A Novel feature selection paradigm in identifying prominent variables from a high-dimensionality dataset

K.Chaitanya Deepthi[1], B.Homer.Benny[2]
Sir C. R. Reddy College of Engineering, Andhra Pradesh, India

Abstract - Feature selection methods have been widely used in high dimensionality datasets. Of all supervised feature selection methods, LASSO was implemented on a dataset of 76 PTP1b inhibitors with dimensions being n=76 and p=358. Feature selection is a very important step in data analysis and more importantly, to gain insight into inherent features of data. Initially OLS and ridge regression analysis was performed followed by LASSO regression. The lambda minimum and standard error were found to be 0.06 and 0.339 respectively. The mean square error (MSE) was used to compare the analysis among OLS, ridge and lasso methods. The MSE of lasso was found to be 0.16 which is much lower than OLS or ridge regression methods. The coefficients obtained from LASSO method are computed and it was observed that 8 explanatory variables are selected as important such as EV, Dx, Dy, TL, Lx, Ly, MR and KC3. Further, the coefficients at different steps of regression were analyzed by least angle regression, forward stagewise, and forward stepwise algorithms.

Keywords: LASSO, shrinkage, regression, coefficients, lambda

I. INTRODUCTION

The feature selection is the process that chose reduced number of explanatory variable to describe a response variable. The variable selection is even more important for the high-dimensional datasets; here the number of features is very high [1]. On the other hand, it is difficult, due to dimensionality issues, to build and interpret a model that takes into consideration all the variables. For these reasons the feature selection is an important task [2]. Feature selection models are easier to interpret as the method employed removes redundant variables. The over fitting is reduced by eliminating irrelevant variables that are not associated with the response variable, which enables algorithm to run faster and handles high-dimensional data [3]. In the literature several types of methods are reported to perform feature selection paradigm. Feature selection algorithms can be categorized into supervised [4], unsupervised [5] and semi-supervised feature selection [6]. Supervised feature selection methods can further be broadly categorized into filter models, wrapper models and embedded models. First the *Filter Methods* select the features by ranking them on how useful they are for the model, to compute the usefulness score statistical test and correlation results are used. Secondly *Wrapper Methods* generates different subsets of features, each sub- set is then used to build a model and train the learning algorithm. The best subset is selected by testing the algorithm. To select the features for the subsets different criteria are used (e.g. Forward and Backward selection). Finally, the *Embedded Methods* are a combination between the two previous methods. LASSO (Least Absolute Shrinkage and Selection Operator) is an example of Embedded method which performs regularization and feature selection [7]. Embedded methods have the advantage that they include the interaction with the classification model, while at the same time being far less computationally intensive than wrapper methods [8]. The method applies a shrinking (regularization) process where it penalizes the coefficients of the regression variables shrinking some of them to zero. During features selection process the variables that still have a non-zero coefficient after the shrinking process are selected to be part of the model. An advantage of LASSO is shrinking and removing the coefficients can reduce variance without a substantial increase of the bias. The tuning parameter λ controls the strength of the penalty. When λ is sufficiently large then coefficients are forced to be exactly equal to zero. Moreover, the bias increases and variance decreases when λ increases [9]. Therefore, LASSO helps to increase the model interpretability by eliminating irrelevant variables that are not associated with the response variable, which would otherwise reduce overfitting. Feature selection is a very important step in data analysis and more importantly, to gain insight into inherent features of data. The initial step of data analysis is to select the relevant features or chose a model which automatically identifies the relevant features. After collecting the data and extracting the features, the relevant features are selected.

In this paper, we report application of feature selection algorithm by LASSO on a dataset of 76 PTP1b inhibitors with dimensions being n=76 and p=358 respectively.

