Paper CC20

VIF Regression: A SAS® Application to Feature Selection in Large Data Sets

Ruiwen Zhang, SAS Institute Inc.; Feng Liu, Univ. of North Carolina at Chapel Hill

ABSTRACT

Data size and dimensionality can grow big easily in data mining problems. Feature selection plays an increasingly crucial role in modern industry to improve predictive model interpretability, avoid overfitting and multicollinarity. More important is that we need the feature selection to be done fast in big data. VIF regression is a fast algorithm which does feature selection in large regression problems. VIF regression handles big number of features streamwise. Such streamwise regression method has its advantages over traditional stepwise regression as it offers faster computational speed without loss of its accuracy. We implement the algorithm in SAS language and provide a comprehensive example so that SAS users can get benefits from SAS platform or server which usually stores their big data sets and also from the VIF regression, a much-needed fast feature selection for large data sets.

Keywords: Variable selection; Stepwise regression; Variance Inflation Factor (VIF); Model selection

INTRODUCTION

Big Data containing millions of observations and huge number of features are quite common in data mining problems, especially from such areas as gene sequencing, sensor data, image processing and finance related data, etc. Though R-Squared or similar measure of goodness of fit for regression always increase as the number of predictors grows, modeling with reasonable number of explanatory variables is more desirable because the issues of collinearity and overfitting. Parsimonious models also offer better interpretability which is critical in real business. There are numerous model selection approaches and selection criteria can be based on prediction, fit, etc. However, it is most desired to develop and implement fast algorithms applicable to real big datasets. Stepwise regression using criteria such as he AIC (Akaike, 1973), BIC (Schwarz, 1978), Mallow's Cp (Mallows, 1973), cross-validations, etc., can be very slow for large sets because all remaining variables are evaluated at each stage. Fast algorithms for stepwise regressions are also available, for example, Lasso/LAR (Efron et al, 2004), the Dantzig Selector (Candes and Tao, 2007), or coordinated descent (Friedman, Hastie and Tibshirani, 2010). But the speed is not scalable because the penalty λ needs to be computed and often via cross-validation (Dupuis and Victoria, 2013). VIF regression proposed by Lin, Foster and Ungar in 2011 uses an improved streamwise regression approach, which search over the predictors only one-pass, and a computationally efficient method of testing each potential predictor for addition to the model. For a classical linear model,

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

With n observations $\mathbf{y}= (y_1, \dots, y_n)'$ and p predictors $\boldsymbol{x}_1, \dots, \boldsymbol{x}_p$, $p \gg n$, where $\boldsymbol{X} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_p)$ is an $n \times p$ design matrix of features, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$ is the vector of coefficient parameters, and error $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$. The computation complexity of VIF is $O(pn)$ under the sparsity assumption that only a subset of k of the p predictors in (1) has nonzero coefficients, and $k \gg p$ (Miller, 2002). This property enables the VIF regression to handle larger datasets as illustrated by Lin, Foster and Ungar in 2011. VIF regression algorithm also guarantee good control of the marginal false discovery rate (mFDR) (Foster and Stine, 2008) with no overfitting.

In this paper, you will see a brief introduction to VIF Regression algorithm as well as a real data example which tests the performance of VIF regression as well as some comparisons, like LARS and stepwise regression.

VIF REGRESSION

We can formulate variable selection algorithms generally as estimating $\boldsymbol{\beta}$ that minimize the penalized sum of squared errors

$$\textbf{argmin}\{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda_q \|\boldsymbol{\beta}\|_{l_q}\}, \tag{2}$$

where $\|\boldsymbol{\beta}\|_{l_q} = (\sum_{i=1}^{p} |\beta_i|^q)^{1/q}$ for $q > 0$ and $\|\boldsymbol{\beta}\|_{l_0} = \sum_{i=1}^{p} I_{\{\beta_i \neq 0\}}$.

Stepwise regression like forward selection, backward selection and forward-backward combination evaluates variables only using marginal correlations stagewise. Also its siblings such as LASSO and LARS, the small step-size forward stagewise regression using l_1 norm, they suffer from collinearity among the predictors. VIF regression

corrects this bias by pre-sampling a small set of data to compute the variance inflation factor (VIF) of each candidate.

