Application of Bayesian Logistic Regression to Mining Biomedical Data

Viji R. Avali PhD[1], Gregory F. Cooper MD, PhD[1,2,3], and Vanathi Gopalakrishnan PhD[1,2,3]
[1]Department of Biomedical Informatics, [2]Intelligent Systems Program, [3]Department of Computational and Systems Biology, University of Pittsburgh, Pittsburgh, PA

Abstract

Mining high dimensional biomedical data with existing classifiers is challenging and the predictions are often inaccurate. We investigated the use of Bayesian Logistic Regression (B-LR) for mining such data to predict and classify various disease conditions. The analysis was done on twelve biomedical datasets with binary class variables and the performance of B-LR was compared to those from other popular classifiers on these datasets with 10-fold cross validation using the WEKA data mining toolkit. The statistical significance of the results was analyzed by paired two tailed t-tests and non-parametric Wilcoxon signed-rank tests. We observed overall that B-LR with non-informative Gaussian priors performed on par with other classifiers in terms of accuracy, balanced accuracy and AUC. These results suggest that it is worthwhile to explore the application of B-LR to predictive modeling tasks in bioinformatics using informative biological prior probabilities. With informative prior probabilities, we conjecture that the performance of B-LR will improve.

Introduction

Biomedical data tend to have many variables and a scarcity of samples. Mining such high dimensional data with existing classifiers is challenging and the predictions are often inaccurate. Logistic regression (LR) is often applied in making predictions. However, it is difficult to include prior biological knowledge into the analysis when using LR. Almost all biomedical domains have associated domain knowledge. It would be helpful to be able to include such additional knowledge when building predictive models. For example, if a predictor variable is already known as a biomarker for a disease, it will be prudent to use this information when trying to come up with a model for classification and prediction for that disease. Prior knowledge can be incorporated into Bayesian Logistic Regression (B-LR) and the method is computationally efficient. B-LR has been applied successfully in text categorization [1], in integrating early physiological responses to predict later illness severity in preterm infants[2], and in early prediction of the response of breast tumors to neoadjuvant chemotherapy [3]. But in both [2] and [3], the number of predictor variables is small. We want to study the performance of B-LR on classifying high dimensional data and compare its performance to other existing classifiers. This paper uses a B-LR implementation that is readily available in the WEKA data mining environment[4]. Our goal is to understand the extent to which B-LR performs on par with other classifiers in WEKA according to the following performance measures – accuracy, balanced accuracy (i.e., average of sensitivity and specificity), and the area under the ROC curve (AUC).

Background

Linear logistic regression is a probabilistic classification model used for predicting a target variable depending on one or more predictor variables. It can give accurate predictions, but it often does not handle high dimensional data well. One way to overcome the shortcomings of LR is to apply a Bayesian approach with a prior probability distribution over predictor variables. In Bayesian analyses, the three steps involved are (1) specifying prior probabilities for the parameters, (2) determining the marginal likelihood of the data, (3) and using Bayes theorem to determine the posterior distribution of the parameters. B-LR captures the nonlinear relationships between the predictor variables and the outcome variable using Bayesian modeling. In B-LR, the equation for calculating the posterior probability of a sample belonging to a specific class is generated by the traditional logistic function:

$$P(Class|a_1, a_2, a_3, \dots, a_n) = \frac{1}{(1 + \exp\,(b + w_0 * c + \sum_{i=1}^{n} w_i * f(a_i)))} \quad (1)$$

Where, $'a_i'$ denotes the predictor variables, 'c' is the prior log odds ratio $\left(c = log\ \frac{P(\text{class}=0)}{P(\text{class}=1)}\right)$, the bias '$b$' and weights w_0 and w_i are learned from the training data, and the i^{th} attribute a_i is used to calculate the feature $f(a_i)$, using $f(a_i) = \log\left(\frac{P(a_i \mid \text{class}=0)}{P(a_i \mid \text{class}=1)}\right)$ (for binary class outcome variables). In the Bayesian approach to logistic regression, a univariate Gaussian prior with a mean '0' and a variance of 'σ_i' over the weights is commonly used. By

using a mean of '0', we assert our prior belief that the weights are close to zero. The values of σ_i are positive, with small values indicating our confidence in the values of the weights and larger values indicating the lack there of. Though this Gaussian prior favors weights with values close to zero, it does not favor the values exactly being zero. Maximum a posteriori (MAP) estimate of these weight values is similar to ridge regression for the logistic model.