II. MATERIALS AND METHODS

Dataset

A dataset of anti-diabetic inhibitors that are intended to interact and bind with specific protein target such as Protein tyrosine phosphatase 1B (PTP1B) were extracted from

literature. Further, the bio activity data of 76 inhibitory compounds are treated as response variable (dependent variable) and nearly 358 properties of compounds comprising 2-dimensional and/or 3-dimensional features are considered as explanatory (independent) variables. These variables explain how the response variable is influenced by the change in property values. Certain independent variables show positive correlation within themselves or with response variable, while few may be negative or neutral.

LASSO algorithm

Lasso was originally formulated for least squares models, however, lasso regularization is easily extended to a wide variety of statistical models including generalized linear models, generalized estimating equations, proportional hazards models, and M-estimators [10]. Lasso's ability to perform subset selection relies on the form of the constraint and has a variety of interpretations including in terms of geometry, Bayesian statistics, and convex analysis [11].

Glmnet is a package that fits a generalized linear model via penalized maximum likelihood. The regularization path is computed for the lasso or elastic net penalty at a grid of values for the regularization parameter lambda. The algorithm is extremely fast, and can exploit sparsity in the input matrix. It fits linear, logistic and multinomial, poisson, and Cox regression models. A variety of predictions can be made from the fitted models. It can also fit multi-response linear regression. LASSO is widely recognized to be the alternative in solving high-dimensional problems, where high-dimensional refers to number of unknown parameters to be estimated, p, is of much larger order than the number of observations, n.

III. RESULTS AND DISCUSSION

The process of generating features from the raw data is called feature extraction. Extracting features becomes difficult when the data has more number of features than the observations. The dataset of PTP1b inhibitors with 15 independent variables were selected to perform linear regression analysis. A model matrix was constructed with x and y parameters followed by creating a vector of lambda values. An OLS model was constructed and the following coefficients are obtained where it was observed from regression data that an R-squared value of 0.854 and F-value 23.4 were found to be reasonable as the data analyzed was devoid of outliers (n=76).

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-2.337e+00	2.909e-01	-8.034	4.28e-11	***
MV	3.034e-03	6.010e-03	0.505	0.615450	
EV	-6.365e-05	4.844e-05	-1.314	0.193854	
TD	6.391e-02	2.358e-02	2.711	0.008739	**
Dx	-1.546e-02	2.269e-02	-0.681	0.498479	
Dy	3.501e-02	1.867e-02	1.875	0.065639	.
Dz	-2.624e-02	2.343e-02	-1.120	0.267295	
TL	5.514e-03	1.231e-02	0.448	0.655813	
Lx	1.115e-02	1.098e-02	1.016	0.313835	
Ly	-3.527e-02	9.579e-03	-3.682	0.000498	***
Lz	-1.998e-02	1.223e-02	-1.634	0.107476	
MR	5.710e-03	1.571e-02	0.363	0.717579	
KC0	-6.413e-01	2.280e-01	-2.813	0.006619	**
KC1	1.037e+00	5.182e-01	2.001	0.049954	*
KC2	-3.018e-01	3.993e-01	-0.756	0.452720	
KC3	9.924e-01	4.923e-01	2.016	0.048283	*

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3365 on 60 degrees of freedom
Multiple R-squared: 0.854, Adjusted R-squared: 0.817
5
F-statistic: 23.4 on 15 and 60 DF, p-value: < 2.2e-16

Regular OLS is able to determine few variables as significant such as TD, Dy, Ly, KC0, KC1 and KC3. Further, ridge regression was performed in order to compare the OLS regression result. The lambda minimum and standard error were found to be 0.06 and 0.339 respectively.

The coefficients of regression from ridge method are comparable with OLS method. However, to improve the estimate, ridge regression was performed with a subset of dataset and the best lambda minimum was obtained. The mean square error (MSE) was used to compare the analysis among OLS, ridge and lasso methods. In ridge regression, a penalty was added by tuning parameter called lambda which is chosen using cross validation which makes the fit apparent by producing small residual sum or squares while adding a shrinkage penalty. The shrinkage penalty is refers to the lambda times the sum of squares of the coefficients, which means that the large coefficients are penalized. As lambda approaches high value, the bias is unchanged but the variance reduces. The main drawback of ridge regression is that it does not select variables instead includes all of the variables in the final model.

