AL- Rafidain
University College

PISSN: (1681-6870); EISSN: (2790-2293)

Journal of AL-Rafidain
University College for Sciences

Available online at: https://www.jrucs.iq

JRUCS

Journal of AL-Rafidain
University College for
Sciences

Fast and Robust Majority Voting Algorithm for Big Data Regression Model

Hassan S. Uraibi

hassan.uraibi@qu.edu.iq

Department of Statistics - College of Administration and Economics - University of Al-Qadisiyah, Al-Qadisiyah, Iraq

Article Information

Article History:

Received: December, 20, 2022
Accepted: March, 3, 2023
Available Online: December, 31, 2023

Keywords:

Big data, Divide and conquer, Robust variable selection, Majority Vote

Correspondence:

Hassan S. Uraibi
hassan.uraibi@qu.edu.iq

Abstract

The majority voting approach is one of divide and conquers technologies that have the capacity to analyzing, understanding, and then decision making of big data. Big regression data is massive not only in terms of volume where P and n tend to infinity, but in terms of intensity, and complexity too. For instance, the collection of massive data from different subpopulations results in a heterogeneity problem that definitely leads to the presence of outliers. Moreover, tackling such a volume of data exceeds the capacity of standard analytic tools. therefore, the processing requires the proposed algorithm to have two main properties, rapid and accurate. Unfortunately, the majority voting approach is not reliable where outliers are present in the data. Furthermore, the curse of dimensionality causes the multicollinearity problem which yields misleading results This paper proposes a fast and robust majority voting approach that is a new version of the original one with some steps. The first step is to vertically divide the design matrix into a number of blocks to conquer the curse of ultra-dimensionality. Voting on choosing the best subset of variables should be considered by using a robust variable selection method as a dimensional reduction procedure and resistant to the presence of outliers. The second step is to aggregate all best subsets in one linear regression model, and then using majority vote algorithm to get the best variables e. A simulation study has done in this paper to know the performance of the proposed technique is compared with Big-lasso which is well-known as the fastest method in the statistical literature right now. The result shows outperforming of the proposed algorithm which is faster than Big-lasso and more accurate than it.

https://doi.org/10.55562/jrucs.v54i1.616

1. Introduction

Analysis of the daily streaming of data from multiple online and other sources exceeds the capacity of the known statistical analytics methods due to its volume, intensity, and complexity. This type of large-scale data is so-called Big data that formed a big challenge for the statisticians in many scientific fields. The problem of big data has given the attention of researchers in the statistical literature that presented some of the methodologies to discover its pattern and analysis. The proposed methodologies can be classified into three groups: subsampling-based, divide and conquer, and online updating for stream data. see, Bag little bootstrap (Kleiner et al. [2]), Leveraging (Ma and Sun, [3]), Majority voting (Chen and Xie [5]), and Screening with ultrahigh dimensions (Song and Liang,[6]).

It is obvious that the databases of big data are growing in volume per day, so employing statistical tools requires integration between computer networks and programming styles, such as employing parallel programming with high-performance computing. Another style is used with a high-quality computer which possesses ultra-high storage and parallel processors such as the mainframe computer with high specifications.

Regardless of the distribution and processing systems of big data such as Hadoop and spark or something else, the researcher's objective is to reduce the ultra-high dimensional of features and deal with properly the huge number of observations. One important attractive dimensional reduction method in the statistics literature is Iterated sure independence screening (ISIS), which was proposed by Fan and Lv (2008). ISIS is a very effective procedure to tackle ultrahigh dimensional feature problem. In the context of least squares regression, the SIS algorithm starts with a very simple procedure that is so-called screening. The screening means ranking features that having the best marginal correlation with the response and then pick up the top features that indexed from the first rank to the feature that has been ranked at d where $d = n/log(n)$. Lasso or SCAD can be applied in the second stage to selects the important features among d of them. Wang et al. (2016) pointed out that the ISIS method is particularly suited for big data where the number of covariates P_n is much larger than the sample size n , $P_n \gg n$, and possibly increasing with n.