Another drawback of stepwise regression is the computational complex. Optimally resolving (2) with a l_0 penalty requires searching over all 2^p possible subsets, which is NP hard (Natarajan, 1995) and thus computationally expensive even when p is small. VIF regression incorporates its evaluation step with a streamwise regression algorithm using an α-investing rule. Streamwise regression (Zhou et al., 2006) evaluates each candidate variable on a single pass, and so is extremely fast.

VIF algorithm is essentially characterized by the following two components:

- Evaluation step: approximate the partial correlation of each candidate variable with the response variable by correcting the marginal correlation using variance inflation factor (VIF) calculated from a small pre-sampled set.
- Search step: test each variable sequentially using an α-investing rule (Foster and Stine, 2008). Variables will be added only when they are able to pay the price of reducing a statistically sufficient variance in the predictive model. The α-investing rule guarantees no model overfitting and provides highly accurate models.

1. α-Investing and Sequential Testing

An α-investing rule is an adaptive, sequential procedure for testing multiple hypotheses (Foster and Stine, 2008). The rule works as follows. Suppose that this is a game with a series of tests. A gambler begins his game with initial wealth, w_0; intuitively, this is an allowance for type I error. In the ith test (game), at level α_i, if a rejection is made, then the gambler earns a pay-out Δw; otherwise, his current wealth w_i will be reduced by $\alpha_i/(1-\alpha_i)$. The test level α_i is set to be $w_i/(1+i-f)$ where f is the time at which the last hypothesis was rejected. Thus, once the gambler successfully rejects a null hypothesis, he earns more to spend the next few times. Furthermore, the game becomes easier to play in the near future, in the sense that α_i will remain inflated in the short term. The game continues until the player goes bankrupt, that is, $w_i \leq 0$.

2. Fast Evaluation Procedure

At each step of the regression, suppose that a set of predictors, $C = \{\boldsymbol{x}_1, \dots \boldsymbol{x}_k\}$, has been chosen in the model. We assume that all of the variables $\boldsymbol{x}_i$ are centered.

i) Obtain residuals $\boldsymbol{r} = \boldsymbol{y} - \boldsymbol{X}_C(\boldsymbol{X}_C'\boldsymbol{X}_C)^{-1}\boldsymbol{X}_C'\boldsymbol{y}$ and RMSE $\hat{\sigma}_{null} = \|\boldsymbol{r}\|/\sqrt{(n-|C|-1)}$ from the previous step.

ii) Sample a small subset $I = \{i_1, \dots, i_m\} \subset \{1, \dots, n\}$ of observations; let $\boldsymbol{x}^I$ denote the corresponding subsample of predictors $\boldsymbol{x}$.

iii) Fit $\boldsymbol{r}$ on $\boldsymbol{x}_{new}/\|\boldsymbol{x}_{new}\|$ and compute the coefficient estimate $\hat{\gamma}_{new} =< \boldsymbol{r}, \boldsymbol{x}_{new} >/\|\boldsymbol{x}_{new}\|$.

iv) Fit $\boldsymbol{x}_{new}^I$ on $\{\boldsymbol{x}_1^I, \dots, \boldsymbol{x}_k^I\}$ and compute $\boldsymbol{R}_I^2 = \boldsymbol{x}_{new}'\boldsymbol{X}_C^I \times ((\boldsymbol{X}_C^I)'\boldsymbol{X}_C^I)^{-\mathbf{1}}(\boldsymbol{X}_C^I)'\boldsymbol{x}_{new} >/ \|\boldsymbol{x}_{new}\|^2$.

v) Compute and return the approximate t-ratio as $\hat{t}_{new} = \hat{\gamma}_{new}/(\hat{\sigma}\sqrt{1-\boldsymbol{R}_I^2})$.

3. Streamwise Variable Selection

Using an α-investing rule allows us to test an infinite stream of hypotheses while controlling mFDR. In the context of variable selection, this implies that we may order the variables in a sequence (possibly dynamically) and include them into the model in a streamwise manner without overfitting.

