Explore the causes of cancer

Hang Liu
Southampton University

ABSTRACT

Generally, the early detection of cancer relies on the experience of physicians. The machine learning algorithm is considered as a powerful method to classify cancer patients. This report focuses on cancer feature selection and cancer type classification using some interpretive machine learning algorithms. It proved that the genes related to pro-apoptotic, transcriptional activation and life cycle are useful features in this project.

KEYWORDS

cancer prediction, gene, computational biology

ACM Reference Format:
Hang Liu. 2018. Explore the causes of cancer. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

The early detection of cancer is an important part of cancer prevention. The machine learning technology is considered to be a potential way to cancer prediction. In this report, three machine learning algorithms are used to explore the causes of cancer. Firstly, Section 2 gives an overview of the project. It also shows the analysis of the dataset. Section 3 talks about how to illustrate the important features using the machine learning algorithm. Section 4 illustrates the result and analysis of the performance of the algorithm. Section 5 is an evaluation and discussion part, in which the comparison of classifiers, challenges, prior knowledge are all described. Finally, conclusion and future development will be drawn.

Woodstock '18, June 03–05, 2018, Woodstock, NY
2018. ACM ISBN 978-1-4503-9999-9/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

2 ANALYSIS

2.1 Problem Analysis

From the specification of this coursework, the task is to find what features are most important for predicting cancer types and build some interpretive algorithms to classify the cancer type. Namely, The objective of this task is:

(1) Cancer data-set analysis
(2) Machine learning model selection
(3) Performance evaluation

2.2 Data Pre-processing and Analysis

Data sets used in this project are TCGA breast invasive carcinoma (BRCA) dataset and TCGA colon and rectum adenocarcinoma (COADREAD) dataset. The reason why these two datasets are chosen is explained in Section V. They saved cancer transcriptome profiling data, which can be downloaded from the University of California Santa Cruz Xena data centre. In all, these datasets are legal and reliable.

These gene expression profiles were measured experimentally using the Illumina Hiseq 2000 RNA Sequencing platform [13]. They are all mean-normalized across all TCGA cohorts in log2(x+1) transformed RSEM.

In TCGA dataset, the label is expressed as barcode (BCR). Each specifically in BCR identifies a TCGA data element. Referring to given information from Encyclopedia [14], all sample codes of barcode in range from 01 to 09 will be marked as cancer A (breast cancer patient) or B (colon and rectum cancer patient). others will be labeled as non-cancer.

Figure 1 shows the overview of these two datasets.

Figure 2 shows the information about new dataset which is combined dataset.

This new dataset has 20530 features and 3 labels. Figure2 shows that this dataset has 1652 entries and no null value. Shortly, it has low-level background noise.

The next step is analysing the balance of this training dataset.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20530 entries, 0 to 20529
Columns: 1219 entries, sample to TCGA-B6-A0X1-01
dtypes: float64(1218), object(1)
memory usage: 190.9+ MB
None
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20530 entries, 0 to 20529
Columns: 435 entries, sample to TCGA-AG-3592-01
dtypes: float64(434), object(1)
memory usage: 68.1+ MB
None
```

Figure 1: the Overview of this Dataset

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1652 entries, 0 to 2
Columns: 20530 entries, ARHGEF10L to SELS
dtypes: object(20530)
memory usage: 258.8+ MB
```

Figure 2: the Overview of Combined Dataset

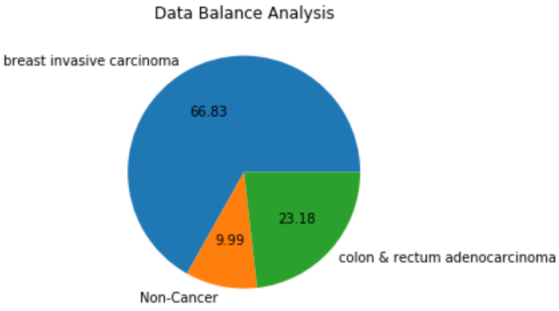

Figure 3: The distribution of labels

Figure 3 shows the distribution of these labels. The imbalance of this dataset affects the way to evaluate the performance of the algorithm.

