Feature Selection and Classification using Modified Bayesian Information Creiterion

Name: Jayanth Sivakumar

0.1 Overview

Classification for high-dimensional dataset is a challange every data scientist face. Handling the big data is time consuming and the memory leaks are costly if the right algorithm and garbage collection is not used. As going through the competition,trying different algorithms initially had a lot of issues unlike the small datasets.The stack overflow error happens all the time when trying different algorithms. Building a model by keeping the dataset in the disk is one of the new approaches and is very fast preventing unnecessary memory leaks. An approach that overcame these challenges is Parallel Computing. It is very fast, with efficient garbage collection methods and is robust. After exploring multiple parallel computing algorithms with the support of machine learning algorithms, the competition became interesting. With the use of this approach, trying to fit a model for this dataset was not a hard task.But choosing the right algorithm which is stable and reproducible came later in the competition after variable selection. Dimensionality Reduction was hard to do for this dataset. In the end, the regularized logistic regression with feature selection was the successful and fast method that contributed to the highest model accuracy.

0.2 Base Learners

After trying different base learners like Lasso, Elastic Net, SVM and LDA, Lasso outperformed all of the methods with the help of variable selection. Initially, the dataset was trained using lasso (*glmnet*). It is very slow for this dataset and it poorly performed with the accuracy of 75%. After trying Ridge Regression for binomial classification, the accuracy came down a bit. Since regularized logistic regression like Lasso works better for sparse matrices, that was the personal choice for this dataset. SVM gave stack overflow error because of the high dimensionality. Likewise for LDA and QDA. Logistic regression performed poorly after letting it run for a long time. *biglasso*, an R package with Parallel computing support, improved the accuracy to 80%[2]. Apart from this, Some linear models for classification with different penalties like L1- or L2-regularized logistic regression, L1- or L2-regularized L2-loss support vector classification, L2-regularized L1-loss support vector classification and multi-class support vector classification[3] were used to assess the error rate of the method. In addition to that, after feature selection, all the above mentioned algorithms worked with only 20 variables in the dataframe. Principle Component Analysis for dimensionality reduction and the aforementioned algorithms were combined to improve the accuracy.

Besides the large dimension, choosing the tuning parameters was very difficult. The method were tuned to fit the 20 variables after variable selection. The best performing

Table 1: Private Leaderboard Accuracy

	Kaggle Private Score			
	Lasso	Ridge	Adap-Lasso	SVM
Without Variable Selection	0.79404	0.67500	0.60952	Stack Overflow
With Variable Selection	**0.97857**	-	-	0.97500

	Kaggle Private Score			
	biglasso	Elastic-Net	logReg	Penalized LogReg
Without Variable Selection	0.81428	0.70714	Stack Overflow	Stack Overflow
With Variable Selection	-	-	0.97857	0.67380

model was chosen using the cross validation and the classification error rate. Most of the algorithms fit were assessed using classification error and the best model with very less error was chosen to be the best performing model.

0.3 Ensembles

After the variable selection, Classification trees and boosting methods were implemented for this dataset after variable selection. Classification trees and Gradient boosting performed well. Adaboost and Logitboost didn't perform well to the dataset compared to other ensemble methods. Gradient boosting help choose the variables by the variable important graph. The recursive feature selection helped judge better about the variables selected. The 20 variables selected had very less errors. Before choosing the model and assess better, these 20 variables were selected from the dataset.Although these ensemble methods performed well, the regularized logistic regression worked better for the dataset. The model accuracy is very high compared to the ensemble methods (96%).

0.4 Model Selection and Assessment

Before the variable selection, model selection and assessment was very difficult and time consuming. Almost all the methods implemented failed with the dataset due to very large dimension. To get an idea about how the models implemented work, the whole dataset is being used for model fitting. The system crashed. Trying different models for dimensionality reduction like PCA, stepwise selection and information criterias like AIC, BIC. These methods fail to work for big data. With the help of parallel computing, using modified BIC and stepwise procedure, the 20 variables were selected. Filtering these 20 variables from the dataset, the model fitting was done only with these variables. Data is partitioned into training and validation set. Based on the less classification error and the repeated CV, the tuning parameters were selected from the best model. The best model selected is trained with these selected tuning parameters on the validation set to assess the classification error. The Lasso model for classification performed really well and the classification accuracy on the validation set is 98%.