VIF regression procedure is stated in Figure 1. The ability to test the variables in a streamwise way has many advantages. First, the one-pass algorithm can save a great amount of computation if the data are massive. In most search algorithms, adding each new variable necessitates going through the whole space of candidates; the computation is expensive if the data size, $n \times p$, is huge. VIF regression alleviates this burden by reducing the loops to only one round. Second, this allows one to handle dynamic variable sets. These include the cases

where p is extremely large or unknown, resulting in a problem in applying static variable selection criteria. This also allows one to first test the lower-order interactions and then decide which higher-order interactions need to be tested.

The boosted Streamwise Regression using an $\boldsymbol{\alpha}$-investing

Input: data $\boldsymbol{y}$, $\boldsymbol{x_1}$, $\boldsymbol{x_2}$,… (centered);

Set: initial wealth $\boldsymbol{w_0 = 0.05}$ and pay-out $\Delta\, \boldsymbol{w = 0.05}$, and subsample size $\boldsymbol{m}$;

$$\boldsymbol{C = \{0\}; r = y - \bar{y}; \hat{\sigma} = sd(y); i = 1; w_1 = w_0; f = 0.}$$

Sample: $\boldsymbol{I = \{i_1, \ldots, i_m\} \subset \{1, \ldots, n\}}$

Repeat

set threshold $\boldsymbol{\alpha_i = w_i/(1 + i - f)}$

attain $\boldsymbol{\hat{t}_i}$ from the Fast Evaluation Procedure

if $\boldsymbol{2\Phi(|t_i|) > 1 - \alpha_i}$ //compute p-value to threshold

then $\boldsymbol{C = C \cup \{i\}}$ //add feature to model

update $\boldsymbol{r = y - \hat{y}_C}$, $\boldsymbol{\hat{\sigma} = RMSE_C}$

$\boldsymbol{w_{i+1} = w_i + \Delta w}$, $\boldsymbol{f = i}$

else $\boldsymbol{w_{i+1} = w_i - \alpha_i/(1 - \alpha_i)}$

end if

i=i+1

until maximum CPU time or memory is reached

Figure 1. VIF Regression Algorithm

REAL APPLICATION IN SAS

We present a real analysis using the communities and crime data available at https://archive.ics.uci.edu/ml/datasets/Communities+and+Crime+Unnormalized. This data set combines socio-economic data from 90' Census, law enforcement data from 1990 Law Enforcement Management and crime data from 1995 FBI UCR. It consists of 2215 observations on 125 predictors, and the regression model is to predict the number of assaults in 1995. In the EM flow as shown in Fig. 1, we make comparable results for stepwise regression, LARS and VIF regression. In the comparison, VIF regression and stepwise regression select similar number of variables, however, the results from stepwise regression implies overfitting problem. LARS yields the biggest Average Squared Error in both train and validation data after 20 iteration steps.

Figure 2 demonstrates that VIF Regression outperforms the other two in terms of providing the minimum Average Square Error from validation. It selects 20 predictors in the setting of m=200, $\boldsymbol{w_0 = 0.05}$ $\Delta\, \boldsymbol{w = 0.05}$. Stepwise regression using AIC selection criterion was able to scale down the set of predictors to 21; however, it suffers the overfitting problem. It has the minimum error in train data among all the three methods, 147205.3, while the second lowest is 157860.5. In validation data, in contrast, its error is obviously higher than VIF regression.