This dataset is the normalized version of the "gene expression RNAseq" data. Therefore, the object standardized and data cleaning is avoided.

3 METHODOLOGY

Since the reformed data set have three labels, this project is the multinomial classification problem.

The design of the project is described in Figure 4. Shortly, this project is using iterative processing to find the relevant feature.

This project employed three machine-learning techniques, namely linear support vector machines (SVM), decision tree(DT) and random forest(RF), to select a small subset of gene alterations that are most informative for cancer-type prediction.

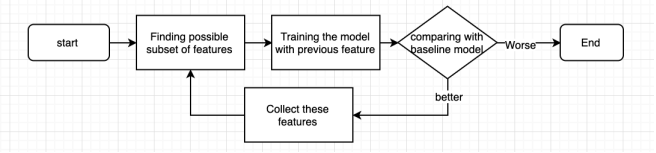

Figure 4: The Flowchart of Programming

3.1 Feature Extraction

For this project, one challenge is high dimensional of the gene expression data. Therefore, it should remove some unrelated features. The Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) are used to feature extraction.

LDA is a classifier with a linear decision boundary, reduce the dimensionality of the input by projecting it to the most discriminative directions [11].

PCA is a Linear dimensionality reduction using Singular Value Decomposition of the data to project it to a lower dimensional space [10].

These two methods return the weight of each predictor in the linear combination that is the discriminant function.

3.2 Models

Since the character of this dataset is a small sample but a large feature, three wrapper models are suitable for this project: SVM, DT and RF. This project will use the k-fold cross-validation to save the samples.

The linear SVM is a possible way to uses a subset of training points in the decision function. SVMs can work on sparse and unbalanced data as it has penalty parameter C of the error term [12].

The DT using CART provided better performance than C4.5, C5.0. Figure 7 shows the visualizing result after training the DT model [2].

RF has become popular in bio-informatics in recent years. Generally, they perform well in a wide variety of situations [4].

In this project, the deep learning methods are abandoned, as this task typically requires interpretive machine learning algorithm.

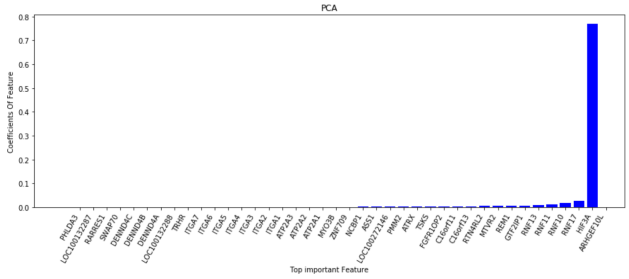

Figure 5: The most important feature of PCA

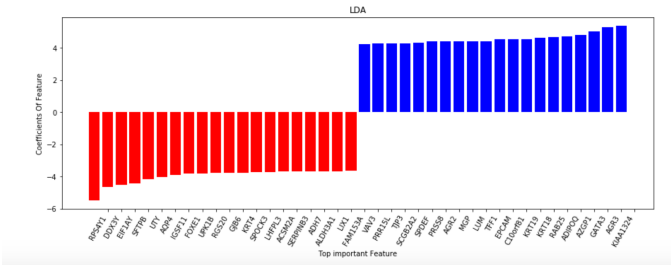

Figure 6: The most important feature of LDA

4 RESULT

4.1 Feature Selection

Figure 5 shows the top 20 features given by PCA. From this Figure, the ARHGEF10L is the most important feature for predicting cancer types.

Figure 6 shows the top 20 features produced by LDA. AKIAAI324 is the most important positive feature. RPS4Y1 is the most important negative feature. Because the difference of these features is small than 1, the LDA shows the canonical correlation of these features is close. Actually, LDA maximizes the separability between these three classes but PCA maximizes explained variance. Therefore, the PCA technique is unsuitable in this project.

