Current Trends in Biomedical Engineering & Biosciences

Juniper Publishers key to the Researchers

Research Article
Volume 1 Issue 2 - January 2017

Curr Trends Biomedical Eng & Biosci

Controlling Informative Features for Improved Accuracy and Faster Predictions in Omentum Cancer Models

Damian R Mingle*

WPC Healthcare, Nashville, USA

Submission: December 09, 2016; **Published:** January 03, 2017

***Corresponding author:** Damian Mingle, Chief Data Scientist, WPC Healthcare, 1802 Williamson Ct, Brentwood, TN 37027, USA Email: dmingle@wpchealthcare.com

Abstract

Identification of suitable biomarkers for accurate prediction of phenotypic outcomes is a goal for personalized medicine. However, current machine learning approaches are either too complex or perform poorly. Here, a novel feature detection and engineering machine-learning framework is presented to address this need. First, the *Rip Curl* process is applied which generates a set of 10 additional features. Second, we rank all features including the *Rip Curl* features from which the top-ranked will most likely contain the most informative features for prediction of the underlying biological classes. The top-ranked features are used in model building. This process creates for more expressive features which are captured in models with an eye towards the model learning from increasing sample amount and the accuracy/time results. The performance of the proposed *Rip Curl* classification framework was tested on omentum cancer data. *Rip Curl* outperformed other more sophisticated classification methods in terms of prediction accuracy, minimum number of classification markers, and computational time.

Keywords: Omentum, Cancer, Data science, Machine learning, Biomarkers, Phenotype, Personalized medicine

Introduction

In recent years, the dawn of technologies like microarrays, proteomics, and next-generation sequencing has transformed life science. The data from these experimental approaches deliver a comprehensive depiction of the complexity of biological systems at different levels. A challenge within the "-omics" data strata is in finding the small amount of information that is relevant to a particular question, such as biomarkers that can accurately classify phenotypic outcomes [1]. This is certainly true in the fold of peritoneum connecting the stomach with other abdominal organs known as the omentum. Numerous machine learning techniques and methods have been proposed to identify biomarkers that accurately classify these outcomes by learning the elusive pattern latent in the data. To date, there have been three categories that assist in biomarker selection and phenotypic classification:

A. Filters

B. Wrappers

C. Embedding

In practice, time-to-prediction and accuracy of prediction matter a great deal.

Filtering methods are generally considered in an effort to spend the least time-to-prediction and can be used to decide which are the most informative features in relation to the biological target [2]. Filtering produces the degrees of correlation with a given phenotype and then ranks the markers in a given dataset. Many researchers acknowledge the weakness of such methods and take careful note to observe the selection of redundant biomarkers. In addition, filtering methods do not allow for interactions between biomarkers. An example of a popular filtering method is Student's t-test [3].

In order to optimize the predictive power of a classification model, wrapper methods iteratively perform combinatorial biomarker search. Since this combinatorial optimization process is computationally complex, a NP-hard problem, many heuristics have been proposed, for example, to reduce the search space and thus reduce the computational burden of the biomarker selection [4].

With the exception of performing feature selection and classification simultaneously, embedded methods are similar to wrapper methods. Recursive feature elimination support vector machine (SVM-RFE) is a widely used technique for analysis of microarray data [5,6]. The SVM-RFE procedure constructs a

classification model using all available features, with the least informative features being pruned from the model.This process continues iteratively until a model has learned the minimum number of features that are useful. In the case of "-omics" data, this process becomes impractical when considering a large feature space.

Our research used a hybrid approach between user and machine that dramatically reduced the computational time required by similar approaches while increasing prediction accuracy when comparing other state-of-the art machine learning techniques. Our proposed framework includes

A. Ranking and pruning attributes using information gain to extract the most informative features and thus greatly reducing the number of data dimensions;

B. A user to view histograms on attributes where the information gain is 0.80 or higher and creating new binary features from continuous values;

C. Re-ranking both the original features and the newly constructed features; and

D. Using the number of instances to determine how many top-n informative features should be used in modeling. The *Rip Curl* framework can be used to construct a high-dimensional classification model that takes into account dependencies among the attributes for analysis of complex biological -omics datasets containing dependencies of features. The performance of the proposed four-step classification framework was evaluated using datasets from microarray. The proposed framework was compared with SVM-RFE in terms of area under the ROC curve (AUC) and the number of informative biomarkers used for classification.

