Discovering Bladder Cancer Biomarkers With Machine Learning

Joel Rorseth
School of Computer Science
University of Windsor
Windsor, Ontario
Email: rorsethj@uwindsor.ca

Michael Bianchi
School of Computer Science
University of Windsor
Windsor, Ontario
Email: bianchi1@uwindsor.ca

December 13, 2017

Abstract

In an attempt to better understand the genetic nature of cancer, researchers are employing machine learning algorithms to aid in identifying patterns and correlations invisible to the human eye. In this paper, we aim to analyze and learn using sample data from patients experiencing various stages of bladder cancer. We begin by discussing the difficulties in using the cancer dataset, and how the imbalance or unequal distribution of the samples makes accurate predictive classification extremely difficult. Due to the large range of specific bladder cancer stages, we determine that correlation is infrequent and at times illogical. More insight between specific cancer subtypes is, in comparison, much more accurate, and is the basis for our proposed solution to the multi-class problem. With the help of Feature Importance algorithms, Random Forest classification models and some ingenuity, we can identify genes that are statistically characteristic of the various subtypes of this form of cancer. By treating the cancer stages as classes for the genetic data, almost sixty thousand sampled genetic biomarkers can be ranked as important indicators or completely meaningless. Moreover, we investigate the procedure of dimensionality reduction, which in combination with feature selection, yields the hypothesis that a shocking amount of the given genetic data is statistically uncorrelated with any type of bladder cancer.

1 Introduction

With the intent of discovering insight into the nature of bladder cancer, this paper presents a research findings using machine learning analysis and techniques. Using a dataset of genetic biomarkers from a small sample of diagnosed patients, we allow mathematical algorithms to determine important information easily overlooked by the human eye. In our research, we will use the Scikit-Learn and Weka machine learning libraries and tools to format, preprocess, feature select and classify the samples in question. This research aims to solve the problem of identifying meaningful biomarkers as they relate to certain stages (classes) of bladder cancer. Along the path to attaining this information, our research is met with several smaller, machine learning dilemmas, to be discussed. We will consider the machine learning feature selection procedure to be instrumental in discerning the important biomarkers, while these smaller problems including classification, rebalancing and the multi-class problem will help to support and obtain that information.

2 Bladder Cancer

In order to fully understand the nature of the bladder cancer dataset we will be analyzing, we must first establish an interpretation and definition of bladder cancer and its identifiable forms. Bladder cancer is a specific form of cancer, defined as a deadly disease involving abnormal cell growth that may spread and invade others parts of the human body [1]. Bladder cancer, in specific, originates from the bladder of the subject. When a patient has been diagnosed with bladder cancer, doctors begin the process of staging, the task of determining how far the cancerous cells have spread in the body. Furthermore, a specific stage of the cancer is established, representing the extent of the spread [2]. The stage is a vital factor in deciding treatment and likelihood of success, and is thus understood to be a valid classification of bladder cancer samples. For the scope of this paper, samples of bladder cancer subjects are categorized and labelled by their specific, standardized stage. The given dataset presents few samples, with various different specific stages. Before we can employ machine learning techniques to analyze the data, proper interpretation and formatting must be explored in the context of the bladder cancer data.

3 Interpreting the Dataset

From a logical standpoint, the bladder cancer dataset contains genetic samples from sixty patients suffering from the disease. For each patient sample, the stage of their cancer is given according to the TNM staging system established by the American Joint Committee on Cancer [2]. Approximately fifty-eight thousand of the same, labelled genetic biomarkers are observed for each sample. The values given for each biomarker represent the amount of cellular activity occurring at the molecular level. With this data, the remainder of this paper

focus on interpreting the genetic information, looking for correlations between stages and certain biomarkers, and determining which genes have no influence in particular stages. The stages assigned to each sample describe themselves using the letter T, which in combination with a given number, describes how far the original tumour has grown through the bladder wall (possibly into nearby tissues) [2]. Each T sample gives several alphanumeric characters to describe itself. However, in an attempt to find commonality and correlation between the samples, we must group them into sufficiently large groupings. To perform the analysis detailed in this paper, samples were grouped into the following three dominant subgroups:

1. Ta - Non-invasive papillary carcinoma, where growth has reached hollow center of bladder but not the connective tissue or bladder wall.
2. T1 - Growth into the layer of connective tissue under lining layer of bladder, but not reaching the bladder wall muscle.
3. T2 - Growth has reached the inner or outer muscle layer of the bladder wall, but has not breached the fatty tissue surrounding bladder.

With each successive stage in the sequence Ta, T1, and T2, the observed cancerous cells have spread more.

3.1 Formatting

To format the data, the labels (stages) of the sixty classes were truncated to simplify the classification to the three groups aforementioned. Of the sixty samples, twenty-seven are Ta, twenty-three are T1 and eight are T2 stage. Two Ti samples were discarded due to their lack of representation and decision of scope for the machine learning analysis. In our research, we have utilized the Scikit-Learn machine learning library for the Python programming environment. As such, the primary dataset is read by our program, which reorganizes the into an array of genes (the machine learning features or parameters), a two-dimensional array of sample data (samples) and an array of cancer stages (labels). The NumPy scientific computing package for Python is used, allowing the data to be stored, written to and manipulated in the most efficient manner. To add perspective, the Weka machine learning tool was used in tandem to verify and perform operations. The performance and results throughout this paper will cite the tool used to obtain the data.

4 Inherent Dataset Problems

Upon initial observation, the dataset is quite irregular. With only 58 samples, the feature space containing 57,820 features is overwhelmingly large. From a machine learning perspective, training an algorithm with limited samples incurs bias and variance in the performance of the classifier, relative to the same

process trained with an infinite sample size [3]. Moreover, this dilemma is a limiting factor in determining the most effective features. To further complicate matters, the distribution of classes among all 58 samples is imbalanced. In machine learning, much bias can be introduced when this issue is unaccounted for. This class imbalance leads to overfitting, due to a lack of representation for all classes equally. To work around this, our research investigates different approaches to handling this multi-class problem. Most importantly, by tackling these problems, we are able to address the primary objective of this research: identifying meaningful biomarkers and their influence in bladder cancer stages.

4.1 Curse of Dimensionality

Undoubtedly, the most challenging aspect of analyzing the bladder cancer dataset is the overshadowing feature space size in comparison to the number of samples. It is known that machine learning classification ideally expects the lowest feature space size to sample size ratio possible. When this ratio is much larger than 1.0, the problem is susceptible to The Curse of Dimensionality (Bellman, 1961). This designation refers to the fact that the convergence of any estimator to the true value of a smooth function defined in high dimensional space is extremely slow, taking exponentially longer with larger dimensional space [4]. The bladder cancer dataset will unquestionably suffer in classification performance as a consequence of this problem. More evidence of this issue is discussed in the Classification section. In the established study of machine learning, few accepted methods exist to aid in handling this circumstance. Although no single method may necessarily improve classification accuracy, solutions may include the following:

- Feature selection - The process of selecting an optimal subset from the original feature set.
- Dimensionality Reduction - The process of reducing or compressing the feature space, mapping a higher dimensional space onto a lower one.
- Gathering more sample data to increase the sampling size

Thanks to the importance of feature selection in finding meaningful biomarkers, we investigate feature selection not only as a method of reducing the feature space, but in ranking the most important features. This feature ranking and selection process is discussed in the Feature Selection Problem section.