The screening step of ISIS transforms the dimensionality feature space from ultrahigh P_n to ultra-low d in which the sample size n is more than d. As a result of this procedure selection of important features cannot exceed d, where $d < n$ and $d \ll p$ (Uraibi, 2020). A new cutoff point which is denoted as $d_2^* = n + \left(\frac{n}{\log(n)}\right)$ introduced by Uraibi (2020) to modify the ISIS cutoff point of screening step. With big data where $n \to \infty$, both cut-off points d and d_2^* are not relevant for reasonable dimension reduction due to this situation yields $n \gg d$. In other word, it transforms the dimensionality feature space from ultrahigh P_n to high dimension or perhaps to other extreme dimensions too. Consequently, when the screening step rely on n leads to the curse of dimensionality, multicollinearity problem which yields misleading results.

It is well known in the statistical literature that variable selection is hard to argue, the problem not only providing faster and more cost-effective predictors and improving the prediction performance, but with definitely extended to providing a better understanding of the underlying process that generated the data, see (Hesterberg *et al* ;2008, Isabelle Guyon; 2003, Uraibi et al;2015). Here, obtaining a short model with a small number of predictors can be satisfied this objective, therefore, thinking should be riveted for choosing a reasonable cutoff point to reduce the ultra-dimension of independent variables.

The VIF-regression for ultra-high feature space algorithm has been proposed by Uraibi (2020) as an alternative to the ISIS method and showed it is more efficient than it. Unfortunately, collecting large scale data from different subpopulations results-in heterogeneity problem which leads to the presence of outliers. Consequently, VIF-regression, ISIS and other statistical methods of dealing with such data would breakdown. Unfortunately, these methods are sensitive to the presence of outliers and leverage points. In spit of robust variable selection methods are presented in robust statistics literature such as Khan et al. (2007), Uraibi et al. (2017), Uraibi (2019), Uraibi and Midi (2020), but the robust VIF-regression that suggested by Dupuis et al. (2013) is adopted in this paper.

In this paper, we suggest a new methodology in three steps, first splitting features into number blocks, the block size should not exceed fifteen variables and are associated with a dependent variable. In each block, all variables are standardized and then the p-values of robust regression coefficients have to be computed. The second step is to aggregate the significant p-values of all blocks in one vector and choose only those ones that are less Bonferroni significant level. The last step is using robust VIF-regression with a new design matrix of features to avoid collinearity problem and select the best features.

2. Fast and Robust Majority Voting Algorithm (FRMV)

The main target of this algorithm is to get interpretable model for big data in short time. Moreover, this algorithm is not need to memory mapping technique or computer network for computing. It is high speed computational algorithm is implemented on core i7 personal computer with 8 GB installed RAM. It can be considered as a divide and conquer algorithm for ultra-high dimensional data sets. The FRMV algorithm can be formulated into three steps, reducing the ultra-high dimensional variable space by using screening procedure, divide the new dimension of variables into some blocks which would be used with approximate robust VIF-regression results should be conquered using majority vote, and finally the non-zero variables have to select when the average of its selection exceeds 10% using majority vote.

Marginal correlation based on screening

Marginal correlation-based ranking is widely used for variable selection methods to reduce the ultra-high dimensional variables space. Run et al (2020) proved that the marginal correlation screening may miss the true variables in the presence of correlated predictors. It is noted that the results of Run et al (2020) were associated with three cutoff points, $\left\{\frac{n}{log(n)}, n^{1/2}, n^{1/3}\right\}$, through which it was proved that some important variables exceed the cutoff points, and therefore they are not selected in this step. The size of the simulation sample did not exceed 1000 observation, in other words, the highest cut-off point does not reach 145, and there is certainly an important variable after this point that cannot be imagined less than the other cut-off points. Employing real data with sample sizes higher than the simulated sample does not prove the purpose of the researchers. This is because the results showed the superiority of the cut-off point $\frac{n}{log(n)}$ over its counterparts $\left\{n^{1/2}, n^{1/3}\right\}$, and this is inevitable because the values of the other cut-off points when the sample size 3951 did not exceed (64) and (16), respectively. From the foregoing, this paper suggests fixing the cut-off point of marginal correlation at 200 when the sample size at least 1000 observation as follows.