Figure 7 shows the top feature in the orange block, which is SPY2. The information gain will be decreased to 0.459 when data-set is split on CDC14B. However, DT is an unstable model. Sometimes, GLP2R is at the top of this tree. Figure 7 just shows one status of results. After counting the frequency of top 5 features, CDC148, GLP2R, FIGF LOC57558 and DST appeared frequently. This feature seems the most important feature for DT.

The visualising top features of SVM is shown in Figure 8, in which some features appear in other gene lists.

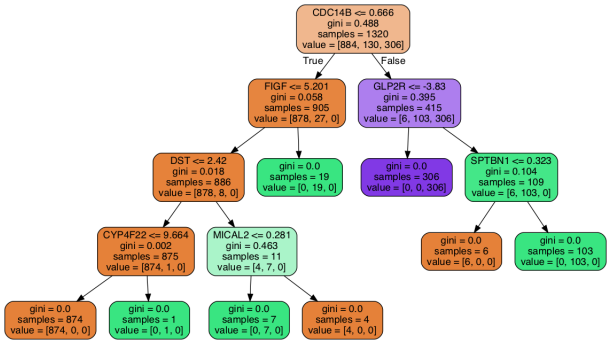

Figure 7: Features at the top of the tree

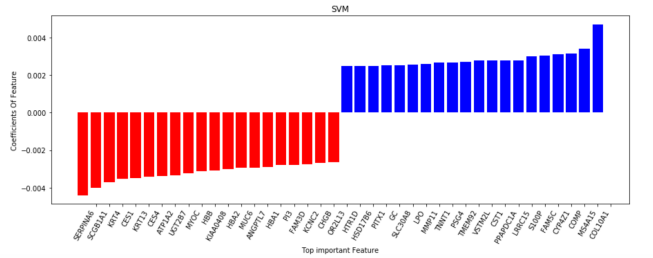

Figure 8: The most important feature of SVM

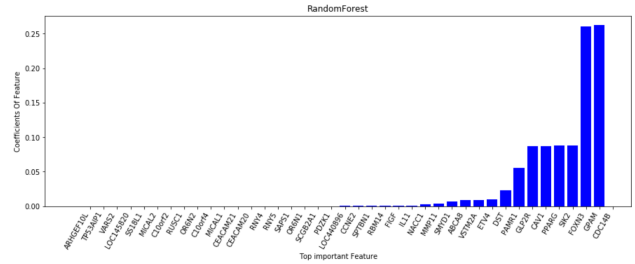

Figure 9: The most important feature of RF

Figure 9 shows the considerably attributes in RF. CAV1, PPAGR, SIK2 and FOXN3 seems to have the same weight in this model.

From Figure 6 to 9, it said these top features have an intersection. It inferred that these features have more influence than other genes for predicting cancer types. There are 142 genes decide the cancer type after counting the number of these features. 18 genes appear twice in this gene list. Table 1 shows the partition of statistical results.

Gene	DT	RF	LDA	SVM	Col_Sum
'ERN2'	1.0	1.0	NaN	NaN	2.0
'ARHGEF10L'	1.0	1.0	NaN	NaN	2.0
'OR6N2'	1.0	1.0	NaN	NaN	2.0
'COL10A1'	NaN	1.0	NaN	1.0	2.0
'SS18L1'	1.0	1.0	NaN	NaN	2.0
'FIGF'	1.0	1.0	NaN	NaN	2.0
'TP53AIP1'	1.0	1.0	NaN	NaN	2.0
'C10orf4'	1.0	1.0	NaN	NaN	2.0
'C10orf2'	1.0	1.0	NaN	NaN	2.0
'KBTBD11'	1.0	1.0	NaN	NaN	2.0
'MICAL2'	1.0	1.0	NaN	NaN	2.0
'RUSC1'	1.0	1.0	NaN	NaN	2.0
'LOC145820'	1.0	1.0	NaN	NaN	2.0
'VARS2'	1.0	1.0	NaN	NaN	2.0