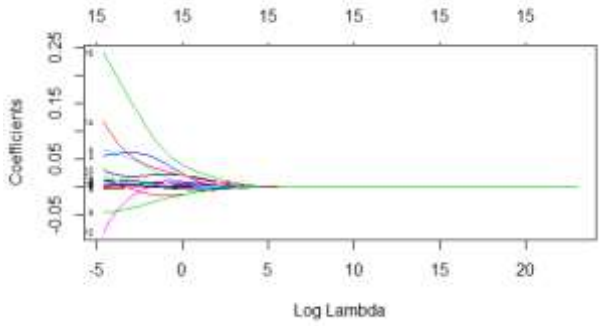

Figure 1: Ridge regression plot of coefficients vs lambda

The above plot (Figure 1) shows that when lambda values get small, it gets unregularized. However, a ridge cross validation picks the best value for lambda and the resulting plot indicates that the unregularized full model does pretty well in this case (Figure 2).

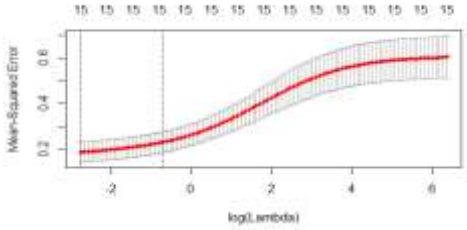

Figure 2: Cross-validation analysis of ridge regression

With s=bestlam, in other words, with best lambda minimum, the coefficients obtained are more reasonable than earlier runs for ridge regression. Data obtained suggests that the coefficient estimates are more conservative and lambda minimum provided better coefficients than the remaining.

Further, lasso was implemented where the shrinkage term was employed as the lasso takes the absolute value of the coefficient estimate. In other words, the penalty is the sum of the absolute values of the coefficients. Lasso method shrinks the coefficient estimates towards zero and when lambda is large it sets variables exactly equal to zero while ridge does not. Hence, much like the best subset selection method, lasso performs variable selection. The tuning parameter lambda is chosen by cross validation. When lambda is small, it results in least squares estimates and as lambda increases, shrinkage occurs and the variables at zero are therefore excluded. An advantage of lasso is that it is a combination of both shrinkage and selection of variables. When a dataset has large number of features, lasso finds an efficient sparse model which involves only a small subset of the features.

The MSE of lasso was found to be 0.16 which is much lower than OLS or ridge regression methods (Table 1), which means that the LASSO method is superior to the remaining.

Table 1: MSE (Mean Square Error) estimates of three methods.

Method	**MSE**
OLS	0.233
Ridge Regression	0.194
LASSO	0.163

Further, the coefficients obtained from LASSO method are computed and it was observed that 8 explanatory variables are selected as important such as EV, Dx, Dy, TL, Lx, Ly, MR and KC3. The data can be evidenced by a plot between coefficients and log lambda (Figure 3). The plot in Figure 4 suggests the deviance explained by the model which is similar to R-squared value.

```
coef(lasso.mod, s=best_se))
16 x 1 sparse Matrix of class "dgCMatrix"
                    1
(Intercept) -2.666451e+00
MV           .
EV          -6.583015e-05
TD           .
Dx           3.856798e-02
Dy           4.683199e-02
Dz           .
TL           9.176919e-03
Lx           .
Ly          -4.137166e-02
Lz           .
MR           1.269523e-02
KC0          .
KC1          .
KC2          .
KC3          1.356044e-01
```

The output shows that *only* those variables that had determined to be significant on the basis of p-values have non-zero coefficients. The coefficients of all other variables have been set to zero by the algorithm. Lasso has reduced the complexity of the fitting function. The simpler function (8 non-zero coefficients) should be preferred than the original one (15 non-zero coefficients) because it is less likely to overfit the training data.