Results and Discussion

Using the omentum dataset we conducted the *Rip Curl* process of setting the target feature (in this case it was one-versus-all), characterized the target variable, loaded and prepared the omentum data, saved the target and portioning information, analyzed the omentum features, created cross-validation and hold-out partitions, and conducted exploratory data analysis.

Table 1: Different types of descriptive features.

Type	Description
Predictive	A predictive descriptive feature provides information that is useful in estimating the correct value of a target feature.
Interacting	By itself, an interacting descriptive feature is not informative about the value of the target feature. In conjunction with one or more other features, however, it becomes informative.
Redundant	A descriptive feature is redundant if it has a strong correlation with another descriptive feature.
Irrelevant	An irrelevant descriptive feature does not provide information that is useful in estimating the value of the target feature.

In an effort to increase performance and accuracy we opted for an approach of feature selection to help reduce the number of descriptive features in the omentum dataset to just the subset that is most useful for prediction. Before we begin our discussion of approaches to feature selection, it is useful to distinguish between different types of descriptive features (Table 1):

The goal of any feature selection approach is to identify the smallest subset of descriptive features that maintains overall model performance. Ideally a feature selection approach will return the subset of features that includes the predictive and interacting features while excluding the irrelevant and redundant features.

Using conventional methods, it is not efficient to find the ideal subset of descriptive features used to train an omentum model. Considerd features. There are 2^ddifferent possible feature subsets, which is far too many to evaluate unless d is very small. For example, with the descriptive features represented in the omentum dataset,there are $2^{10,960}$ which produces a 3,300 digit integer as the possible feature subsets.

Material and Methods

Datasets

The dataset used in the experiments were provided by the Gene Expression Machine Learning Repository (GEMLeR) [7]. GEMLeR contains microarray data from 9 different tissue types including colon, breast, endometrium, kidney, lung, omentum, ovary, prostate, and uterus. Each microarray sample is classified as tumor or normal. The data from this repository were collated into 9 one-tissue-types versus all-other-types (OVA) datasets where the second class is labeled as "other." All GEMLeR microarray datasets have been analyzed by SVM-RFE, the results of which are available from the same resource.

Methods

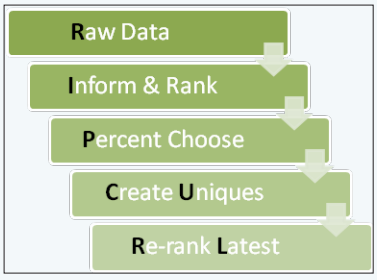

Figure 1: Represent protein and LDH level in different types of meningitis.

Figure 1 demonstrates the *Rip Curl* framework and its dependencies on the prior stage.

In applying the *Rip Curl* framework, we initially ran the omentum microarray data through to gain informative feature feedback and then rank those features from most to least important. We confirmed the number of instances available in the dataset and then applied 1% (1,545 instances X 0.01 = 154

How to cite this article: Damian R M. Controlling Informative Features for Improved Accuracy and Faster Predictions in Omentum Cancer Models. Curr Trends Biomedical Eng & Biosci. 2017; 1(2): 555559.

features) to discover how many top informative features we would make use of in our framework. Where the features were both in the top 1% and expressed informativeness at or above 80%, we created unique features that followed some meaningful thresholds which grouped biomarker data into bins of "0" or "1". Finally, we sent the enhanced data back through the informative feature test to reduce the feature space to the top 1% and then removed all other features, modeling this subset using Random Forest (Entropy).