Table 1: Partition of statistical results

	DT	SVM	RF	DT*	SVM*	RF*
F1	0.982	0.995	-	0.988	0.943	-
A	0.984	0.998	-	0.994	0.913	-
P	0.979	0.9907	-	0.982	0.981	-
MSE	0.00903	0.00301	0.0067	0.015	0.102	0.0052

Table 2: Comparison of performance of models using no-processing data and data with selected feature

These features are collected using forward selection or backward elimination. These models will be compared with the baseline model. Section 4.2 describes and evaluate their performances.

4.2 Performance of Classifiers

The score of these classifiers are shown in Table II.

After comparing several scores of models, it found that the model fitted the training set, but did poorly on the test set in the DT model and SVM. However, the score come from RF, trained by selected feature, is better than the score of the based-line model.

5 EVALUATION AND REFLECTION

6 EVALUATION

Generally, the SVM model is suitable for a small amount of data. RF needs more instances to work its randomization concept and generalize to the test data. It thought the result from the SVM model is more reliable.

From the result, 142 genes are enough to achieve an overall accuracy of 98 %. Evenly, the MSE of RF is achieved to 0.0052. Since the total dataset is imbalance and this is a cancer-related task, the F1 score is an optimal metric to evaluate model performance, which considers both the precision and the recall. The F1 score of DT model using reformed dataset is larger than the DT model using total data without feature reduction.

Table 2 shows that the F1 score of these classifier remains robust when the features are decreasing to 142. This result demonstrated that the selected features are strongly related to cancer type.

COL10A1 and KRT4 are the common genes that hold higher weight in the SVM model and LDA. According to [1], KRT4 is specifically found in differentiated layers of the mucosal and oesophagal epithelia together with keratin 13. COL10A1 is associated with poor prognosis in colorectal and cancer breast cancer[8]. However, most of the selected genes do not have high mutation rates. Namely, some selected features may not include cancer driver genes or cell cycle gene. The mutation of these gene affects the life cycle of a cell and tend the cell to accumulate exponentially mutations. This is a considerable factor to feature selection in cancer dataset.

6.1 Challenges

Two challenges in this project are dataset selection and gene complexity.

The first one is how to find valuable dataset in many large and complex datasets given by UCSC Xena. Santa Cruz provided several cancer genomics resources, such as The Cancer Genome Atlas (TCGA), GDC, etc. They recorded 122 cohorts and 1533 datasets. TCGA lists 32 cancer Type for Study. The pan-cancer dataset has 10535 samples and covers 58, 582 transcripts. The reason why choose BRCA and COADREAD dataset are these dataset are all subsets of the pan-cancer dataset. The memory usage of Pan-cancer set is 4.6 GB. The size

of the dataset effect efficiency and accuracy. The total time that is taken to build the training model with the pan-cancer dataset is unacceptable. BRCA and COAD-READ provide 10 level datasets, such as gene expression RNAseq, copy number, etc. This project uses gene expression RNAseq dataset as it enables the investigation and comparison of gene expression levels at unprecedented resolution [5].

The second challenge is caused by the complexity of gene expression. Cancer can be caused by many reasons [6], including exposure to certain behaviours, age, and inherited genetic mutations . These dataset does not include information about these features. Another considerable point is the practical value of this project. One of the motivations of this project is improving the accuracy of the algorithm with selected features. If the selected feature is related to death data, it will reduce the practical value of these features.

6.2 Common sense judgement backed by evidence

From the research [7], the development of cancer is viewed as a multistep process involving mutation and selection for cells with progressively increasing capacity for proliferation, survival, invasion, and metastasis at the cellular level.

It inferred that gene related to cell life and cell death affects the feature type probably.