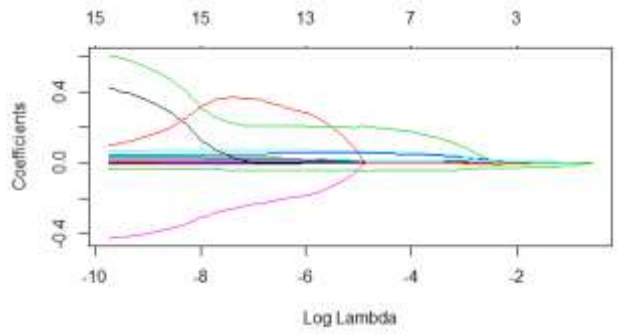

Figure 3: LASSO log lambda Vs coefficients

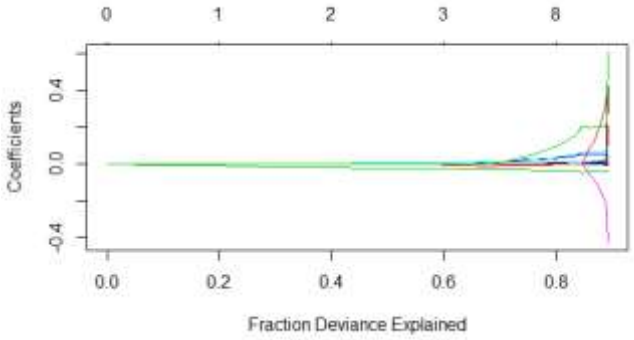

Figure 4: Fraction deviance similar to R-squared explained by the model.

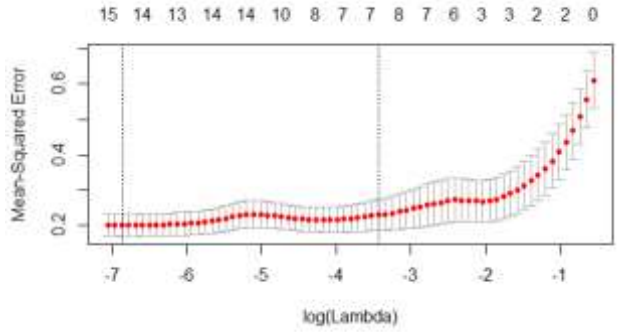

Figure 5: Cross-validation analysis of LASSO regression

The plot in Figure 5 shows that the *log* of the optimal value of lambda (i.e. the one that minimizes the root mean square error) is approximately -7.

The exact value can be viewed by examining the variable *lambda_min* and the objective of regularisation is to balance accuracy *and* simplicity. From the analysis it was observed that the value of lambda that gives the simplest model was found to be 0.00103 but also lies within one standard error of the optimal value of lambda, 0.03236, respectively.

LASSO is less prone to overfitting and hence generalizes better. The sequence of lasso moves is given below and the coefficients that obtained certain values at certain steps given in Figure 6.

Sequence of LASSO moves:

Ly MR Lx TL KC2 Lz TD Dy EV Lx Dz KC3 KC0 Lx Dx KC1 KC2 MV KC2

Var 9 11 8 7 14 10 3 5 2 -8 6 15 12 8 4 13 -14 1 14

Step 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19

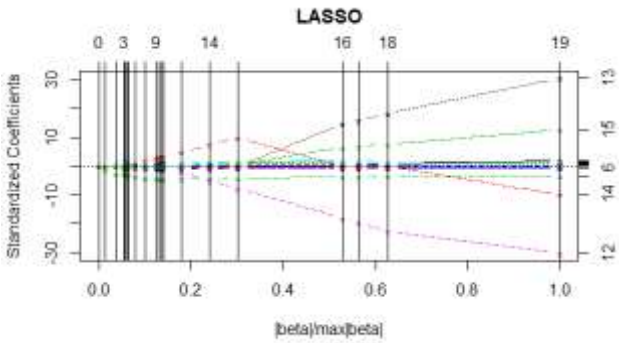

Figure 6: The coefficients that obtained certain values at certain steps of LASSO.