The B-LR implementation in WEKA is based on [1] and has Gaussian parameter priors and Laplace parameter priors as the two options. Domain knowledge related to the datasets can be incorporated by specifying a prior, thereby defining a distribution over the values of the weights. Since the WEKA implementation of B-LR has Gaussian and Laplace priors as the two options available, we used only these non-informative priors in our analysis.

Experimental Method

Twelve datasets with binary class variables were chosen. Eleven are publicly available and one is a private dataset collected in the LungSPORE project[5]. The LungSPORE dataset contains as yet unpublished data that was collected to validate the results of an earlier study [5]. This study identified a panel of ten serum biomarkers that distinguished lung cancer from controls and have the potential to aid in the early detection of lung cancer and more accurate interpretation of indeterminate pulmonary nodules detected by CT screening.

Table 1. Details of the 12 datasets that were analyzed.
G/P indicates if the data is Genomic or Proteomic. P/D shows whether the data is Prognostic (P) or Diagnostic (D). The number of variables (Original) gives the total variables in the original dataset. The number of variables (PAIFE) gives the total number of variables after processing the dataset through our irrelevant feature elimination algorithm 'PAIFE'. The Sample (Class1, Class2) gives the total number of samples and class distribution, and 'Reference', the relevant reference to the dataset.

ID	G/P	P/D	Number of variables (Original)	Number of variables (PAIFE)	Sample(Class1,Class2)	Outcome variable	Reference
1	G	D	6584	1972	61(40,21)	Colon Cancer	Alon et al [6]
2	G	P	5372	858	86(69,17)	Lung Cancer	Beer et al. [7]
3	P	D	70	15	205(66,139)	Lung Cancer	Bigbee, et al. [5]
4	G	D	7129	2288	72(47,25)	Leukemia	Golub, et al. [8]
5	G	D	7464	1880	36(18,18)	Breast Cancer	Hedenfalk et al. [9]
6	G	P	7129	699	60(20,40)	Hepatocellular carcinoma	Iizuka et al. [10]
7	G	P	7399	1084	240(138,102)	Lymphoma	Rosenwald, et al. [11]
8	G	D	7129	1927	77(58,19)	Lymphoma	Shipp, et al. [12]
9	G	P	24481	4251	78(44,34)	Breast cancer	Van't Veer, et al. [13]
10	G	D	7039	1230	39(35,4)	Ovarian Cancer	Welch, et al. [14]
11	G	P	12625	1166	249(201,48)	Leukemia	Yeoh, et al. [15]
12	P	D	16	12	583(184,401)	Lung cancer	LungSPORE (unpublished)

For high dimensional biomedical data, the presence of uninformative variables in the dataset introduces noise and adversely affects the performance of classifiers. As a preprocessing step, we used our in-house developed algorithm called 'Partitioning based adaptive irrelevant feature eliminator' (PAIFE) to remove presumptive non-informative features from the datasets[16]. PAIFE evaluates predictor variable – outcome variable relationships over not only a whole dataset, but also the partitioned subsets and is effective in identifying variables whose relevance to the outcome are conditional on certain other variables. In experiments with synthetic datasets, PAIFE had outperformed other state-of-the-art feature selection methods in retaining relevant features and eliminating irrelevant ones [16]. PAIFE successfully removed irrelevant features when tested on proteomic and genomic datasets and the models developed from the PAIFE processed datasets performed either better or on par with the models built without any processing.

The PAIFE processed datasets were then normalized using the unsupervised attribute filter 'normalize' in WEKA's preprocessing step. The following methods were applied with 10-fold cross validation on each of these PAIFE processed, normalized datasets: B-LR with both Gaussian and Laplace priors, B-LR with Gaussian priors and cross-validation-based hyperparameter selection [1], C4.5 (J48 in WEKA) [17], naïve Bayes [18], simple logistic regression [19, 20], ridge logistic regression [21], CART (SimpleCart in WEKA) [22], Random Forest [23], and SVM (SMO in WEKA[24]), as implemented in WEKA 3.6.10 [4]. Simple logistic regression (LR_{simple}) in WEKA is the linear logistic regression and ridge logistic regression (LR_{ridge}) is the logistic regression model with a ridge estimator. For all the classifiers, except B-LR with Gaussian priors and hyperparameter selection based on cross validation (CV), default parameter values in WEKA were used. That classifier was chosen by selecting the 'CV based hyperparameter' option in WEKA's B-LR classifier.

The statistical significance of the results was analyzed by paired two-tailed t-test and by non-parametric Wilcoxon paired-samples signed ranks test. We used the alpha value of 0.05 for significance testing.

