Using Tumor DNA Signatures and Multinomial Naive Bayes Classifiers to Predict Cancer Types

Tiana Pereira

Abstract—Identifying the best treatment for various cancers depends on knowing the primary site of the cancer; however, for the subset of cancer patients whose primary site is unknown, histopathological examination cannot be used for cancer diagnosis. Since cancer stems from genetic alterations, tumor DNA serves as a potential identifier for classifying cancer types. Such genetic alterations include, but are not limited to, somatic mutations and copy number alterations. Support Vector Machines (SVM), in combination with Recursive Feature Elimination (RFE), have been proven to predict cancer types from tumor DNA signature alone, as demonstrated by Soh et al. This work demonstrates the potential and limitations of such alterations through the use of a variety of machine learning models [1]. The aim of this paper is to determine how prediction of cancer types using a Multinomial Naive Bayes classifier compares to these results.

1 Introduction

The main method of diagnosing cancer is histopathological examination, in which tissue biopsies are retrieved directly from the patient. However, this requires knowing the primary site of the tumor. Carcinoma of unknown primary (CUP) is a disease in which malignant tumors are discovered in the body yet the site where they originated from cannot be determined [2]. In such situations, tissue biopsies are unable to be collected, and therefore proper treatment cannot be determined, leading to poor survival rate. [3].

Despite this, genomic analysis has emerged as a potential solution: by identifying genomic characteristics, such as gene expression and DNA methylation and their correlations with various cancer types, researchers have been able to better understand the associations between aberrations and specific cancer types. Previous work has explored the use of machine learning to diagnose cancer types based on gene expression and DNA methylation. This has enabled the potential for cancer diagnosis without the need for invasive methods such tissue extraction.

While promising in theory, there remains the need for further research into how efficient genomic analyses can predict cancer types. While previous research has focused on gene expression data and DNA methylation markers, Soh et al. explores genomic analysis and cancer prediction with respect to somatic mutations and copy number alterations. The reasoning behind investigating these two types of genomic alterations is that some types of cancer, including breast and colorectal cancer, typically possess somatic point mutations as the most common form of genomic alterations, which is usually enough to identify these two cancers [1]. However, there is evidence that some cancers mainly possess mutations in the form of copy number alterations; these alterations may provide better insight into identifying more cancer types beyond only examining somatic mutations.

Soh et al. found that analyzing copy number alterations and somatic mutations together can lead to more predictive power in comparison to classifying cancer types based on copy number alterations or somatic mutations alone. Through the use of Support Vector Machines

(SVMs) with a linear kernel, combined with Recursive Feature Extraction (RFE), the authors were able to achieve 77.7±0.3% accuracy with as few as 50 genes and 88.4±0.2% with 900 top-ranked genes containing both copy number alterations and somatic mutations.

The authors only analyzed three classification models with separate feature selection methods – an SVM-RFE, L1-regularised logistic regression, and random forest classifiers using a selection method presented by Díaz-Uriarte et al.[3] The latter two did not perform as well as the SVM-RFE.

One model that they did not use was a naive Bayes classifier and in particular, the Multinomial naive Bayes (MNB) classifier. In general, naive Bayes classifiers are relatively simple, quick models in comparison to SVCs that utilize "one-versus-one" comparisons. They take discrete values for input, such as word counts for text classification, and calculate conditional probabilities of observing features between classes. Predictions are then decided using these conditional probabilities. In this project, I aim to determine whether a Multinomial naive Bayes Classifier can achieve similar results to an RFE-SVM as used in Soh et al. I also intend to examine the relationship between somatic mutations and copy number alterations with regard to predicting cancer types.

2 METHODS

2.1 Data Collection and Preprocessing

A total of 6640 tumor samples from 28 cancer types were analyzed for both the Soh et al. study and this paper (the distribution of sample sizes can be found in their paper [1]). The raw data consists of counts of somatic mutations and copy number alterations that appeared in a gene for each sample, across 28 studies of each cancer type. This data was downloaded from cBioPortal for Cancer Genomics [4] [5], and the preprocessed data for somatic mutations and copy number alterations is made available by Soh et al [6]. The authors further filtered the data by removing pseudogenes, non-coding genes, and genes that had no mutations at alls.