Further, least angle regression (LAR) [12] was employed to study the coefficients at different steps for the PTP1b data set fitted by LASSO, least angle regression, forward stagewise, and forward stepwise algorithms. Hence, the different solution paths for different methods are listed (LASSO, least angle regression, forward stagewise, and forward stepwise) below and it was observed that LASSO performed better than other methods.

Sequence of LAR moves:

Ly MR Lx TL KC2 Lz TD Dy EV Dz KC3 KC0 Dx MV KC1

Var 9 11 8 7 14 10 3 5 2 6 15 12 4 1 13

Step 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15

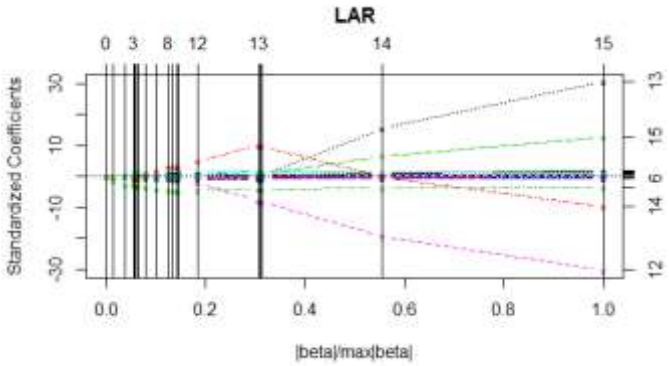

Figure 7: The coefficients obtained at certain steps of LAR.

Sequence of Forward Stagewise moves:
Ly MR Lx TL Lx Lx Lx KC2 MR Lz TD Dy KC3 KC2 E V Dx KC2 Dz Lx TD Dx TD TD MV
Var 9 11 8 7 -8 8 -8 14 -11 10 3 5 15 -14 2 4 14 6 8 -3 -4 3 -3 1
Step 1 2 3 4 4 5 5 6 6 7 8 9 10 10 11 12 13 14 15 15 15 16 16 17
KC3 KC3 KC3 KC0 Ly TL Lz EV EV Lz EV Lz Dx MV TD MR TL MV KC1 MR Ly Lz Dy Dy
Var -15 15 -15 12 -9 -7 -10 -2 2 10 -2 -10 4 -1 3 11 7 1 1 3 -11 9 10 -5 5
Step 17 18 18 19 19 19 19 19 20 20 20 20 21 21 22 23 24 25 26 26 27 28 28 29
Dy KC3 KC2 MV Dy EV KC2 TL EV EV EV MV Lx EV MR TL Lx
Var -5 15 -14 -1 5 2 14 -7 -2 2 -2 1 -8 2 11 7 8
Step 29 30 30 30 31 32 33 33 33 34 34 35 35 36 37 38 39

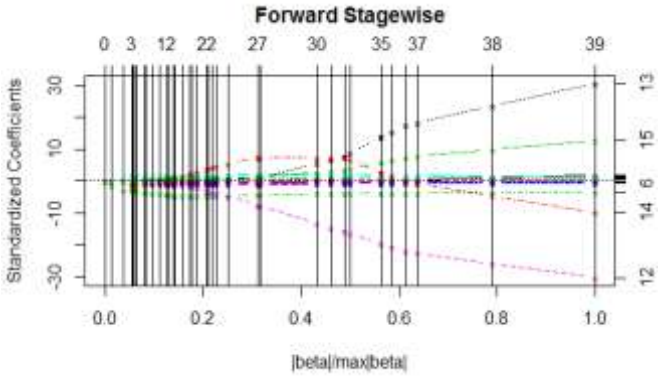

Figure 8: The coefficients that obtained at different steps by forward stagewise.