Results

In our performance analysis, we compared the accuracy, balanced accuracy (BACC) and percentage AUC values of each classifier.

Table 2: Comparison of the accuracy (percentage) of the classifiers from 10-fold cross validation.

ID	B-LR_{GP1}	B-LR_{GP2}	B-LR_{LP}	J48	NB	LR_{simple}	LR_{ridge}	CART	SVM	RF
1	96.72	96.72	96.72	98.36	96.72	98.36	96.72	**100.0**	98.36	88.52
2	93.02	93.02	88.37	66.28	84.88	81.40	86.05	76.74	**93.02**	79.07
3	78.05	74.63	73.17	80.98	78.54	82.93	**83.90**	79.02	80.49	80.49
4	98.61	97.22	98.61	84.72	98.61	90.28	94.44	83.33	**98.61**	94.44
5	97.22	97.22	94.44	91.67	**100.0**	94.44	97.22	94.44	97.22	91.67
6	**91.67**	**91.67**	75.00	55.00	88.33	71.67	78.33	65.00	**91.67**	78.33
7	70.00	**71.67**	70.42	60.00	65.83	67.50	65.42	57.08	69.17	65.42
8	**97.40**	**97.40**	**97.40**	70.13	89.61	94.81	**97.40**	68.83	**97.40**	85.71
9	96.15	85.90	89.74	82.05	83.33	93.59	93.59	80.77	**98.72**	92.31
10	**100.0**	**100.0**	**100.0**	92.31	97.44	97.44	**100.0**	87.18	**100.0**	97.44
11	89.16	85.14	82.33	72.69	79.92	80.72	86.75	80.72	**89.96**	81.12
12	91.21	92.43	80.37	93.46	90.59	95.50	95.50	94.89	92.43	**96.11**
Avg	91.60	90.25	87.21	78.97	87.82	87.39	89.61	80.67	**92.25**	85.89
s.d.	8.61	8.82	10.19	13.49	9.61	9.92	9.65	12.16	8.70	8.98

Table 3: Comparison of BACC (percentage) values of the classifiers from10-fold cross validation.

ID	B-LR_{GP1}	B-LR_{GP2}	B-LR_{LP}	J48	NB	LR_{simple}	LR_{ridge}	CART	SVM	RF
1	96.37	96.37	96.37	97.73	96.37	98.78	95.65	**100.0**	97.73	90.21
2	92.73	92.73	**93.67**	48.04	76.42	70.34	85.90	39.76	92.73	40.00
3	82.65	83.38	**85.82**	78.32	76.89	82.08	83.38	76.45	82.70	77.79
4	**98.96**	96.94	**98.96**	82.95	**98.96**	89.05	93.33	81.48	**98.96**	96.08
5	97.37	97.37	95.00	92.86	**100.00**	95.00	97.37	94.44	97.37	92.86
6	**92.41**	**92.41**	74.12	52.29	90.00	67.78	75.64	55.39	91.08	75.67
7	69.28	**71.03**	69.92	59.25	65.50	66.80	64.71	54.68	68.46	64.60
8	96.51	96.51	96.51	62.34	83.78	93.01	**98.33**	58.08	96.51	84.74
9	96.26	87.47	89.88	81.75	84.64	94.13	93.38	81.18	**98.89**	93.13
10	**100.0**	**100.0**	**100.0**	78.53	98.61	90.00	**100.0**	44.74	**100.0**	98.61
11	87.88	**92.23**	78.79	58.79	68.20	67.15	82.02	40.36	86.99	73.98
12	92.93	93.25	87.56	92.94	90.37	95.61	95.50	94.78	94.17	**95.65**
Avg	91.95	91.64	88.88	73.82	85.81	84.15	88.77	68.45	**92.13**	81.94
s.d,	8.27	7.59	9.54	16.30	11.49	12.08	10.25	21.22	8.73	16.24

Table 4: AUC (percentage) values for the classifiers from 10-fold cross validation.