4.2 Imbalanced Classes

Similarly to the uneven ratio of features to samples, the poor distribution of represented classes in the given samples also hinders the performance of classification and machine learning analysis. This situation refers to the sample data as being imbalanced, meaning that the samples do not represent all classes

```
SVM C = 0.1051  Classification on original dataset
Average accuracy over CV runs: 0.467948717949
Confusion Matrix:
 [[ 0  0 23]
 [ 0  0  8]
 [ 0  0 27]]
```

Figure 1: A confusion matrix from a sampled run of the SVM classifier with RBF kernel function. The vertical axis represents the true classes, while the horizontal axis represents the corresponding classes predicted by classifier.

equally [5]. In our bladder cancer dataset, there are a significantly smaller number of samples classified as T2 (only 8), in comparison to the relatively similar numbers in the remaining classes (23 and 27). Due to this misrepresentation, most classifiers will optimize accuracy by predicting most samples to be of the class which is best represented, thus yielding the highest performance accuracy. There exists several well known countermeasures to deal with this scenario [5]:

- Collect more data to boost under-represented classes
- Additive or subtractive dataset resampling
- Generating synthetic samples
- Try classifiers that work better with imbalanced data
- Penalized classification

This problem is the basis for discussion of the Multi-Class Problem, to follow. In an attempt to achieve notable accuracy, several of the aforementioned techniques are documented in the Classification section. In addition to classification of all classes, our research attempts to evaluate between-class classification performance using subsets of the original data. Thus, a deeper understanding of relationships between specific cancer stages is presented.

4.3 Overfitting

Due to the lack of class distribution and sample size, the bladder cancer dataset is prone to overfitting. Overfitting occurs when, due to a lack of broad, generalized data for each class, a model is fit too closely to be considered accurate. In the dataset, the overshadowing of Ta samples largely influences most models to classify all samples as this class. For example, in Figure 1, we see the performance evaluation of an SVM RBF classifier through accuracy score and confusion matrix. In this example, the overfitting is visible by the fact that every sample has been classified as the majority class. The accuracy corresponds almost exactly to the percentage of these majority samples in the original dataset. With more investigation, we determined that this problem still exists despite drastic dimensionality reduction and feature selection, consistently overlooking the minority T2 samples.

5 The Multi-Class Problem

As described in Interpreting the Dataset, our research examines the bladder cancer samples from the perspective of belonging to one of three classified stages. Classification of the dataset now falls into the category of being a Multi-Class Problem. This is defined as being a machine learning problem that is to analyze a dataset where samples each belong to a single class, but there exists more than two classes across all samples [6]. Solving this problem is one of the primary objectives of our research. We'll consider two popular solutions that attempt to solve this problem by essentially transforming the dataset into a binary (class) problem:

1. One-Vs.-All - Train a single classifier on each class, treating samples of that class positive and other negative.
2. One-Vs.-One - Fit one classifier for every possible pair of classes, while using a voting algorithm to decide class.

With the Scikit-Learn library, all classifiers are already equipped to perform multi-class classification [7]. Although we will evaluate and compare performance of these inherently multi-class classifiers, we will also utilize the *sklearn.multiclass* class to wrap regular classifiers in meta-estimators corresponding to the mutlti-class solutions listed above. This class facilitates improved accuracy for multi-class problems by allowing One-Vs.-All and One-Vs.-One algorithms to parameterize their base classifiers. This solution is described in the Classification Problem section.

5.1 Fixing the Imbalance

As a first step in achieving reasonable classification accuracy, the distribution of samples had to be equalized by class. In the Scikit-Learn research, a custom algorithm was written to implement a typical SMOTE (Synthetic Minority Over-Sampling Technique) resampler. This method duplicates minority samples to normalize class distribution, while avoiding any bias of allowing samples to co-exist in the cross-validation train and test sets simultaneously. In the Weka experiment, the Resample Filter was used to perform this operation. Immediately following application, the resampling increased classification accuracy by almost 30% for the tested classifiers. As seen in Table 2, the superior Random Forest classifier saw an increase from 51.72% to 84.48% by this process alone This experiment pitted five unique classifiers against each other, showing their accuracy in three stages. The dramatic increase is a testament to the severity of this issue in our bladder cancer dataset, however it proved to be easily accounted for.