Our general approach to model selection was to run several model types and to select the best performing based on the highest AUC from the cross-validation results. Once those models were selected, we confirmed that the models were learning models based on sample sizes of the data 16%, 32%, and 64% and are reported in Figure 2. If the model for each sample size did not increase then the model was discarded as a non-learning model (Figure 2).

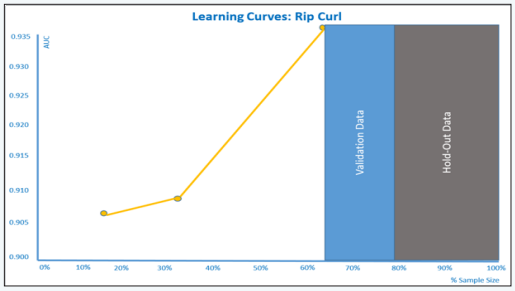

Figure 2: A demonstration of a proper learning model.

We chose area under the ROC curve for its immediate understanding and calculated it as follows

$$ROC\ index = \sum_{i=2}^{|T|} \frac{(FPR(T[i]) - FPR(T[i-1])) \times (TPR(T[i]) + TPR(T[i-1]))}{2} \quad (1)$$

Where T is a set of thresholds, |T| is the number of thresholds tested, and TPR(T[i]) and FPR(T[i]) are the true positive and false positive rates at threshold i respectively. The ROC index is quite robust in the presence of imbalanced data, which makes it a common choice for practitioners, especially when multiple modeling techniques are being compared to one another.

In addition, we ran a second evaluation measure, Gini Norm, which is calculated as follows

Gini coefficient=(2 ×ROC index)-1 (2)

The Gini coefficient can take values in the range of (0,1), and higher values indicate better model performance.

In our experiment with the omentum microarray data, we wanted to pay particular attention to reducing complexity and thereby improving time-to-prediction. This is especially true with an *M X N* dimensional dataset, where *M* is the number of samples and *N* is the features respectively, more specifically where N is orders of magnitude greater than *M*, as is the case in our experiment.

We selected Random Forest to represent the general technique of random decision forests, an ensemble learning method for classification, regression, and other tasks. Random Forest operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of over fitting to their training set. Figure 3 demonstrates visually the increase in predictive accuracy as it relates to complexity of that prediction.

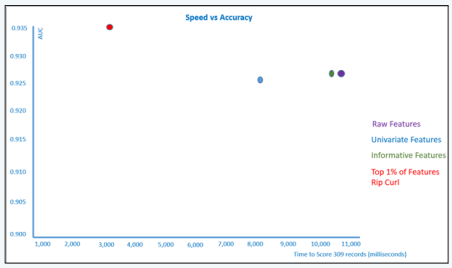

Figure 3: Comparison of Rip Curl vs other methods.

Table 2: Comparative analysis of model performance.

Data	Model	Number of Variables	AUC (Validation)	AUC (Cross -Validation)	AUC (Hold-Out)	Gini Norm (Validation)	Gini Norm (Cross-Validation)	Gini Norm (Hold-Out)	Time (milliseconds)
Raw Features	Random Forest (Entropy)	10,935	0.9520	0.9427	0.9269	0.9040	0.8855	0.8537	10,859.25
Univariate Features	Random Forest (Entropy)	8,165	0.9592	0.9492	0.9232	0.9184	0.8984	0.8463	8,233.39
Informative Features	Random Forest (Entropy)	8,283	0.9520	0.9427	0.9269	0.9040	0.8855	0.8537	10,732.36
Top 1% of Features	Random Forest (Entropy)	15	0.9379	0.9201	0.9344	0.8757	0.8401	0.8689	3,374.21

Table 2 emphasizes the disparity of different results in prediction and time by holding constant the model type, Random Forest (Entropy).