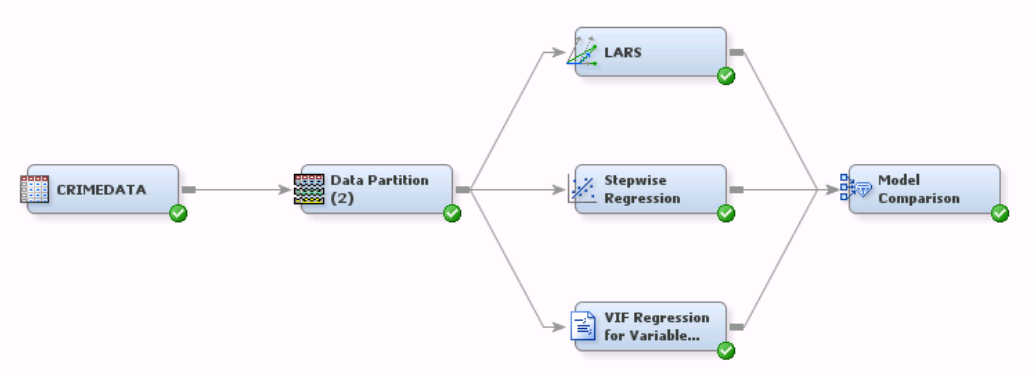

Figure 2. EM flow for the comparison analysis

Fit Statistics

Selected Model	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Average Squared Error	Train: Akaike's Information Criterion	Train: Average Squared Error	Train: Average Error Function	Train: Degrees of Freedom for Error	Train: Model Degrees of Freedom
Y	VIF Regres...	VAR136		258548.5	15815.76	157860.5	157860.5	1298	20
	Stepwise R...	VAR136		264309	15725.65	147205.3	147205.3	1297	21
	LARS	VAR136		393343.5	.	175166.6	.	.	.

Output 1. Output from Model Comparison node

Multicollinarity is a known danger for causing overfitting in regression analysis. It produces large standard errors in the related independent variables and may introduce large error in prediction. Removing such data redundancy (variables with high correlation) improves statistical robustness for regression models.

CONCLUSION

VIF regression is a streamwise regression approach to select variables based on VIF and fast robust estimates. It has been proven to be an efficient algorithm in finding good subsets of variables from a huge space of candidates, and such algorithm is applicable to solve some online problems when features are generated and added to the model dynamically. We implement the algorithm using SAS and please contact us for the code.

REFERENCES

Akaike, H. (1973) “Information theory and an extension of the maximum likelihood principle”. *Second International Symposium on Information Theory* (Tsahkadsor, 1971), pp. 267–281. Akadémiai Kiadó, Budapest.

Schwarz, Gideon. (1978) “Estimating the dimension of a model”. *Ann. Statist*. 6, no. 2, 461–464.

Mallows, C. L. (1973). “Some comments on Cp”. *Technometrics* 15 661–675.

Efron, Bradley; Hastie, Trevor; Johnstone, Iain; Tibshirani, Robert. (2004) “Least angle regression”. With discussion, and a rejoinder by the authors. *Ann. Statist*. 32, no. 2, 407–499.

Candes, Emmanuel; Tao, Terence. (2007) “The Dantzig selector: statistical estimation when p is much larger than n”. *Ann. Statist*. 35, no. 6, 2313–2351.

Friedman, J. H., Hastie, T. and Tibshirani, R. (2010). “Regularization paths for generalized linear models via coordinate descent”. *Journal of Statistical Software* 33 1-22.

Dupuis, Debbie J.; Victoria-Feser, Maria-Pia. (2011) “Fast robust model selection in large datasets”. *J. Amer. Statist. Assoc.* 106, no. 493, 203–212.

Lin, Dongyu; Foster, Dean P.; Ungar, Lyle H. (2011) “VIF regression: a fast regression algorithm for large data”. *J. Amer. Statist. Assoc*. 106, no. 493, 232–247.

Foster, Dean P.; Stine, Robert A. (2008) “α -investing: a procedure for sequential control of expected false discoveries”. *J. R. Stat. Soc. Ser. B Stat. Methodol*. 70, no. 2, 429–444.

Miller, A. (2002), Subset Selection in Regression (2nd ed.), Boca Raton, FL: Chapman & Hall.

Zhou, J., Foster, D. P., Stine, R. A., and Ungar, L. H. (2006), “Streamwise Feature Selection,” *Journal of Machine Learning Research*, 7, 1861–1885.

Natarajan, B. K. (1995), “Sparse Approximate Solutions to Linear Systems,” *SIAM Journal on Computing*, 24, 227–234.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Ruiwen Zhang
SAS Institute Inc.
100 SAS Campus Dr.
Cary, NC 27513
Ruiwen.Zhang@sas.com