5.2 Inter-Class Separation

In an alternative take on the multi-class problem, the dataset was manually separated into binary class sets, where the dataset would contain samples belonging

to only two classes. Specifically, new datasets were produced for T1 vs. T2, and Ta vs. T1. This approach allows important insight to the classification problem, where we may better understand the difference between cancer stages more effectively than when considered together. More importantly, inter-class separation allows us to discover meaningful biomarkers via feature selection from reduced sample space. By reducing to binary problems, the biomarkers that most significantly identify and differentiate between two stages of bladder cancer are effectively calculated. In Table 3, the accuracy of the eventual Random Forest classification of these cases is given. Although we obtain fair accuracy when evaluated over all three classes, increased accuracy and insight is gained by taking this explorative approach.

6 Feature Selection Problem

Combined with the role of resampling, feature selection is undoubtedly the most integral component of accurate classification using this dataset. With almost sixty-thousand features, classifiers will have difficulty correlating and finding patterns. Feature selection is the method by which we chose to reduce our feature space, in in doing so, determine the most important biomarkers. To facilitate optimal selection, scoring algorithms were chosen and evaluated comparatively to determine the best ranking.

6.1 Information Gain

The clear winner of those selected, Information Gain is able to produce a rank for every feature in the feature space, specifying how important it is in the context of the dataset [8]. Using this algorithm, the feature set scored 252 non-zero scores, effectively eliminating almost every feature. From these, various combinations were tested. After much research, the top three were selected and used. As seen in the figure below, Information Gain yielded the best classifier accuracy when applied.

6.2 Pearson Correlation

Using the Weka *CorrelationAttributeEval* class, we attempted to rank features by means of Pearson's correlation coefficient. This is a popular ranking algorithm in Weka, and it performed well when applied to the dataset. Of almost 58,000 features, Pearson Correlation effectively reduced the feature space to 8602 features by scoring the remaining features 0.

6.3 Recursive Feature Elimination

Using Scikit-Learn's RFE ranker from the $feature\_selection$ class, a list of the top ranked features was calculated over a lengthy amount of time. This algorithm works by recursively considering increasingly smaller sets of features

Table 1: Feature Selection Comparison (Random Forest Classifier)

Scoring Method	Optimal Top K	Top K Ranked Genes	Accuracy
Information Gain	3	ENSG00000143970.12 ENSG00000169756.12 ENSG00000130299.12	91.37%
Pearson Correlation	6	ENSG00000165055.11 ENSG00000175414.6 ENSG00000117868.11 ENSG00000140830.4 ENSG00000168259.10 ENSG00000100393.9	86.21%
Recursive Feature Elimination	3	ENSG00000005156.7 ENSG00000005810.13 ENSG00000029363.11	74.14%
Chi^2	3	ENSG00000065361.10 ENSG00000113522.9 ENSG00000143799.8	75.86%

for eventual selection, by means of recursion [9]. This algorithm took the longest of all, and in turn, lacked in applied classification accuracy.

6.4 Chi^2

As a final scoring metric, the Chi^2 scoring function was used in our Scikit-Learn analysis. After the top features were generated, they too were judged alongside others for classification accuracy. This function was chosen due to its ability to measure dependence between stochastic variables [10], meaning it will only work with only non-negative features.

6.5 Evaluating the Best Scoring Algorithm

In a matter of extreme importance, the scoring method selected is the primary factor in the decision of important biomarkers. To compare these functions, we put them to the test in a Random Forest classifier, with the dataset adjusted for resampling. Using the confusion matrix and accuracy scores, Information Gain was determined to be the most reliable and accurate scoring method for the feature selection process. Moving forward, Information Gain's feature ranking is applied in all three cases of the multi-class problem. After much testing, it was decided that the top three features from this ranking be selected for optimal accuracy. The results of our comparison are given in Table 1, along with the relative performance and selections for each.