It is commonly known that cancer caused by abnormal cell proliferation. In cancer, as a result of genetic mutations, this regulatory process malfunctions, resulting in uncontrolled cell proliferation [3]. The normal cell cycles are broken, which means the cell has no response to control cellular growth and death.

According to Cancer Australia [9], TP53 and RB1 are key genes for inhibiting cell proliferation. Cells with mutated TP53 evade the apoptosis mechanisms normally responsible for eliminating impaired cells. TP53AIP1 is selected in the previous algorithm. However, these algorithms ignore RB1.

From the Cell Report [15], it said that DNA CNAs associated with SVs would show the most influence on gene expression. The representative genes are ERBB2 and STARD3. This related information should be considered in programming design.

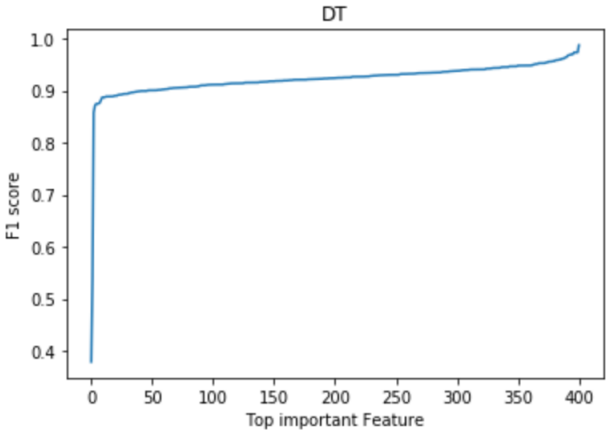

Figure 10: F1 score at the different features

7 CONCLUSION

7.1 Summary

Actually, the classification problem is one of the most fundamental problems in machine learning. From this project, it finds SVM is sufficiently classifiers to feature selection and cancer prediction. Moreover, RF can get acceptable performance when numerous redundant genes are removed. This report proved that using the limited feature to predict cancer types is acceptable when the selected features are related to pro-apoptotic, transcriptional activation and life cycle.

7.2 Future Development

For exploring features which affect the cancer type, one possible way is increasing the number of efficient features and remove the redundant feature. Figure 10 shows the performance of the predictor when the number of genes used is increasing. The performance becomes steady when the number of the gene is more than 200. In this project, the number of selected feature is 142, it can be considered that adding more gene-set from prior knowledge and common sense judgement to train the model.

For improving the performance of the algorithm, the RNN model is a suitable way to be sued to predict cancer type. Figure 11 shows the cost of each epoch in a neural network with 5 layers. It can be found that the cost is decreasing rapidly at the beginning of training processing. After taking 200 iterations, the cost is 0.039

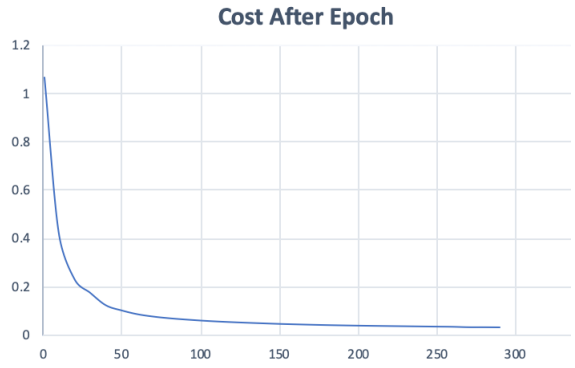

Figure 11: The Cost of Epoch

and achieve a steady state. At the same time, the F1 score is 0.986, which is better than DT. Actually, after 300 iterations, the MSE value of RNN model is small than SVM.

One advantage of RNN is the RNN model can reduce the weight of each weight and make the same feature dead. Shortly, RNN can learn what to store and what to ignore. Since the concept of RNN is based on gene regulatory network, it thought RNN can have a shared representation of features in this project.