We observed that *Rip Curl* (Top 1% of Features) made use of the best parameters, which we found to be max_depth: None, max_features: 0.2, max_leaf_nodes: 50, min-samples_leaf: 5, and

How to cite this article: Damian R M. Controlling Informative Features for Improved Accuracy and Faster Predictions in Omentum Cancer Models. Curr Trends Biomedical Eng & Biosci. 2017; 1(2): 555559.

min_samples_split: 10. *Rip Curl* improved time-to-prediction by a range of 59.02% to 68.93%, increased the hold-out AUC by 0.81% to 1.21%, and increased the hold-out Gini Norm by a range of 1.78% to 2.67%.

Our framework makes use of Claude Shannon's entropy formula which defines a computational measure of the impurity of the elements in a set. Shannon's idea of entropy is a weighted sum of the logs of the probabilities of each possible outcome when we make a random selection from a set. The weights used in the sum are the probabilities of the outcomes themselves so that outcomes with high probabilities contribute more to the overall entropy of a set than outcomes with low probabilities. Shannon's entropy is defined as

$$H(t) = -\sum_{i=l}^{l}(P(t=i) \times \log_s(P(t=i))) \quad (3)$$

where P(t=i) is the probability that the outcome of randomly selecting an element t is the type i, l is the number of different types of things in the set, and s is an arbitrary logarithmic base (which we selected as 2) (Shannon, 1948).

Once we established the in formativeness of each feature we visually explored the histograms of each variable that expressed an informative value of ≥80%. Below is an example of the histogram for 206067_s_at, indicating where the target's signal was most concentrated within this biomarker (Figure 4).

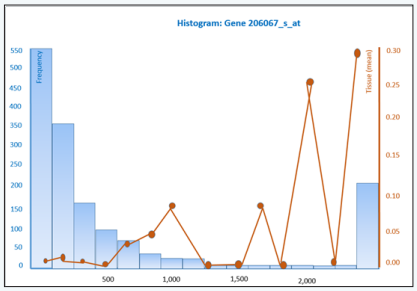

Figure 4: Histogram of Gene 206067_s_at in raw expression.

In an effort to concentrate the omentum tissue signal, we generated a rule that stated

$$f(x) = \begin{cases} 1, & if\ 206067\_s\_at \geq 1800 \\ 0, & otherwise \end{cases} \quad (4)$$

Which rendered a new feature that generated an additional histogram (Figure 5)?

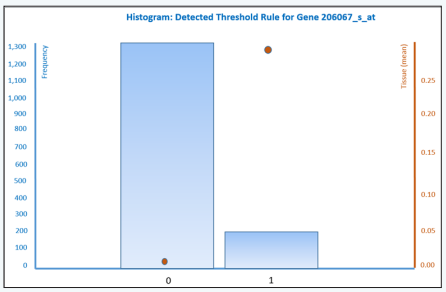

Figure 5: Demonstration of Rip Curl feature engineering using visual thresholds.

Allowing us to pass a different, possibly more understandable context to our algorithm.

We repeated this process above, applying rules based on our observation of the training data.

$$f(x) = \begin{cases} 1, & if\ 216953\_s\_at \geq 400 \\ 0, & otherwise \end{cases} \quad (5)$$

$$f(x) = \begin{cases} 1, & if\ 219454\_at \geq 1100 \\ 0, & otherwise \end{cases} \quad (6)$$

$$f(x) = \begin{cases} 1, & if\ 214844\_s\_at \geq 1300 \\ 0, & otherwise \end{cases} \quad (7)$$

$$f(x) = \begin{cases} 1, & if\ 227195\_s\_at \geq 4000 \\ 0, & otherwise \end{cases} \quad (8)$$

$$f(x) = \begin{cases} 1, & if\ 213518\_at \geq 1100 \\ 0, & otherwise \end{cases} \quad (9)$$

$$f(x) = \begin{cases} 1, & if\ 204457\_s\_at \geq 3500 \\ 0, & otherwise \end{cases} \quad (10)$$

$$f(x) = \begin{cases} 1, & if\ 219778\_at \geq 900 \\ 0, & otherwise \end{cases} \quad (11)$$

Some additional features were simple descriptive statistics such as Min and Mode while others were a bit unconventional such as

$$Lensum = \lfloor \log_{10} \sum_{i=1}^{n} X_i \rfloor + 1 \quad (12)$$

Where X a gene is feature and i represents the placement within the feature index.