Table 2: Random Forest classification accuracy in preprocessing stages

	No Resampling No Feature Selection	Resampled No Feature Selection	Resampled Feature Selected
Random Forest	51.72%	84.48%	91.38%
SVM RBF	46.55%	81.03%	81.03%
SVM Linear	29.31%	74.14%	46.55%
Naive Bayes	46.55%	65.52%	67.24%
Bagging	50%	68.97%	74.14%

7 Classification Problem

Several attempts at improving initial classification of the dataset were made through using a variety of classifiers. Support Vector Machine (SVM), Random Forest (RF), K Nearest Neighbour (KNN), and numerous tree classifiers were tested. Although initial attempts before preprocessing produced shameful results, it was clear even at this point that SVM and RF were likely most effective. By means of preprocessing, performance was increased to impressive levels. Before rebalancing, SVM classified with constant 43% accuracy due to overfitting. Due to the imbalance problem, most classifiers would originally classify all samples as Ta, the largest class. Several SVM kernels are detailed below, however their performance almost never improves after feature selection. To provide context, Random Forest and SVM RBF were compared against several other relevant classifiers

7.1 SVM

Upon initial classification with no feature selection or resampling, SVM classifiers yielded a constant accuracy, regularly below 50%. This was also the case with the linear kernel SVM. These results make sense for the dataset because there are three distinct class, and without manipulation, RBF and Linear SVMs (compatible only with One-vs.-One scheme) are unable to properly classify. A number of problems were tackled in order to produce better results. As discussed in the Feature Selection section, ranking algorithms were used to obtain a small amount of meaningful biomarkers. After this, the samples were rebalanced (as described) to address the imbalance problem. Linear and RBF SVMs were then tested using a One-vs-One algorithm. Notably, the linear SVM was able to obtain an accuracy of 83.4%, outperforming the RBF. Although surprising, this occurred because linear kernels are the only inherently compatible kernels for SVM in Scikit-Learn's feature selection class. Moreover, after successfully and drastically reducing the feature space, the data was more simply split in a linearly split fashion than before this procedure.

Table 3: Meaningful Biomarkers determined using Random Forest, Information Gain

Classification	Top Biomarkers	RF Accuracy
Ta vs. T1 vs. T2	ENSG00000143970.12 ENSG00000169756.12 ENSG00000130299.12	91.38%
Ta vs. T1	ENSG00000231822.1 ENSG00000051180.12 ENSG00000255438.2	92.00%
T1 vs. T2	ENSG00000156968.8 ENSG00000157764.8 ENSG00000236051.2	96.77%

7.2 Random Forest

Among others, Random Forest (RF) was evaluated on identical conditions to other classifiers. Without any doubt, classifiers from Scikit-Learn and Weka's *tree* class outperformed all other classifiers at all stages of the research. With Scikit-Learn, several tree classifiers were tested from this class. Shown below, the Extra Trees Classifier (ET) yields accuracy of 83.4% on resampled, feature selected data. Due to their ability to fit a number of randomized decision trees using samples and averaging, tree classifiers are able to regularize to compensate for overfitting in a way other classifiers will not. Using the Weka environment, the Random Forest classifier with default parameters was determined to be the most accurate and correct classifier of all throughout the research. Using the Weka resampling filter and Information Gain ranking discussed above, 91.38% accuracy was achieved, as seen in Table 3.

7.3 Performance

As shown in Table 2, Random Forest classifier performs the best at all stages of the processing pipeline. After resampling and paired with Information Gain, it has proven to be the most accurate. More evidence of this can be seen in the screenshots and source files from the research. Having established the best classifier, resampler and feature selection algorithm for optimal performance, we now have the tools necessary to extract and calculate the meaningful biomarkers from the original bladder cancer dataset.

