Detection of Outliers in High Dimensional Data with Lasso Regression

Ch. Anuradha, M. Ramesh

Abstract: ***Detecting Outliers has become a significant research area in data mining in last few years. The focus of this research has been to identify patterns or objects in huge data sets of a database that are exceptional from normal pattern, specifically dissimilar, and unpredictable with reference to the most of the datasets. As billions of personal computers, and internet users rose phenomenally, huge data sets of real life applications have been created for new challenges as well as explorations in research for Outlier detection. Many traditional techniques to detect outliers have unable to yield good results in such environments. So, developing a method to detect Outliers has become a critical task. A method to identify anomalies in high dimensional data based on Lasso Regression has been study in this research. This framework has been implemented in the open source JMP software. The parameters such as RSquare 0.001162, RMSE 0.031806 and Mean Response 0.007889 are calculated using Spambase dataset. The results from the experiments have shown that the proposed method detects Outliers in high dimensional data with potentially higher accuracy.***

Keywords: ***Outlier detection, Lasso regression, High-dimensional data, JMP, Spambse Dataset.***

I. INTRODUCTION

The occurrence of an Outlier has small chance probability in a given data set but it has to be detected. As Data Science evolved with analysis of massive data, Outliers are data pixels which are significantly different from other data pixels of similarity. An outlier may occur as an experimental or simulation error or measurement error. Detecting an outlier has become a major research in various implementations using data mining techniques. In huge data examination, the no of mutable both independent and dependent observed or sampled or simulated have been significantly large. One has to perform a rational examination which is the recognition of annotations gathered. Though anomalies have specific or key information that can be ignored just like an error or noise, sometime this is important. Aberrant data has been identified through detected outliers and this will create misspecification model adversely, incorrect results and bias in parameter estimation. So, the key is to detect Outliers before the process of modeling and analysis. A challenging area to detect Outlier is the case of high-dimensional data. Many dimensions contain noisy data, and detecting anomalies is irrelevant in high-dimensional data. This may raise the tendency for pairwise distances to become more alike. One has to keep in mind that Accuracy of distance computations can have a dilution effect with irrelevant attributes and this may lead to inaccurate scores for the resulting outliers.

Detecting Outliers has become a key process in data sets and subsequently any data set must exclude outliers for significant use. Methods of varies types to detect Outlier have been in use for research, academic and commercial solicitations, like credit card scam recognition, prevention of heart attack, detecting breast cancer, many difficult clinical diseases require suit, electing abnormality investigation, data cleansing, data hackers, severe climate of cyclones or typhoon predictions, earth quake monitoring, volcanic eruptions, ravaging tornadoes, performance analysis of athletes in many sporting events, and other data-mining tasks. There are four approaches to computer-based methods for outlier detection: Density-based approaches, Distance-based approaches, statistical-based approaches and deviation-based approaches. In this research, we presented an outliers detection method in high-dimensional data streams based on Lasso Regression.

This research has been presented introduction in 1st section, description of the different methods and approaches on detecting Outliers in 2nd section. Section 3 discusses about outlier detection technique. Section 4 provides a compact survey of existing outlier detection techniques using Lasso regression model. Finally the last section 5 gives conclusions of this research.

II. PRVIOUS RESEARCH

Detecting Outliers has been the key in finding out data points or observations that have been deviated substantially from rest of the data point in any data set. A review of previous research of methods some may be state of the art, to detect outliers has been presented here. Also offer some idea of the settings in which detecting Outliers might work well. Hartigan & Wong (1979) presented an algorithm based on clustering with p-means. This procedure distributes datasets into P groups with M points in N dimensions. The key is to minimize the sum of squares within each cluster. It was not realistic to obtain solution with sum of squares being minimum against all partitions, unless when M points, and N dimensions were small with two clusters.

Revised Manuscript Received on October 30, 2019.
* Correspondence Author

Ch. Anuradha*, Research Scholar, Dept. of CSE, ANU, Guntur, Andhra Pradesh, India.
Dr. M. Ramesh, Research Supervisor, Dept. of CSE, ANU, Guntur, Andhra Pradesh, India.

Retrieval Number: A1478109119/2019©BEIESP
DOI: 10.35940/ijeat.A1478.109119
Journal Website: www.ijeat.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

The authors presented "local" optima and solutions based reduction of sum of squares within-cluster when no movement of a point from one to another cluster. Ha et al., (2015) presented unsupervised technique using distance to detect Outliers. It uses iterative random sampling.

This techniques takes inspiration from the simple idea that outliers have not been selected easily as Inliers in unsighted arbitrary sampling. Therefore the objects selected are provided with more inlierness scores. A new factor called Observability has been developed making use of this idea. Moreover entropy of scores is presented in turn to provide heuristic guideline to get the optimum size of the closest neighbourhood. But performance of this technique decreases for the biggest entropy values. But overall the outliers are found in an effective manner and could be used with the combination of other techniques for getting better results. Methods based on Density have been discussed to take the local density as criteria to search for Outliers. These methods define outliers based on the local structure of the dataset. Zhong& Huang (2012) have proposed detecting Outlierr based on density with lofty computation and smaller accuracy, density (VDD) measure and distance changing.