Both the somatic mutation and copy number alteration datasets were binary matrices, such that in i-th row (representing the samples) and in the j-th column $A[i, j] = 1$, there was a mutation in that particular gene, and $A[i, j] = 0$ otherwise. While the somatic mutation data only contained one column per gene, the copy number alteration dataset included two columns per gene: one to account for amplification and the other for deletion (labels representing the columns included the gene name and the suffix ".p" and ".n" for amplication and deletion, respectively). Rows and columns containing only zeros were filtered out, as they indicate samples with no gene aberrations and genes with no mutations.

While the authors used a total of three datasets - somatic mutation data, copy number alteration data, and the combination of these two - I only utilized the concatenated matrix of the somatic mutation data and copy number alteration data. This resulted in a binary matrix of 6640 samples and 13151 features.

2.2 Multinomial Naive Bayes Classifier Model and Feature Selection

The Multinomial Naive Bayes classifier is typically used for text classification and works by calculating the probabilities of features occurring within a given class. It usually require discrete values for features, which fits the somatic mutation and copy number alteration data format since each element of the matrix includes a discrete value of 0 or 1.

These counts are used to calculate conditional probabilities of observing a count of some feature x that appears in a class c. The classifier performs these calculations according to the naive Bayes algorithm for multinomial data [7][8]. In the context of gene alteration data used in this project, the classifier calculates the

conditional probabilities of seeing each somatic and/or copy number alteration within a gene that appears in each class.

The MNB classifier was implemented through the `MultinomialNB` class from `sklearn`. The parameters for such class include additive (Laplace/Lidstone) smoothing parameter, whether to use a uniform prior or to learn the class prior probabilities, and the option to specify the class prior probabilities [9]. Since the dataset contains many zeros, the additive parameter was set to 1. The classifier was also set to learn the prior probabilities since no class prior probabilities were calculated and passed to the classifier.

In addition to using various models, Soh et al. utilized forms of feature selection to reduce the number of features needed in the classification. There were two reasons behind this: first, having extra features that do not contribute any useful information towards towards the final prediction create noise that can hinder the model. Secondly, the authors mention that in a practical sense, reducing the number of genes that are sequenced and analyzed ultimately lowers the cost for cancer patients.

The feature selection performed with the MNB classifier uses `SelectKBest` from `sklearn`'s `feature_selection` package. `SelectKBest` is a filter-based method of filtering unnecessary features using a scoring function. A χ^2 test was implemented using the `chi2` method from `sklearn`'s `feature_selection` package to select the highest-scoring features.

2.3 Splitting the Dataset and Training the MNB Classifier

In order to prevent overfitting, 25% of the dataset, 1661 samples, was reserved as a validation set. During training, the remaining 75% was split into training and testing sizes of 80 and 20%, respectively. This is the same method that was used in Soh et al. Due to the uneven distribution of cancer types in the dataset, splitting was stratified such that each set of training and testing, as well as the validation, would contain an equal proportion of cancer types.

The first step in training the MNB classifier was to first select the best features. Features were ranked using `SelectKBest` and the χ^2 test to sort the genes based on their importance, with higher-scoring genes appearing at the beginning. This was repeated 50 times, and the scoring for each iteration was stored such that by the end of the iterations, genes could be selected based on their importance across all iterations. This was to ensure that the gene ranking did not depend on one randomly sampled training set.

Next, the MNB is fitted and trained according to an increasing number of top-ranked genes, starting with the highest-ranked gene and ending with the top 4500 genes as scored by the χ^2 test. The MNB models then underwent prediction, and the resulting accuracy was stored for each set of genes. This method replicates the analysis of classification performance as in Soh et al. [1]. While the authors decided to repeat this process 50 times and take the average over all such iterations, I only performed this accuracy collection once due to restrictions in storage and time.

2.4 Examining an Optimal Set of Features

Once the accuracy for each set of genes was collected, the sets of genes that yielded the best results, as well as their associated features, were then fitted and trainged again through the MNB for further investigation. This was repeated 100 times, and the overall accuracy, precision, and recall scores for each class were collected.

To investigate the impact of various mutations on cancer predictability, the distribution between mutation types across the top 900 ranking genes were counted.

