accuracy of the tumor cite model.

4.2 Metastasis prediction

Similarly, upon training and tuning for best hyperparameters, the logistic regression model resulted in the selection of a fully lasso model ($\alpha = 1$) with an optimal λ value of 0.0058. Using this model, we obtain a prediction accuracy of 90.33%. Table 2 illustrates the confusion matrix for this prediction. Our model better predicts when there is no metastasis than when there is, with the former yielding a precision of about 91.5% and the latter being 84%.

Table 2. Confusion matrix for metastasis prediction

		Ground Truth		
		No	Yes	Total
Prediction	No	1286	119	1405
	Yes	42	218	260
	Total	1328	337	1665

4.3 Top Influential genes for metastasis

We analyzed the coefficients of the predictors to obtain the top 25 influential genes determined by absolute values of their coefficients and their contribution towards metastasis (fig. 3). Of the top 25 genes, there are 6 genes positively correlated with metastasis events and 19 negatively associated with metastasis.

Of note, we see an abundance of mitochondrially associated genes in the oxidative phosphorylation pathway such as ATP5MD, MT-ND1, and MT-CO1 that are negatively associated with prediction of metastasis. It is known that cancers often have higher metabolism requirements compared to normal cells, and out results suggests that cancers with high metastatic potential may decrease their reliance of oxidative phosphorylation. Additionally, for positively associated genes, we see an increase of AXL, a gene part of the Gas6/AXL pathway associated with invasion and metastasis of cancer.

These results suggest to us that influential genes that our model used to predict metastasis status was in line with known literature, and that it was not influenced by artifacts.

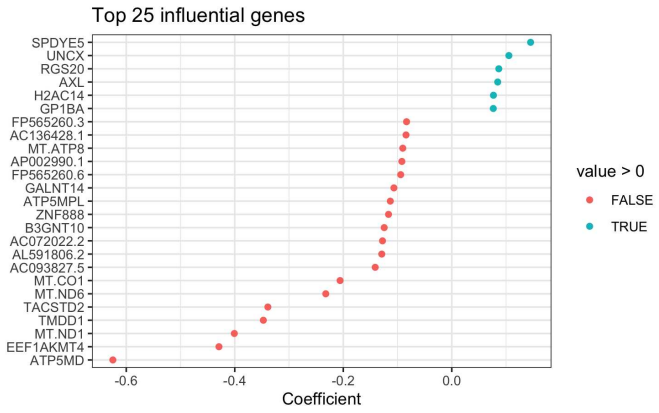

Fig. 3. Top Influential genes for metastasis prediction

4.4 Overall prediction accuracy

Finally, we assess the performance of our hierarchical classification algorithm. While the ultimate goal is to accurately predict metastasis given the tissue source, the hierarchical or conditional model can be rather intricate in it's overall prediction accuracy. Our final prediction could be one of these cases:

1. Accurately predict metastasis given an erroneously predicted tissue source and vice-versa which we code as ***semi-accurate***.
2. Accurately predicting metastasis given accurate prediction of the tissue source, coded as ***accurate***, and
3. Inaccurately predicting metastasis given an erroneously predicted tissue source, coded as ***inaccurate***.

Using this metric, we obtain the overall prediction accuracy of our hierarchical classification algorithm as ***accurate*** $= 97\%$, ***inaccuare*** $= 3\%$ and 0 for ***semi-accurate***, justify how potent our hierarchical classification algorithm is.

5 Discussion

The expression profiles of cancer can be very distinct between tissue of origin as well as between individuals, which makes establishing general trends among different cancers difficult. Our model was able to achieve an accuracy of 97% in predicting tissue of origin and a 90% for predicting whether the tumor has metastasis given the tissue of origin.

It is know that recurrent oncogene mutations are often used as a biomarker in cancer classification. What is unique about our study is the usage of expression profiles without including gene mutations. The complexity of annotating and understanding the effects of different gene mutations makes developing models based on gene mutations difficult for all but the most recurrent mutations. However, by using expression data regardless of mutation status, we provide a model that can be more easily understood and representative of the cell biology of cancer cells.

However, there are many confounders which limits the accuracy of the model. Although we tried to include as many samples as possible across different studies, batch effects within studies may result in differences between samples that were collected as primary and samples that were collected for metastasis studies. Additionally, patients for whom both primary and metastasis samples are sequenced may not represent the whole population of patients with metastasis, as primary tissues of patients with metastasis are typically late-stage cancers while normal primary tissues may be gathered from any stage. These effects may result in those primary samples with metastasis representing severity and development of cancers rather than metastasis potential.

Overall, the results are promising in showing that there may be sufficient evidence in expression profiles of primary tumors that can predict metastasis events. Further studies and incorporation of additional datasets may help with improving the accuracy of the model.

Funding

This project was not funded by anyone in particular.

References

Badal B, Solovyov A, D. C. S. C. J. e. a. (2017). Transcriptional dissection of melanoma identifies a high-risk subtype underlying tp53 family genes and epigenome deregulation. *The Journal of Clinical Investigation*, **2(9)**.

Boehmke, B. C. and Greenwell, B. M. (2019). Hands-on machine learning with r.

Hastie T, Qian, J. T. K. (2021). An introduction to glmnet.

Hunter, K. W., C. N. P. . A. J. (2008). Mechanisms of metastasis. *Breast cancer research: BCR*, **10**, S1.

Kim SK, Kim SY, K. J. R. S. e. a. (2014). A nineteen gene-based risk score classifier predicts prognosis of colorectal cancer patients. *Molecular Oncology*, **8**, 1653–1666.

McDonald OG, Li X, S. T. T. R. e. a. (2017a). Epigenomic reprogramming during pancreatic cancer progression links anabolic glucose metabolism to distant metastasis. *Nature Genetics*, **49(3)**, 367–376.

McDonald OG, Li X, S. T. T. R. e. a. (2017b). Recurrently deregulated lncrnas in hepatocellular carcinoma. *Nature Communications*, **8**, 14421.

Menck K, Wlochowitz D, W. A. C. L. W. A. S. A. K. U. W. S. S. H. B. H. W. E. P. T. H. K. B. T. B. A. (2022). High-throughput profiling of colorectal cancer liver metastases reveals intra- and inter-patient heterogeneity in the egfr and wnt pathways associated with clinical outcome. *Cancers (Basel)*, **14(9)**, 2084.

Riihimäki, M., T. H. S. K. S. J. . H. K. (2018). Clinical landscape of cancer metastases. *Cancer medicine*, **7(11)**, 5534–5542.

Rothwell DG, Li Y, A. M. T. C. e. a. (2014). Evaluation and validation of a robust single cell rna-amplification protocol through transcriptional profiling of enriched lung cancer initiating cells. *BMC Genomics*, **15(1)**, 1129.

Seyfried, T. N., . H. L. C. (2013). On the origin of cancer metastasis. *Critical reviews in oncogenesis*, **18(1-2)**, 43–73.

Siegel MB, He X, H. K. H. A. e. a. (2018). Integrated rna and dna sequencing reveals early drivers of metastatic breast cancer. *The Journal of Clinical Investigation*, **128(4)**, 1371–1383.

Wang, B., Z. Y. Q. T. e. a. (2021). Comprehensive analysis of metastatic gastric cancer tumour cells using single-cell rna-seq. *Sci Rep*, **11**, 1141.

Zou, H., . H. T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology*, **67(2)**, 301–320.