Consider the following linear regression model,

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

where $\mathbf{y}$ is an $n \times 1$ vector of the response variable, $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_p)$ is an $n \times p$ known design matrix of independent variables, $\boldsymbol{\beta} = \left(\beta_1, \dots, \beta_p\right)^{\mathrm{T}}$ is a $p \times 1$ vector of unknown regression coefficients, and $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of random errors with mean 0 and variance σ^2. Suppose that$\left|\hat{R}_j\ \right|$ is the absolute values of Spearman's correlations (as robust correlation method) between **X** and **y** are ordered from the largest to lowest value,$\left|\hat{R}_1^*\right| > \left|\hat{R}_2^*\right| >, \dots., > \left|\hat{R}_S^*\right| >, \dots., > \left|\hat{R}_p^*\right|$. The new cut-off point of marginal correlation equivalents to $S = \frac{P^*}{5 \times \gamma \times 10^{[\log_{10}(P^*)-3]}} \approx 200,$ where $\boldsymbol{P}^* = \mathbf{10}^{\mathbf{log_{10}}(\boldsymbol{P})}$, and $\boldsymbol{\gamma} = [\boldsymbol{P}/\boldsymbol{P}^*]$. However, the screening variables are those associated with $\hat{R}_1^*, \dots, \hat{R}_S^*$,will construct the new matrix design X^* , where $X^* = \left\{X_j : \left|\hat{R}_j^*\right| \le \left|\hat{R}_S^*\right|\right\}$.

Majority Vote

Depending on the previous step, the number of blocks can be determined by dividing $\frac{S}{20} = 10$, meaning that each block will contain 20 variables. Since these variables are arranged in descending order according to their marginal correlations with the dependent variable, the variables that have the greatest influence on the dependent variable will be in the first block. As for the second block, it may include some important variables if their number exceeds 20, or if they have a parameter that differs from zero, but their covariances with $\mathbf{y}$ are equal to zero. The probability of this case appearing in other blocks may tend to zero. Based on the foregoing, the variables of each block will be combined with the variables of the first block each time and subjected to a method robust VIF-regression. Then with this procedure being repeated with the number of (9) times and each time a random sub-sample of 1000 is drawn, randomly. The majority vote will be conditioned on a specific significant level to ensure control over the false selection error of robust VIF-regression. This procedure can be described in the following algorithm.

- Start with Screening 200 variables rely on their marginal correlations
- Partition these variables into 10 blocks, b_k where $k = 1,2,\dots,10$
- For $k = 2$ to 10
    - Let $X_b = \left(X^*_{b_1}, X^*_{b_k}\right)$ and ζ is a 1000 random samples drawn from the original one.
    - Implementing robust VIF-regression for $X_b^{(\zeta)}$ and $y^{(\zeta)}$and the best variable are selected.
    - Next
- The threshold of the algorithm of majority vote will be 0.10, the variable that has exceeded this threshold be included with variables that would construct the new model. It is obvious that the majority vote of variable selection role is a dimension reduction for the second time.
- End

3. Simulation

A simulation study similar to the simulation of Frank and Friedman (1993), Khan, Van Aelst, and Zamar (2007) , Agostinelli and Salibian-Barrera (2010) and Alfons et al. (2011), is carried out to assess the performance of our proposed FRMV method.

Assuming that the latent variables are $X_i^{(L)}$ where $i = 1,2,\dots,p$, and p is the of regressors in the true model.

In this study, we consider a multiple linear regression model

$$y = X_1^{(L)} + X_2^{(L)} + \cdots + X_p^{(L)} + \sigma\varepsilon$$

where the random error term ε are generated from a standard normal distribution $N(0,1)$, and the tuning parameter σ is chosen such that the signal-to-noise ratio is 5:

$$\sqrt{\frac{Var\left(X_1^{(L)} + X_2^{(L)} + \cdots + X_p^{(L)}\right)}{Var(\varepsilon)}} = \frac{\sqrt{p}}{\sigma} = 5$$

A set of d independent variables $X_1, X_2, \dots., X_d$ is generated using d independent standard normal variables $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_d$ as follows,

$$\begin{pmatrix} X_1 \\ \vdots \\ X_p \\ X_{p+1} \\ \vdots \\ X_{2p} \\ \vdots \\ X_{k\times p} \end{pmatrix} := \begin{pmatrix} X_1^{(L)} \\ \vdots \\ X_1^{(L)} \\ X_2^{(L)} \\ \vdots \\ X_2^{(L)} \\ \vdots \\ X_p^{(L)} \end{pmatrix} + \tau \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_p \\ \varepsilon_{p+1} \\ \vdots \\ \varepsilon_{2p} \\ \vdots \\ \varepsilon_{k\times p} \end{pmatrix}$$

The tuning parameter τ is chosen such that $\tau < \delta$ to construct the $k \times p$ collinearities variables, since k of low-noise perturbations add to each latent variables. The degree of correlation between the predictors is based on the distance between τ and δ, high distance yields high collinearity and vise versa. Therefore, the correlation of $k \times p$ should be close; since $\mathrm{cor}(X_1, X_2) = \mathrm{cor}(X_3, X_4) \dots = \mathrm{cor}(X_{k-1}, X_k) = \rho$. The parameter δ is sufficiently greater than τ to create k × L weak collinearities variables by adding high-noise perturbations to the latent variables L time. The correlations of this group is low.