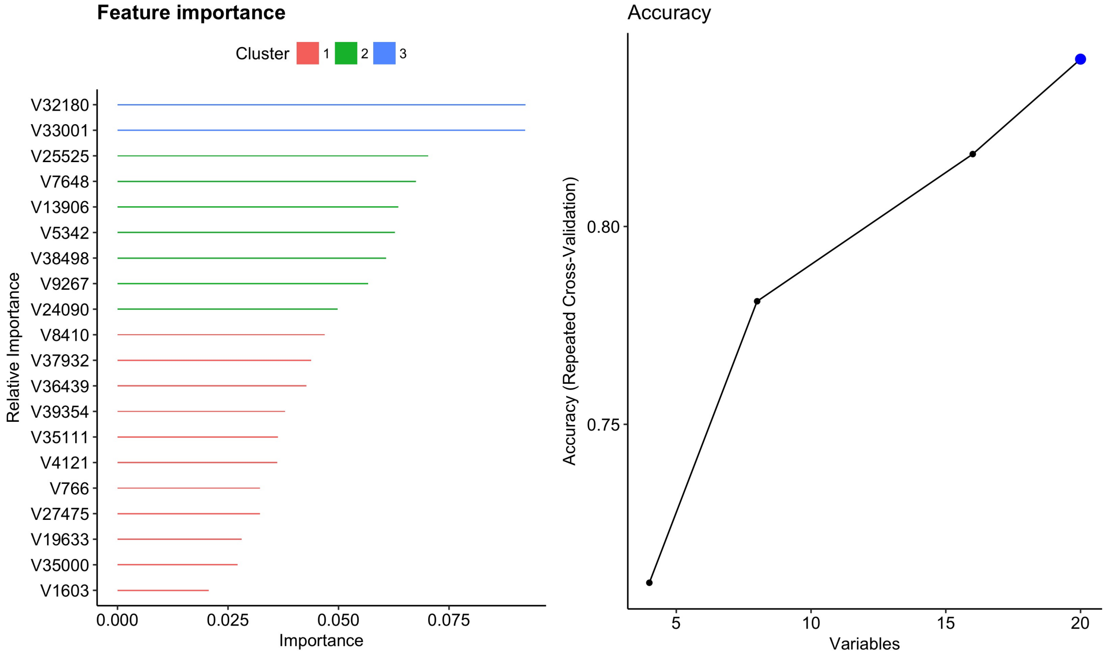
0.5 Best Scoring Model

Lasso for classification with the modified stepwise procedure for variable selection[4] was the best scoring model. It improved the accuracy to 98%. SVM performed equally well, but the model fitting was unstable and the error rate kept changing for this dataset. Going with the choice of lasso, the reproducibility was a challenge in the end. But implementing repeated CV and LOOOCV by creating different CV folds for each iteration helped to judge the model accuracy better. The total number of iterations was 10. It worked really well.

0.6 Creative Solutions and Interesting Findings

One of the surprising thing to note is, the dimensionality is reduced to 40,000 variables to only 20 variables. Only with these 20 variables, the accuracy was achievable. By adding different variables to this buffer based on the correlation and information gain, the model accuracy dropped down. The correlation with class variable was calculated and the variables with next higher correlation and lesser p-value apart from these variables were added into the dataset along with these 20 features. A point to note here is, the correlation and p-value takes a lot of loop and iterations. So using *bigmemory* and *ff* in R for the dataset and *foreach* in R's *Parallel* package, the correlation matrix was calculated block by block by splitting the dataset[4][5]. The process took lesser time than the traditional looping. The model accuracy kept decreasing and it never achieved the score of 98%. It went down below the highest accuracy. The inference finally made was, these cohorts gave rise to the best accuracy and there is not other combination of variables that is improving the model accuracy.

Figure 1: Variable Importance and Accuracy for 20 Variables



0.7 Conclusion

Variable Selection is one of the most Important steps in terms of dimensionality reduction for building a robust system. To be able to achieve different criterias and different procedures of variable selection methods apart from the traditional Stepwise Procedure for big dataset, a method or an R Package can be implemented using Parellel Computing methods to speed up the computation for big datasets. A vector implementation with *big.matrix* type can be implemented as a function for Filter based approach like correlation or Discriminant Analysis and for Embedded methods like Lasso Regression with the built in support for feature selection options.

0.8 References

1. https://cran.r-project.org/web/packages/bigstep/bigstep.pdf

2. https://cran.r-project.org/web/packages/biglasso/biglasso.pdf

3. https://cran.r-project.org/web/packages/LiblineaR/LiblineaR.pdf

4. https://cran.r-project.org/web/packages/bigmemory/bigmemory.pdf

5. https://cran.r-project.org/web/packages/ff/ff.pdf

6. M. Bogdan, J.K. Ghosh, R.W. Doerge (2004), ”Modifying the Schwarz Bayesian Information Criterion to locate multiple interacting quantitative trait loci”, Genetics 167: 989-999.

7. F. Frommlet, F. Ruhaltinger, P. Twarog, M. Bogdan (2012), ”A model selection approach to genome wide association studies”, Computational Statistics and Data Analysis 56: 1038-1051.

8. F. Frommlet, A. Chakrabarti, M. Murawska, M. Bogdan (2011), ”Asymptotic Bayes optimality under sparsity for generally distributed effect sizes under the alternative”. Technical report at arXiv:1005.4753.