Sequence of Forward Stepwise moves:
Ly MR TD TL Lz KC3 Dy EV Dz Dx MV Lx KC2 KC0 KC1
Var 9 11 3 7 10 15 5 2 6 4 1 8 14 12 13
Step 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15

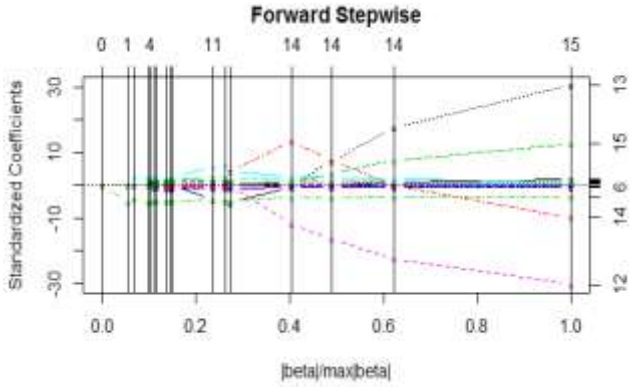

Figure 9: The coefficients that obtained at different steps by forward stepwise algorithm.

IV. CONCLUSION

An attempt has been made to implement LASSO method in extracting features from a high dimensional dataset of PTP1b inhibitors. On a comparative note, OLS method was able to determine few variables as significant such as TD, Dy, Ly, KC0, KC1 and KC3 whereas LASSO resulted in EV, Dx, Dy, TL, Ly, MR and KC3. Evaluation of Mean Square error among three methods resulted in MSE of lasso to be 0.16 which is much lower than OLS or ridge regression methods. Finally, it was observed that LASSO is less prone to overfitting and hence generalizes better. The sequence of lasso moves and the coefficients that obtained certain values at certain steps when compared with the PTP1b data set fitted by least angle regression, forward stagewise, and forward stepwise algorithms showed that the LASSO performed better than other methods.

V. REFERENCES

[1]. Forman G. An extensive empirical study of feature selection metrics for text classification, J. Mach. Learn. Res., 2003, vol. 3 (pg. 1289-1305)
[2]. Daelemans W, et al. Combined optimization of feature selection and algorithm parameter interaction in machine learning of language, 2003Proceedings of the 14th European Conference on Machine Learning (ECML-2003)(pg. 84-95).
[3]. Daelemans W, et al. Combined optimization of feature selection and algorithm parameter interaction in machine learning of language, 2003Proceedings of

the 14th European Conference on Machine Learning (ECML-2003)(pg. 84-95).

[4]. J. Weston, A. Elisseff, B. Schoelkopf, and M. Tipping. Use of the zero norm with linear models and kernel methods. Journal of Machine Learning Research, 3:1439–1461, 2003.

[5]. J.G. Dy and C.E. Brodley. Feature selection for unsupervised learning. The Journal of Machine Learning Research, 5:845–889, 2004,

[6]. Z. Zhao and H. Liu. Semi-supervised feature selection via spectral analysis. In Proceedings of SIAM International Conference on Data Mining, 2007.

[7]. Tibshirani, Robert (1996). "Regression Shrinkage and Selection via the lasso". Journal of the Royal Statistical Society. Series B (methodological). Wiley. 58 (1): 267–88.

[8]. Yvan Saeys et al. A review of feature selection techniques in bioinformatics. Bioinformatics, Volume 23, Issue 19, 1 October 2007, Pages 2507–2517.

[9]. Tibshirani, Robert (1997). "The lasso Method for Variable Selection in the Cox Model". Statistics in Medicine. 16: 385–395.

[10]. Tibshirani, Robert (1996). "Regression Shrinkage and Selection via the lasso". Journal of the Royal Statistical Society. Series B (methodological). Wiley. 58 (1): 267–88.

[11]. Tibshirani, Robert (1997). "The lasso Method for Variable Selection in the Cox Model". Statistics in Medicine. 16: 385–395.

[12]. Efron, Hastie, Johnstone and Tibshirani (2003) “Least Angle Regression”, Annals of Statistics