3 Results

After running the χ^2 test with *SelectKBeset*, genes were ranked according to their importance. 18 of those genes coincided with the top 50 genes found in Soh et al: these were included, but not limited to, *BRAF*, *IDH1*, and *KRAS*. This demonstrates that the model was able to capture some oncogenes, however the model failed to replicate the importance measurements of Soh et al, suggesting the model's limitations.

The MNB classifier was then trained according to increasing numbers of the top-rated genes. The 4500 top-scoring genes were used in the assessments, as displayed in Fig. 1. Smaller sets of less than a thousands genes tended to have more noises, as indicated by the drastic changes in accuracy as the number of genes increased. This is in part due to not sampling each set of genes, which can be implemented in future work. It was only until roughly 800 genes were used that an overall increase in accuracy was observed. It continued to increase at a slower rate until after 3000 genes were used. Despite this observation, the performance of the MNB classifier only reached at most roughly 64.05%. This poor accuracy is singifancly lower than the accuracy achieved from the SVM-RFE that was implemented in Soh et al.

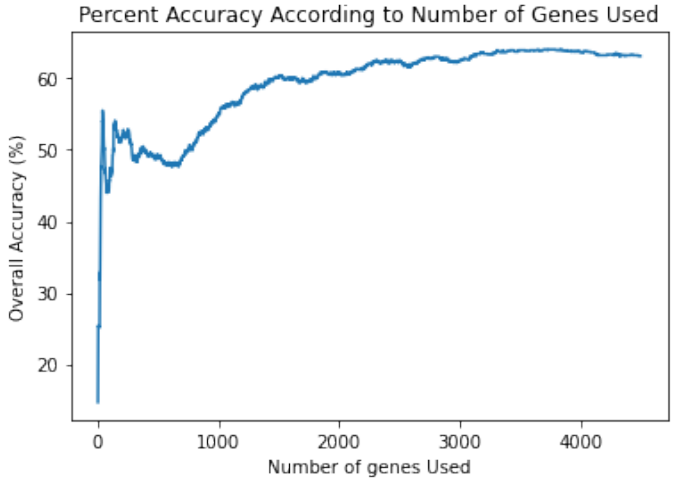

Fig. 1. Overall accuracy across an increasing number of genes used in fitting the MNB classifier.

Using the smallest number of genes that yielded this accuracy, the model was trained with all features relating to these genes, and the precision and recall scores were calculated for each class (Fig. 2) in the reserved validation set. This training occurred 100 items, and the average scores were taken across all results. As anticipated, classes with greater distributions resulted in greater scores for precision and recall, as indicated by kidney renal clear cell carcinoma, which had a total of 418 samples. Likewise, the precision and recall scores for uterine carcinosarcoma are extremely low, since the sample size for this cancer type was merely 55 samples. There are still some counterexamples to the correlation between class distribution and scoring: breast invasive carcinoma, which had a total sample size of 973, does not have very high and is even lower than kidney renal clear cell carcinoma.