Binsum was another engineered feature that was simply

$$Binsum = \sum_{i=1}^{8} bin_i \quad (13)$$

Where bin is one of the generated binary features and i is the index of the bin within the omentum training data. In an effort to develop greater context for the omentum data and the new features that were engineered, we analyzed key values and their respective informativeness (or importance) [8] (Table 3):

Table 3: Rip Curl Feature engineering statistics.

Feature Name	Importance	Unique	Missing	Mean	SD	Median	Min	Max
Binsum	92.88	9	0	1.46	2.07	1	0	8
1800_206067_s_at	84.93	2	0	0.15	0.35	0	0	1
400_216953_s_at	82.15	2	0	0.14	0.34	0	0	1
1100_219454_at	82.09	2	0	0.22	0.42	0	0	1
1300_214844_s_at	78.42	2	0	0.13	0.34	0	0	1

How to cite this article: Damian R M. Controlling Informative Features for Improved Accuracy and Faster Predictions in Omentum Cancer Models. Curr Trends Biomedical Eng & Biosci. 2017; 1(2): 555559.

4000_227195_at	77.04	2	0	0.21	0.41	0	0	1
1100_213518_at	76.07	2	0	0.31	0.46	0	0	1
3500_204457_s_at	75.71	2	0	0.21	0.40	0	0	1
900_219778_at	71.29	2	0	0.09	0.29	0	0	1
Lensum	55.36	3	0	13.82	2.98	16	8	16
Mode	53.04	837	0	214	413	19.80	1.60	4152
Min	51.85	81	0	242	1.41	2.20	0.20	15.40

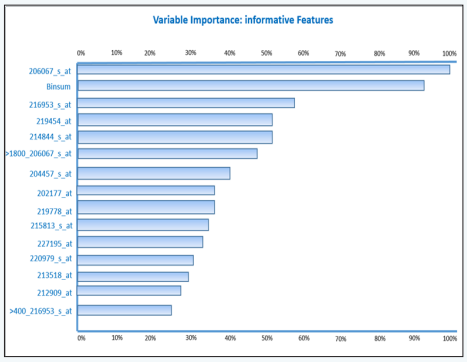

Figure 6: Variable importance rank generated from the Rip Curl framework.

Figure 6 demonstrates visually the variable importance of the final *Rip Curl* model. We observed that 20% of the top informative features were generated through the *Rip Curl* framework: Binsum, >1800_206067_s_at, and >400_216953_s_at with a range of importance between 27% and 93%.

GEMLeR provides an initial classification accuracy value for the omentum dataset. In their experiments designed to generate the state-of-the-art benchmark, all measurements were performed using WEKA machine learning environment. They opted to make use of one of the most popular machine learning methods for gene expression analysis, Support Vector Machines – Recursive Feature Elimination (SVM-RFE) feature selection algorithm. They evaluated (feature selection + classification was done inside a 10-fold cross-validation loop on the omentum dataset to avoid so called selection bias [9] and demonstrates their approach (Figure 7).

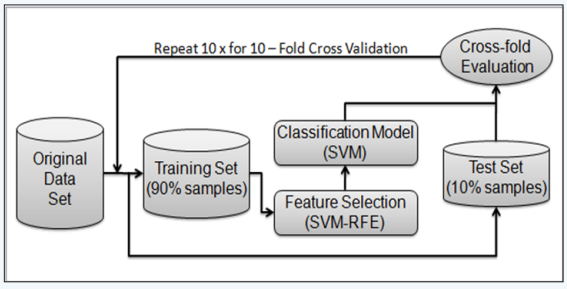

Figure 7: SVM-RFE Process.