REFERENCES

[1] CA Galipeau PC et al. Chao DL, Sanchez. 2008. Cell proliferation, cell cycle abnormalities, and cancer outcome in patients with BarrettâĂŹs esophagus: A long-term prospective study. *Clinical cancer research : an official journal of the American Association for Cancer Research* 14(21) (2008). https://doi.org/10.1158/1078-0432.CCR-07-5063

[2] Chawla Nitesh V. Cieslak, David A. 2008. Learning Decision Trees for Unbalanced Data. In *Machine Learning and Knowledge Discovery in Databases*, Goethals Bart Morik Katharina Daelemans, Walter (Ed.). Springer Berlin Heidelberg, Berlin, Heidelberg, 241–256.

[3] Jill U. Adams Clare O'Connor. 2004. The Development and Causes of Cancer.. In *The Cell: A Molecular Approach*. Nature Education. https://www.nature.com/scitable/ebooks/essentials-of-cell-biology-14749010/122997842

[4] Padideh Danaee, Reza Ghaeini, and David Hendrix. 2017. A Deep Learning Approach for Cancer Detection and Relevant Gene Identification. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* 22 (2017), 219–229.

[5] Di Camillo Barbara Finotello, Francesca. 2014. Measuring differential gene expression with RNA-seq: challenges and strategies for data analysis. *Briefings in Functional Genomics* 14, 2 (09 2014), 130–142. https://doi.org/10.1093/bfgp/elu035 arXiv:http://oup.prod.sis.lan/bfg/article-pdf/14/2/130/715070/elu035.pdf

[6] Beare David Gunasekaran Prasad Leung Kenric Bindal Nidhi Boutselakis Harry Ding Minjie Bamford Sally Cole Charlotte Ward Sari Kok Chai Yin Jia Mingming De Tisham Teague Jon W. Stratton Michael R. McDermott Ultan Campbell Peter J. Forbes, Simon A. 2014. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Research* 43, D1 (10 2014), D805–D811.

[7] Cooper GM. 2000. The Development and Causes of Cancer.. In *The Cell: A Molecular Approach*. Sunderland (MA): Sinauer Associates. https://www.ncbi.nlm.nih.gov/books/NBK9963/

[8] Li T Ye G Zhao L Zhang Z Mo D Wang Y Zhang C Deng H Li G Liu H Huang H. 2018. High expression of COL10A1 is associated with poor prognosis in colorectal cancer. *PubMed Central* (2018). https://doi.org/110.2147/OTT.S160196

[9] Dueensland University of Technology. 2001. Abnormal cell proliferation. Retrieved May 7, 2019 from http://edcan.org.au/edcan-learning-resources/supporting-resources/biology-of-cancer/abnormal-cell-proliferation

[10] scikit-learn developers 2007. *sklearn.decomposition.PCA*. scikit-learn developers. https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html.

[11] scikit-learn developers 2007. *sklearn.lda.LDA*. scikit-learn developers. https://scikit-learn.org/0.16/modules/generated/sklearn.lda.LDA.html.

[12] scikit-learn developers 2007. *sklearn.svm.LinearSVC*. scikit-learn developers. https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html#sklearn.svm.LinearSVC.

[13] The Cancer Genome Atlas 2006. *The Cancer Genome Atlas Program*. The Cancer Genome Atlas. https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga/?redirect=true.

[14] The Cancer Genome Atlas 2006. *TCGA barcode*. The Cancer Genome Atlas. https://docs.gdc.cancer.gov/Encyclopedia/pages/images/TCGA-TCGAbarcode-080518-1750-4378.pdf.

[15] Yang Lixing Kucherlapati Melanie Chen Fengju Hadjipanayis Angela Pantazi Angeliki Christopher A. Bristow Lee Eunjung Alice ... Chad J. Creighton Zhang, Yiqun. 2018. A Pan-Cancer Compendium of Genes Deregulated by Somatic Genomic Rearrangement across More Than 1,400 Cases. *Cell Reports* 10, 24(2) (2018), 515–527.