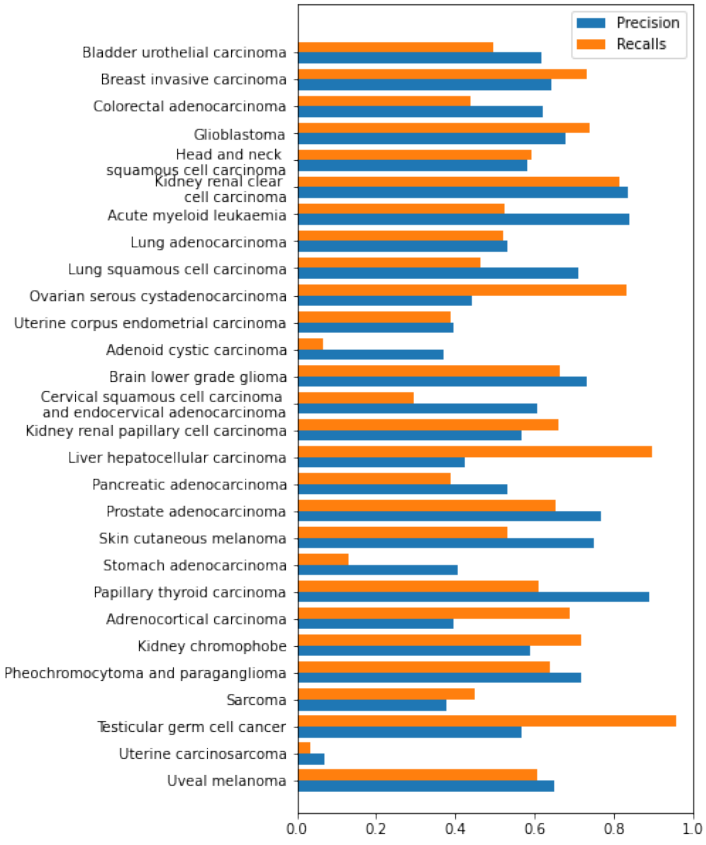

Fig. 2. Resulting precision and recall scores for all classes in the reserved validation set.

Another observation of interest was the fre-

quency of different mutations in top-scoring genes. As was found in Soh et al., the number of copy number alterations in the most influential genes for cancer classification was greater than the number of somatic mutations (Fig. 3). In fact, the frequency of copy number alterations was 1.7 times the number of somatic mutations. This is particularly interesting as this provides insight into the relationship between copy number alterations and genes that influence cancer. This also further demonstrates, as was shown in Soh et al., that there is a benefit to analyzing copy number alterations in cancer diagnosis.

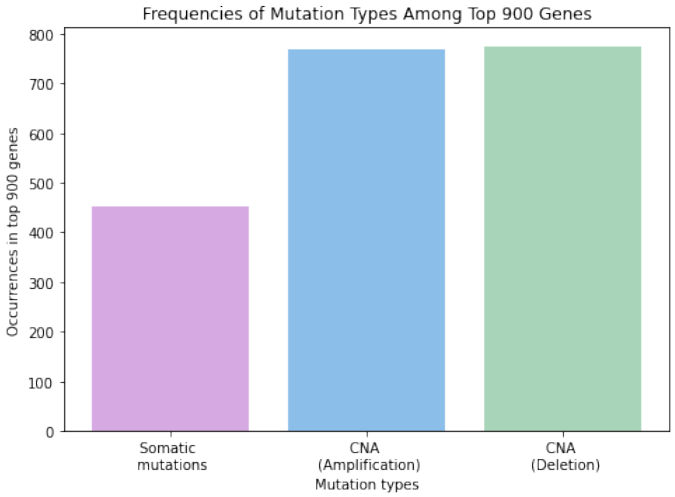

Fig. 3. Distribution of somatic mutations versus copy number alterations (CNA) for both amplification and deletion across the 900 top-ranked genes. As displayed, there is a significant difference between the number of CNA mutations versus somatic mutations, revealing the importance of analyzing such mutations for cancer diagnosis.

4 Discussion

In an effort to determine whether MNB classifiers – combined with feature selections calculated through χ^2 – could outperform SVM-RFE, it is clear that the MNB classifier fails. This is both in respect to accuracy, as well as precision and recall scores for each class (more information in Soh et al.)[1].

One of the drawbacks of using a MNB classifier is that the Naive Bayes algorithm assumes independence between all sets of features, which often does not translate to realistic datasets such as the somatic mutation and copy number alteration datasets.

In utilizing univariate feature selection methods such as χ^2 tests, the importance of features is strictly measured by relatively simple statistics, unlike the the Recursive Feature Elimination utilized paired with SVMs in Soh et al. For future work, it would be interesting to determine how well other feature selections work with naive Bayes classfiers. As an example, Chen et al. present two filtering methods, *Multi-class Odds Ratio* (MOR), and *Class Discriminating Measure* (CDM) to evaluate features for multiclass text classification. These methods adopt log-odds ratios for naive Bayes binary classification – a filtration method that has proven to be particularly efficient – and adapt them for multiclass classification [10].

Additionally, better prediction can potentially be achieved through using deep learning models, such as neural networks that undergo more thorough evaluations. However, as is mentioned in Soh et al., such models require larger datasets than the current dataset. This could be solved through gathering more samples and data, and the pay-off may be worth the extra time collecting data. There are currently neural networks that analyze gene expression profiling, such as in Khan et al. [11]. It would be interesting to see how DNA signatures such as the frequencies of somatic mutations and copy number alterations play a role into prediction.