ID	B-LR_{GP1}	B-LR_{GP2}	B-LR_{LP}	J48	NB	LR_{simple}	LR_{ridge}	CART	SVM	RF
1	96.40	96.40	96.40	98.70	96.30	98.40	99.50	**100.00**	98.70	97.90
2	84.60	84.60	70.60	49.40	85.80	80.60	**88.50**	39.70	84.60	79.50
3	67.10	61.00	58.30	79.00	85.80	85.90	86.40	78.80	71.70	**88.10**
4	98.00	96.90	98.00	81.40	98.70	95.40	**99.30**	79.00	98.00	96.70
5	97.20	97.20	94.40	91.70	**100.00**	93.20	98.60	94.40	97.20	91.70
6	88.80	88.80	66.30	52.20	86.70	70.50	84.70	53.60	**90.00**	87.80
7	68.80	70.50	68.60	62.40	72.30	**74.70**	71.70	54.40	68.50	69.20
8	96.50	96.50	96.50	62.20	88.60	**99.50**	97.50	61.40	96.50	93.30
9	95.90	84.50	89.20	84.40	86.40	97.20	**99.50**	75.50	98.50	95.40
10	**100.00**	**100.00**	**100.00**	84.60	87.50	99.30	**100.00**	44.60	**100.00**	**100.00**
11	75.00	61.50	55.80	61.50	73.10	69.20	**81.60**	48.10	78.70	74.40
12	88.70	90.70	73.90	92.20	96.70	**98.90**	98.80	93.70	90.10	98.60
Avg	88.08	85.72	80.67	74.98	88.16	88.57	**92.18**	68.60	89.38	89.38
s.d.	11.26	13.41	15.95	15.95	8.59	11.35	8.98	20.11	10.63	9.64

In tables 2, 3, 4, 5, and 6, we use B-LR_{GP1} to indicate B-LR with Gaussian priors, B-LR_{GP2} to indicate B-LR with Gaussian priors with cross-validation based hyperparameter selection, B-LR_{LP} for B-LR with Laplace priors, NB for Naïve Bayes, RF for Random Forest, 'Avg' for 'Average', 's.d.' for 'standard deviation'.

Table 2 shows the accuracy of all the classifiers on the different datasets. The bold value on each row indicates the classifier with the highest accuracy for that dataset.

Table 3 gives the balanced accuracy for the different classifiers with the bold numbers indicating the classifier with the maximum BACC value for a specific dataset. Table 4 shows the percentage of AUC values for each of the classifiers for all the datasets.

We evaluated the statistical significance of these performance measures using the paired two-tailed t-test and the non-parametric Wilcoxon signed ranks test. Tables 5 and 6 show the results. The captions of the tables explain the contents of their cells.

Table 5: Comparison of the classifiers using a paired two-tailed t-test.
The numbers shown are p-values; the values below 0.05 are shown in bold. The value in a parenthesis is the mean performance of B-LR minus the mean performance of the listed classifier, expressed as a percentage. The values underlined are those in which the p-value is less than 0.05 and B-LR performed better.

B-LR with Gaussian priors versus each of the following:	Accuracy	BACC	AUC
B-LR_{GP2}	0.18 (1.35)	0.73(0.31)	0.14(2.37)
B-LR_{LP}	**0.01**(4.39)	0.10(3.06)	**0.01**(7.42)
J48	**0.01**(12.63)	**0.00**(18.13)	**0.02**(13.11)
NB	**0.02**(3.78)	**0.01**(6.13)	0.98(-0.08)
LR_{simple}	0.06(4.22)	**0.02**(7.8)	0.86(-0.48)
LR_{ridge}	0.20(1.99)	0.06(3.18)	**0.04**(-4.09)
CART	**0.00**(10.93)	**0.00**(23.5)	**0.01**(19.48)
SVM	0.06(-0.65)	0.57(-0.19)	**0.02**(-1.29)
RF	**0.01**(5.71)	**0.04**(10.00)	0.55(-1.3)

Table 6: Comparison of the classifiers using the Wilcoxon paired-samples signed ranks test. The numbers shown are p-values; the values below 0.05 are shown in bold. The value in a parenthesis is the mean performance of B-LR_{GP1} minus the mean performance of the listed classifier, expressed as a percentage. The values underlined are those in which the p-value is less than 0.05 and B-LR performed better.