The simulation design considers the degree of collinearites $[0.95 < \rho \leq 0.99\,]$, this case can be happened where the pair (τ,δ) is chosen to be $(0.05{,}25)$.

$$\begin{pmatrix} X_{k\times p+1} \\ \vdots \\ X_{k\times p+L} \\ \vdots \\ X_{k\times p+k\times L} \end{pmatrix} := \begin{pmatrix} X_1^{(L)} \\ \vdots \\ X_1^{(L)} \\ \vdots \\ X_p^{(L)} \end{pmatrix} + \delta \begin{pmatrix} \varepsilon_{k\times p+1} \\ \vdots \\ \varepsilon_{k\times p+L} \\ \vdots \\ \varepsilon_{k\times p+k\times L} \end{pmatrix}$$

We consider the remaining variables $X_{k*p+k*L+1}, \dots, X_d$ to be noise; they are generated from a standard normal distribution, where $\mathrm{d} > p \times (k+L)$, as follows,

$$\begin{pmatrix} X_i \\ \vdots \\ X_d \end{pmatrix} := \begin{pmatrix} \varepsilon_\mathrm{i} \\ \vdots \\ \varepsilon_\mathrm{d} \end{pmatrix}$$

where $\varepsilon_i \sim N(0{,}1) \quad \forall\, i = k*p + k*L + 1, \dots, d$

In order to know the performance of the RFMV algorithm with big data, various simulation scenarios are carried out in the presence of 0.05 outliers and leverage points as follows,

Case 1 : $n = 2000, p = 2, d = 1000, K = 1, L = 2$

Case 2 : $n = 2000, p = 2, d = 1000, K = 2, L = 2$

Case 3 : $n = 4000, p = 5, d = 2000, K = 1, L = 2$

Case 4 : $n = 10000, p = 10, d = 5000, K = 2, L = 2$

In order to create outliers in both $X_i^{(L)}$ and y, the first 0.05 observations of $X_1^{(L)}, X_2^{(L)}, X_3^{(L)}$ times by number (10). Similar to the pervious procedure the first 0.05 of ε vector times by 10 too.

The proposed method is compared with biglasso which the fast and more accurate than others available in literature. Biglasso is parallel computing used memory mapping technique.

Table 1 the results of biglasso and FRMV methods with three simulation scenarios

n	d	Method	Length	CS	FPS	Time
2000	1000	biglasso	9	2	7	3.51
			15	2	13	3.82
		RFMV	2	2	9E12	1.42
			2	2	6E12	1.40
4000	2000	biglasso	16	5	11	12.75
			23	10	13	13.1
		RFMV	5	5	3E16	6.00
			5	5	9E16	6.13
10000	5000	biglasso	71	10	61	59.01
			83	10	73	60.6
		RFMV	10	10	7E32	8.68
			10	10	2E32	8.92

Table 1 shows the results of three scenarios of simulation over 50 big data sets. The average of model length (Length), the average of Correct Selection (CS), the average of False Positive Selection (FPS) and the average of computation time (Time) are computed for both methods. It is clear that the biglasso in the first scenario when $K = 1$, involved (7) false positive selection, therefore, the Length becomes (9) even the informative variables are not correlated with each other or with other variables. When $K = 2$ the FPS increases to (13) and its Time little pit increases too, While the performance of RFMV method is perfect even $K = 1, 2$ no FPS and the time of consuming is less than the time of biglasso. From the results of second scenario when $K = 1$ the biglasso method is faulted in choosing (11) variable should be having zero coefficients, and it could not recognize the correlated variables when $K = 2$, so in addition to (5) correlated variables, there are (13) FPS variables are included to final model. On the other hand, the FRMV method is selected the correct variables without FPS and shorter time than biglasso computation time. Finally, the high performance continue with the third scenario, while biglasso suffers from increasing of dimensionality and multicollinearity problem.