Despite the shortcomings of the MNB classifier, there are interesting insights into the frequency of copy number alterations in top-ranked genes. Due to difference in these mutations versus somatic mutations, it would be particularly interesting to use the number of appearances of copy number alterations in predictive genes instead of simply a binary classification as to whether these mutations appeared or not. If there is such an advantage, it has the potential to greatly improve cancer diagnosis from tumor DNA samples.

5 Conclusion

Based on these findings, the Multinomial Naive Bayes classifier does not perform as well the SVM-RFE implemented in Soh et al. However, more thorough testing can potentially lead to more precise results. For example, including more random sampling over testing increasing sizes of gene sets and averaging over those results can lead to a less noisy performance analysis. Additionally, there is further opportunities in choosing feature selection methods that are more efficient than selecting k features across the χ^2, such as the ratios provided by Chen et al [10]. With investigating various machine learning methods and increased sampling of tumor DNA, the accuracy in predicting cancer types can certainly lead to more improved diagnostic tools for patients with CUP.

Data Availability

The data used in this project can be found in the cBio Cancer Genomics Portal [4]. The code used in this project is available here: Project Code.

References

[1] Soh, K., Szczurek, E., Sakoparnig, T. et al. *Predicting cancer type from tumour DNA signatures*. Genome Med 9, 104 (2017). https://doi.org/10.1186/s13073-017-0493-2

[2] PDQ® Adult Treatment Editorial Board. PDQ Carcinoma of Unknown Primary Treatment. Bethesda, MD: National Cancer Institute. Updated 10/23/2019. Available at: https://www.cancer.gov/types/unknown-primary/patient/unknown-primary-treatment-pdq. [PMID: 26389238]

[3] Pavlidis N, Pentheroudakis G. *Cancer of unknown primary site*. Lancet. 2012; 379(9824):1428–35
.

[4] E. Cerami, J. Gao, U. Dogrusoz, B. E. Gross, S. O. Sumer, B. A. Aksoy, A. Jacobsen, C. J. Byrne, M. L. Heuer, E. Larsson, Y. Antipin, B. Reva, A. P. Goldberg, C. Sander, and N. Schultz, "The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data," *Cancer Discovery*, 01-May-2012. [Online]. Available: https://cancerdiscovery.aacrjournals.org/content/2/5/401.

[5] Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, Sun Y, Jacobsen A, Sinha R, Larsson E, Cerami E, Sander C, Schultz N. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal. 2013 Apr 2;6(269):pl1. doi: 10.1126/scisignal.2004088. PMID: 23550210; PMCID: PMC4160307.

[6] N. Beerenwinkel, "Somatic point mutation and copy number alteration matrices used in Soh et al., 'Predicting cancer type from tumour DNA signatures', Genome Medicine," Somatic point mutation and copy number alteration matrices used in Soh et al., "Predicting cancer type from tumour DNA signatures", Genome Medicine - Research Collection, 08-Nov-2017. [Online]. Available: https://www.research-collection.ethz.ch/handle/20.500.11850/206154.

[7] "1.9. Naive Bayes," scikit. [Online]. Available: https://scikit-learn.org/stable/modules/naive_bayes.html.

[8] Artificial Intelligence - All in One. *Lecture 29 - Multinomial Naive Bayes A Worked Example - [NLP —— Dan Jurafsky —— Stanford]*, YouTube, 11-Aug-2019. [Online]. Available: https://www.youtube.com/watch?v=j1uBHvL6Yr0amp;t=210s.

[9] "sklearn.naive_bayes.MultinomialNB," scikit. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.Multinomia

[10] Jingnian Chen, Houkuan Huang, Shengfeng Tian, Youli Qu. Feature selection for text classification with Naïve Bayes, *Expert Systems with Applications*. Volume 36, Issue 3, Part 1, 2009, Pages 5432-5435, ISSN 0957-4174. doi: /10.1016/j.eswa.2008.06.054. Available: https://www.sciencedirect.com/science/article/pii/S095741740800356

[11] Khan, J., Wei, J., Ringnér, M. et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat Med 7, 673–679 (2001). https://doi.org/10.1038/89044

[12] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.