Head-to-Head Comparison with SVM-RFE

Table 4: Comparison results of international benchmark and Rip Curl.

Model	AUC
SVM-RFE (Benchmark)	0.703
Rip Curl	0.934

Table 4 shows a comparison of the SVM-RFE benchmark established in with the *Rip Curl* framework, and the following results were observed [10] (Table 4):

Rip Curl represents a 32.92% gain in prediction accuracy over the GEMLeR benchmark for the same omentum dataset [11].

Conclusion

The *Rip Curl* classification framework outperformed the state-of-the-art benchmark (SVM-RFE) in the GEMLeR omentum cancer experiment. Since the *Rip Curl* classification framework utilizes entropy-based feature filtering and adds more contexts through feature engineering, the complexity of this classification framework is very low permitting analysis of data with many features. Future research would suggest comparisons beyond the omentum cancer data and exploration of other one-versus-all experiments in the areas of breast, colon, endomentrium, kidney, lung, ovary, prostate, and uterus [12-14].

Acknowledgment

We would like to acknowledge GEMLeR for making this important dataset available to researchers and WPC Healthcare for supporting our work. Finally, the authors would like to thank the donors who participated in this study.

References

1. Abeel T, HelleputteT, Van de Peer Y, Dupont P, Saeys Y (2010) Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. Bioinformatics 26(3): 392-398.
2. Mingle D (2015) A Discriminative Feature Space for Detecting and Recognizing Pathologies of the Vertebral Column. International Journal of Biomedical Data Mining.
3. Leung Y, Hung Y (2010) A multiple-filter-multiple-wrapper approach to gene selection and microarray data classification. IEEE/ACM Trans Comput Biol Bioinform 7(1): 108-117.

How to cite this article: Damian R M. Controlling Informative Features for Improved Accuracy and Faster Predictions in Omentum Cancer Models. Curr Trends Biomedical Eng & Biosci. 2017; 1(2): 555559.

4. Mohammadi A, Saraee MH, Salehi M (2011) Identification of disease-causing genes using microarray data mining and Gene Ontology. BMC medical genomics 4(1): 1.

5. Ding Y, Wilkins D (2006) Improving the performance of SVM-RFE to select genes in microarray data. BMC bioinformatics 7(Suppl 2): S12.

6. Balakrishnan S, Narayanaswamy R, Savarimuthu N, Samikannu R (2008) SVM ranking with backward search for feature selection in type II diabetes databases. IEEE pp. 2628-2633.

7. Stiglic G, Kokol P (2010) Stability of ranked gene lists in large microarray analysis studies. BioMed Research International 2010: ID 616358.

8. Breiman L (2001) Random forests. Machine learning 45(1): 5-32.

9. Ambroise C, Mc Lachlan GJ (2002) Selection bias in gene extraction on the basis of microarray gene-expression data. Proc Natl Acad Sci U S A 99(10): 6562-6566.

10. Duan K, Rajapakse JC (2004) SVM-RFE peak selection for cancer classification with mass spectrometry data. In APBC pp. 191-200.

11. Hu ZZ, Huang H, Wu CH, Jung M, Dritschilo A, et al. (2011) Omics-based molecular target and biomarker identification. Methods Mol Biol 547-571.

12. Shannon CE (1948) A note on the concept of entropy. Bell System Tech J 27: 379-423.

13. Stiglic G, Rodriguez JJ, Kokol P (2010) Finding optimal classifiers for small feature sets in genomics and proteomics. Neurocomputing 73(13): 2346-2352.

14. Zervakis M, Blazadonakis ME, Tsiliki G, Danilatou V, Tsiknakis M, et al. (2009) Outcome prediction based on microarray analysis: a critical perspective on methods. BMC bioinformatics 10: 53.

How to cite this article: Damian R M. Controlling Informative Features for Improved Accuracy and Faster Predictions in Omentum Cancer Models. Curr Trends Biomedical Eng & Biosci. 2017; 1(2): 555559.