4. Conclusion and Recommendation

The main purpose of this paper is to develop a novel robust and fast variable selection method for linear regression big data where n and d tends to infinity. The proposed method is constructed in two steps involved divide and conquer technique, the marginal correlation based on screening, splitting the variable space divide into (10) blocks and the majority vote is applied. The simulation study has been done and the proposed method compared with biglasso which memory mapping algorithm. From the result that is displayed in table 1, the FRMV method is not only has the ability to select the most significant model, but has scalability to determine the useful covariates among the extremely correlated covariates in the presence of outliers and high multicollinearity problem. Particular attention is paid to the specific case of multicollinearity problem where the correlation between the covariates ranges from 0.95 to 0.999. The originality of our solution lies in two facts; first, the RFMV has the capability to deal with big data on the personal computer and not requires to be within mapping reducer technique, and the results of table 1 reported that RFMV is faster than biglasso. Secondly, our proposed cut-off point helps to reduce the dimensionality to a reasonable dimension. However, no informative variable lies exceed the proposed cut-off point. Consequentially, this paper prefers that is to recommend using the RFMV method for regression analysis of big data.

References

[1] Agostinelli, Claudio, & Salibian-Barrera, Matias. (2010). Robust model selection with lars based on s-estimators Proceedings of COMPSTAT'2010 (pp. 69-78): Springer.

[2] Alfons, Andreas, Baaske, Wolfgang E, Filzmoser, Peter, Mader, Wolfgang, & Wieser, Roland. (2011). "Robust variable selection with application to quality of life research". Statistical Methods & Applications, 20(1), 65-82.

[3] Chen X, Xie MG. "A Split-and-Conquer Approach for Analysis of Extraordinarily Large Data". Statistica Sinica., Vol. 24, No. 4, 2014.

[4] Dupuis, Debbie J.; Victoria-Feser, Maria-Pia., "Robust VIF regression with application to variable selection in large data sets", Ann. Appl. Stat., Vol. 7, No. 1, (2013).

[5] Fan, J., & Lv, J. (2008), "Sure independence screening for ultrahigh dimensional feature space", Journal of the Royal Statistical Society: Series B (Statistical Methodology), 70(5), 849-911. doi:10.1111/j.1467-9868.2008.00674.x

[6] Frank, LLdiko E, & Friedman, Jerome H. (1993). "A statistical view of some chemometrics regression tools", Technometrics, 35(2), 109-135.

[7] Khan, Jafar A, Van Aelst, Stefan, & Zamar, Ruben H. (2007). "Robust linear model selection based on least angle regression", Journal of the American Statistical Association, 102(480), 1289-1299.

[8] Kleiner A, Talwalkar A, Sarkar P, Jordan MI. "A Scalable Bootstrap for Massive Data", Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2014; 76:795–816. MR3248677.

[9] Liang F, Cheng Y, Song Q, Park J, Yang P., "A Resampling-Based Stochastic Approximation Method for Analysis of Large Geostatistical Data", Journal of the American Statistical Association. 2013; 108:325–339. MR3174623.

[10] Lin, D., Foster, D. P. and Ungar, L. H. (2011). "VIF regression: A fast regression algorithm for large data", J. Amer. Statist. Assoc., Vol. 106, No. 493,

[11] Ma P, Sun X. Leveraging for Big Data Regression. WIREs Computational Statistics. 2014; 7:70–76.

[12] Song Q, Liang F., “A Split-and-merge Bayesian Variable Selection Approach for Ultrahigh Dimensional Regression”, Journal of the Royal Statistical Society: Series B (Statistical Methodology). Vol. 77, No. 5, 2014.

[13] Uraibi, H. S. (2019). “Weighted Lasso Subsampling for High Dimensional Regression”, Electronic Journal of Applied Statistical Analysis, 12(1), 69-84.

[14] Uraibi, H. S. (2020). “VIF-Regression Screening Ultrahigh Dimensional Feature Space”, Journal of Modern Applied Statistical Methods, 19(1), eP2916. https://doi.org/10.22237/jmasm/1608553020

[15] Uraibi, H. S., & Midi, H. (2019). “On Robust Bivariate And Multivariate Correlation Coefficient”, Economic Computation & Economic Cybernetics Studies & Research, 53(2).

[16] Uraibi, H. S., Midi, H. (2020), “Robust Variable Selection Method Based On Huberized Lars-Lasso Regression”. Economic Computation & Economic Cybernetics Studies & Research . 2020, Vol. 54 Issue 3, p145-160. 16p.