8 Important Biomarkers

To obtain the most accurate perspective on the most important features, several processes have taken place. As discussed, the data was preprocessed by resampling to balance the class distribution. The multi-class problem tasked several scoring methods with determining a ranked list of features for classification of

Table 4: Confusion Matrix for Random Forest classification over all classes

		Predicted Class		
		T1	**T2**	**Ta**
Original Class	**T1**	21	0	2
	T2	0	8	0
	Ta	2	1	24

all three classes, T1 vs. T2 and Ta vs. T1. To evaluate their choices, several classifiers were run over these three datasets in a competition of accuracy and performance. It was determined that the Information Gain (IG) ranking algorithm classified using Random Forest (RF) yielded the best classification accuracy across all multi-class cases, as verified with 10-Fold Cross Validation. The actual classification run over this cross validation was superb, only miscalculating four samples from the original sample size of fifty-eight as seen in Table 4. Furthermore, different amounts of features were run from the highest ranked. We concluded that the three highest ranked features were sufficient and optimal in all classification cases. Thus, we present the following highest ranking genes (biomarkers) for each of the three corresponding multi-class cases, evaluated separately using Information Gain:

Knowing that classification performance was sufficiently high for each multi-class case, as shown in Table 3, we concur that these biomarkers are unquestionably indicative of their respective bladder cancer stages. When looking at the associated Information Gain scores of these top picks, they are clear standouts from the other samples, the vast majority of which are determined to be completely useless. These files may be viewed in the source documentation, and show how the top three features are consistently higher than all others in score.

9 Conclusion

Thanks to the mathematical power and intelligence behind the Scikit-Learn and Weka tools, the unfavourable bladder cancer dataset has been transformed and successfully analyzed. Through preprocessing techniques such as resampling, class separation and feature selection, a naturally overpopulated feature space containing 58,000 attributes was reduced to a superior representation using only 3. Using the Random Forest classifier, we are able to easily shift perspective on the data, immediately identifying biomarkers that have the highest correlation to their cancer stages. It goes without saying that the filtering capabilities of machine learning can completely unlock otherwise useless data.

Bibliography

[1] World Health Organization (2010, Dec.) "Cancer Fact sheet N297". [Online.] Available: http://www.who.int/mediacentre/factsheets/fs297/en/

[2] American Cancer Society (2017, Dec.) "Bladder Cancer Stages". [Online.] Available: https://www.cancer.org/cancer/bladder-cancer/detection-diagnosis-staging/staging.html

[3] T. W. Way, B. Sahiner, L. M. Hadjiiski, H. Chan (2010, Jan.) "Effect of finite sample size on feature selection and classification: A simulation study". [Online.] Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2826389/

[4] UCLA Statistics Department (2001, Sep.) "Curse of Dimensionality?". [Online.] Available: http://www.stat.ucla.edu/ sabatti/statarray/textr/node5.html

[5] J. Brownlee (2015, Aug.) "8 Tactics to Combat Imbalanced Classes in Your Machine Learning Dataset". [Online.] Available: https://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/

[6] S. Jain (2017, Aug.) "Solving Multi-Label Classification problems (Case studies included)". [Online.] Available: https://www.analyticsvidhya.com/blog/2017/08/introduction-to-multi-label-classification/

[7] Scikit-learn Developers (2007, Jan.) "1.12. Multiclass and multilabel algorithms". [Online.] Available: http://scikit-learn.org/stable/modules/multiclass.htmlmulticlass

[8] J. Xu, H. Jiang (2015, Sept) "An Improved Information Gain Feature Selection Algorithm for SVM Text Classifier". [Online.] Available: http://ieeexplore.ieee.org/document/7307826/

[9] Scikit-learn Developers (2007, Jan.) "sklearn.feature$_s$election.RFE". [Online.] Available : http : //scikit−learn.org/stable/modules/generated/sklearn.feature$_s$election.RFE.html

[10] Scikit-learn Developers (2007, Jan.) "sklearn.feature$_s$$election.chi2''$.[$Online$.]$Available$: $http : //scikit-learn.org/stable/modules/generated/sklearn.feature_selection.chi2.html$