B-LR_{GP1} versus each of the following:	Accuracy	BACC	AUC
B-LR_{GP2}	0.22 (1.35)	1(0.31)	0.31(2.37)
B-LR_{LP}	**0.02**(4.39)	0.15(3.06)	**0.01**(7.42)
J48	**0.01**(12.63)	**0.00(18.13)**	**0.02**(13.11)
NB	**0.02**(3.78)	**0.01(6.13)**	1.00(-0.08)
LR_{simple}	0.07(4.22)	**0.02(7.8)**	0.89(-0.48)
LR_{ridge}	0.31(1.99)	0.06(3.18)	0.02(-4.09)
CART	**0.01**(10.93)	**0.00**(23.5)	**0.02**(19.48)
SVM	0.09(-0.65)	0.58(-0.19)	**0.03(-1.29)**
RF	**0.01**(5.71)	**0.00**(10.00)	0.70(-1.3)

Discussion

From table 2 of the results, we can see that B-LR with Gaussian prior had the highest accuracy for three of the datasets and an average accuracy of 91.60%. Though SVM has outperformed B-LR with Gaussian prior with an average accuracy of 92.25%, the difference between the two values is very small (0.65%) and the standard deviation is 8.61 and 8.70, for B-LR and SVM respectively. B-LR with Gaussian prior has the maximum BACC value for three of the datasets and an average of 91.95%. SVM has the highest average BACC value of 92.13% leading B-LR with Gaussian prior by 0.18% (Table 3).

It is interesting to observe from table 4 that LR (with ridge estimator) has the highest AUC value of 92.18% when its accuracy and BACC measure were about 2% and 3% behind those of B-LR with Gaussian prior. In comparing the performance using paired two-tailed t-test and the non-parametric Wilcoxon signed ranks test, on accuracy, B-LR with Gaussian priors (B-LR_{GP1}) performed statistically significantly better ($p \leq 0.02$) than B-LR with Laplace priors, J48, Naïve Bayes, CART, and Random Forests. Only SVM had a higher accuracy, which was higher by 0.65% ($p = 0.06$). On BACC, B-LR_{GP1} performed statistically significantly ($p \leq 0.04$) better than J48, Naïve Bayes, LR_{simple}, CART, and Random Forest. No method had a statistically significantly better performance than B-LR_{GP1}, according to the BACC measure. On AUC, B-LR_{GP1} performed statistically significantly better ($p \leq 0.04$) than B-LR with Laplace prior, J48, and CART. LR_{ridge} and SVM had higher AUCs than B-LR_{GP1}, with LR_{ridge} being higher by 4.09% ($p = 0.04$) and SVM being higher by 1.29% ($p = 0.02$).

B-LR_{GP1} was also among the fastest methods, with an average time of 0.13 sec to build the model compared to LR_{ridge}'s average time of 13.45 sec.

In this study, we limited ourselves by using a single type of feature selection method, PAIFE. It would be important to learn the impact of the choice of feature selection method. In the future, we will examine other state-of-the-art feature selection methods and compare its performance to PAIFE. We would also like to observe the results without feature selection. Genkin et al. [1], observed that lasso logistic regression was effective on high dimensional data

analysis problems, it would therefore be interesting to observe the performance of lasso logistic regression on such higher-dimensional and noisier data.

Conclusion

The results from this study provide support that B-LR with a Gaussian prior performs well compared to a set of classifiers that include those often applied in bioinformatics. It provides researchers with an additional classifier from which they can choose when analyzing high dimensional data.

With these promising preliminary results, the next step will be to use biological domain knowledge to develop informative priors to use in B-LR, and then repeat the evaluation of its predictive performance. In this study, we analyzed only 12 datasets with binary class variables. We plan to extend our analysis to more datasets and datasets with multinomial class variables. Our future analysis will also evaluate the performance of B-LR when changing the parameter options in WEKA's B-LR classifier. For example, we used the default value of 100 for the number of iterations. We plan to have our own implementation of B-LR with options to choose the type of informative priors depending on the application domain.

Acknowledgements

The authors gratefully acknowledge the following grants from the National Library of Medicine at the National Institutes of Health: R01-LM010950 and 5T15 LM007059-26. VG was funded in part by grants R01GM100387 and P50CA090440 from the National Institutes of Health. GFC was funded in part by NIH grant R01LM010020 and NSF grant IIS0911032. The content of this paper is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the National Science Foundation.

This project used the UPCI Cancer Biomarkers Facility that is supported in part by award P30CA047904. We thank Dr. William L. Bigbee and his laboratory for the recently produced dataset from the Lung Cancer SPORE project (supported by NCI grant number: P50CA090440) that was also analyzed in this paper. The authors thank Jeya B. Balasubramanian, MS for processing the datasets through PAIFE and for performing the Wilcoxon test analysis.