[17] Uraibi, H. S., Midi, H., & Rana, S. (2015). “Robust stability best subset selection for autocorrelated data based on robust location and dispersion estimator”. Journal of Probability and Statistics, 2015.

[18] Uraibi, H. S., Midi, H., & Rana, S. (2017). “Robust multivariate least angle regression”. J. Science Asia, 43(1), 56-60.

[19] Uraibi, H. S., Midi, H., Talib, B. A., & Yousif, J. H. (2009). “Linear regression model selection based on robust bootstrapping technique”. American Journal of Applied Sciences, 6(6), 1191.

[20] Wang, C., Chen, M. H., Schifano, E., Wu, J., & Yan, J. (2016). “Statistical methods and computing for big data”. Statistics and its Interface, 9(4), 399–414. doi:10.4310/SII.2016.v9.n4.

مجلة كلية الرافدين الجامعة للعلوم (2023)؛ العدد 54؛ 490 - 497

PISSN: (1681-6870); EISSN: (2790-2293)
مجلة كلية الرافدين الجامعة للعلوم

Available online at: https://www.jrucs.iq

JRUCS
Journal of AL-Rafidain University College for Sciences

AL- Rafidain University College

خوارزمية تصويت الأغلبية السريعة والقوية لنموذج انحدار البيانات الضخمة

د. حسن سامي عريبي
hassan.uraibi@qu.edu.iq
قسم الإحصاء - كلية الإدارة والاقتصاد - جامعة القادسية، القادسية، العراق

المستخلص

نهج التصويت بالأغلبية هو أحد تقنيات فرق تسد التي لديها القدرة على تحليل البيانات الضخمة وفهمها ومن ثم اتخاذ القرار بشأنها. تعتبر بيانات الانحدار الكبيرة ضخمة ليس فقط من حيث الحجم حيث تميل P و n إلى اللانهاية، ولكن من حيث الكثافة والتعقيد أيضًا. على سبيل المثال، يؤدي جمع البيانات الضخمة من مجموعات سكانية فرعية مختلفة إلى مشكلة عدم التجانس التي تؤدي بالتأكيد إلى وجود القيم المتطرفة. علاوة على ذلك، فإن معالجة مثل هذا الحجم من البيانات تتجاوز قدرة الأدوات التحليلية القياسية. لذلك، تتطلب المعالجة أن تتمتع الخوارزمية المقترحة بخاصيتين رئيسيتين، سريعة ودقيقة. ولسوء الحظ، فإن نهج التصويت بالأغلبية لا يمكن الاعتماد عليه عندما تكون القيم المتطرفة موجودة في البيانات. علاوة على ذلك، فإن لعنة الأبعاد تسبب مشكلة الخطية المتعددة التي تؤدي إلى نتائج مضللة. تقترح هذه الورقة أسلوب تصويت سريع وقوي للأغلبية وهو نسخة جديدة من النسخة الأصلية مع بعض الخطوات. الخطوة الأولى هي تقسيم مصفوفة التصميم عموديًا إلى عدد من الكتل للتغلب على لعنة الأبعاد الفائقة. وينبغي النظر في التصويت على اختيار أفضل مجموعة فرعية من المتغيرات باستخدام طريقة اختيار متغير قوية كإجراء تخفيض الأبعاد ومقاومة لوجود القيم المتطرفة. والخطوة الثانية هي تجميع أفضل المجموعات الفرعية في نموذج انحدار خطي واحد، ثم استخدام خوارزمية تصويت الأغلبية للحصول على أفضل المتغيرات (على سبيل المثال. تم إجراء دراسة محاكاة في هذا البحث لمعرفة أداء التقنية المقترحة مقارنة مع تقنية Big-Laso المعروفة بأنها أسرع طريقة في الأدبيات الإحصائية في الوقت الحالي. تظهر النتيجة تفوق الخوارزمية المقترحة والتي هي أسرع من Big-lasso وأكثر دقة منها.

معلومات البحث

تواريخ البحث:
تاريخ تقديم البحث: 2022/12/20
تاريخ قبول البحث: 2023/3/3
تاريخ رفع البحث على الموقع: 2023/12/31

الكلمات المفتاحية:
البيانات الضخمة، فرق تسد، اختيار متغير قوي، تصويت الأغلبية

للمراسلة:
د. حسن سامي عريبي
hassan.uraibi@qu.edu.iq

https://doi.org/10.55562/jrucs.v54i1.616