References

1. Genkin, A., Lewis, D., and Madigan, D., *Large-Scale Bayesian Logistic Regression for Text Categorization.* Technometrics, 2007: p. 291-304.
2. Saria S, R., AK, Gould J, Koller D, Penn AA., *Integration of early physiological responses predicts later illness severity in preterm infants.* Science Translational Medicine, 2010. **2**(48): p. 48-65.
3. Mani S, C.Y., Arlinghaus LR, Li X, Chakravarthy AB, Bhave SR, Welch EB, Levy MA, Yankeelov TE. *Early Prediction of the Response of Breast Tumors to Neoadjuvant Chemotherapy using Quantitative MRI and Machine Learning*. in *AMIA Anuual symposiam*. 2011.
4. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H., *The WEKA Data Mining Software: An Update.* SIGKDD Explorations, 2009. **11**(1).
5. Bigbee, W.L., Gopalakrishnan, V., Weissfeld, J. L., Wilson, D. O., Dacic, S., Lokshin, A. E., & Siegfried, and J. M., *A multiplexed serum biomarker immunoassay panel discriminates clinical lung cancer patients from high-risk individuals found to be cancer-free by CT screening.* Journal of Thoracic Oncology, 2012(Apr;7(4)): p. 698-708.
6. Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., & Levine, A. J., *Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays.* Proceedings of the National Academy of Sciences of the United States of America, 1999(96(12)): p. 6745-6750.
7. Beer, D.G., Kardia, S. L. R., Huang, C.-C., Giordano, T. J., Levin, A. M., Misek, D. E., … Hanash, S, *Gene-expression profiles predict survival of patients with lung adenocarcinoma.* Nature Medicine, 2002(8): p. 816–824.

8. Golub, T.R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., … Lander, E. S., *Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.* Science,, 1999(286): p. 531–537.
9. Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., … Sauter, G. , *Geneexpression profiles in hereditary breast cancer.* The New England Journal of Medicine, 2001(344): p. 1-6.
10. Iizuka, N., Oka, M., Yamada-Okabe, H., Nishida, M., Maeda, Y., Mori, N., … Hamamoto, Y., *Oligonucleotide microarray for prediction of early intrahepatic recurrence of hepatocellular carcinoma after curative resection.* Lancet, 2003(361): p. 923–929.
11. Rosenwald, A., et al, *The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma.* New England Journal of Medicine, 2002(346(25)): p. 1937-47.
12. Shipp, M.A., Ross, K. N., Tamayo, P., Weng, A. P., Kutok, J. L., Aguiar, R. C. T., … Golub, T. R, *Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning.* Nature Medicine, 2002(8): p. 68-74.
13. Veer, L.v.t., Dai, H., & Vijver, M. Van De., *Gene expression profiling predicts clinical outcome of breast cancer.* Nature, 2002.
14. Welsh, J.B., Zarrinkar, P. P., Sapinoso, L. M., Kern, S. G., Behling, C. A., Monk, B. J., ... & Hampton, G. and J. M., *Analysis of gene expression profiles in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer.* Proceedings of the National Academy of Sciences of the United States of America, 2001(98(3)): p. 1176-1181.
15. Yeoh, E., et al., *Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling.* Cancer Cell, 2002(1(2)): p. 133-43.
16. Liu, G., Kong, L., & Gopalakrishnan, V. *A Partitioning Based Adaptive Method for Robust Removal of Irrelevant Features from High-dimensional Biomedical Datasets*. in *AMIA Joint Summits on Translational Science*. 2012.
17. Quinlan, R., *C4.5: Programs for Machine Learning*. 1993: Morgan Kaufmann Publishers Inc.
18. John, G.H., Langley P. *Estimating Continuous Distributions in Bayesian Classifiers. In: Eleventh Conference on Uncertainty in Artificial Intelligence*. in *Conference on Uncertainty in Artificial Intelligence*. 1995. Morgan Kaufmann.
19. Landwehr, N., Hall, M., and Frank, E., *Logistic Model Trees.*, in *Machine Learning*. 2005, Springer-Verlag. p. 241-252.
20. Sumner, M., Frank, E. , and Hall, M.A. *Speeding up logistic model tree induction*. in *9th European Conference on Principles and Practice of Knowledge Discovery in Databases*. 2005. Springer.
21. le Cessie, S., van Houwelingen, J.C. , *Ridge Estimators in Logistic Regression.* Applied Statistics, 1992: p. 191-201.
22. Breiman, L., et al., *Classification and regression trees*. 1984: CRC press.
23. Breiman, L., *Random forests.* Machine learning, 2001. **45**(1): p. 5-32.
24. Platt, J., *Fast Training of Support Vector Machines using Sequential Minimal Optimization.* Advances in Kernel Methods - Support Vector Learning, 1998.