Aggarwal and Singh (2013) designed a K-Mean and hybrid distance technique for detecting outlier in Multi-Dimensional Data Set. In the designed system, outliers are detected by partitioning the dataset with the clustering method.teh clustering is performed by using the K – Mean method. After clustering the dataset outliers are to found from the each cluster by using the mean of Euclidean and Manhattan technique. There after calculate the mean of all the outliers that are found from each clusters and by comparing the each outlier with the mean of the previous outliers, real outlier will be separated and that are the real outliers that are different from the other cluster data. Abraham & Chuang (1989) had proposed statistic measures based on regression analysis to detect Outliers in time series. They consider statistics had been Approximations and asymptotic distributions. To distinguish an innovational outlier from one an observational a method had been identified. To detect Outliers presence in time series model, a four step procedure had been defined. The methodology had been demonstrated with an example.

It is clear from the literature survey, detection rate and detection accuracy is needed to be improved.

III. LASSO-BASED OUTLIER DETECTION

Least Absolute Shrinkage and Selection Operator (LASSO) method of statistical analysis has been presented in this research. This would to tackle the key issue as part of with high dimensional data in outlier’s detection. In particular, this research, in specific, has been used LASSO for linear regression models with high dimensional data. The M-estimator which had the Bayesian interpretation of a linear model with Laplacian prior:

$$\hat{\beta} = \arg \underset{\beta}{min} \; \|X - X\beta\|_2^2 + \lambda\|\beta\|_1 \; \ldots\ldots\ldots\ldots\ldots\ldots (1)$$

have multiple names: Lasso Regression and L1-penalized Regression. The Lasso Regression Estimate has a key elucidation in the bias-variance situation. For simplicity, consider the special point where X′X = Ip. In this point, the aim of the Lasso regression decouples:

$$\|X - Y\beta\|_2^2 + \lambda\|\beta\|_1 = X'X + \beta'Y'Y\beta - 2X'Y\beta + \lambda\|\beta\|_1$$
$$= X'X + Pp\ j{=}1\ \beta\ 2j - 2Y'Yj\beta j + \lambda|\beta j \ldots\ldots\ldots (2)$$

Where Xj is the data representing columns of matrix X, specifically jth column. And because it decouples we can solve the optimization problem separately for each term in the summation. In the first point, βj > 0 and so the derivative has been set to zero, gives

$$2\hat{\beta} j - 2Y'Xj + \lambda = 0 \Rightarrow \hat{\beta} j = Y'Xj - \lambda/2 \ldots\ldots\ldots\ldots\ldots (3)$$

In the second point, βj < 0 and so derivative has been set to zero, gives

$$2\hat{\beta} j - 2Y'Xj - \lambda = 0 \Rightarrow \hat{\beta} j = Y'Xj + \lambda/2 \ldots\ldots\ldots\ldots\ldots (4)$$

In the third point, βj = 0.

One important characteristic of Lasso Regression is consistency in the setting with high dimensional data. Assume that Xj is column-normalized, meaning that

$$Xj / \sqrt{}\, n \le 1, \forall j = 1, \ldots, p \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (5)$$

We have two results regarding sparse models.

If some technical conditions hold for the s-sparse model, then with probability at least 1 − c1 exp(−c2 log p) we have for the s-sparse model that

$$\|\hat{\beta} - \beta\|_2 \le c3 \sqrt{}\, s\sqrt{\log p / n} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (6)$$

Where c1, c2, c3 are positive constants.

If some technical conditions hold for the approximately-sq-sparse model (recall that q ∈ [0, 1]) and β belongs to a ball of radius sq such that √sq(log p /n) 1/2−q/4 ≤ 1, then with probability at least 1 − c1 exp(−c2 log p) we have for the approximately-sq-sparse model that

$$\|\hat{\beta} - \beta\|_2 \le c3 \sqrt{}\, sq\ (\log p / n)\ 1/2{-}q/4 \ldots\ldots\ldots\ldots\ldots\ldots (7)$$

Where c1, c2, c3 are positive constants. Compare this to the classical (fixed p) setting in which the convergence rate is Op ($\sqrt{p}$ /n).

In the case of high dimensional data, the Ordinary Least Squares Estimator has not been useful. Assumptions have to be identified to solve problems in an optimum way. Sparsity has been assumed such that true model has been sparse. A Sparse solution has been sought with zero values for many components of the vector β̂. Sparse solutions have been provided by The LASSO. Some measurements have been precisely fixed to 0 by the LASSO based on the amount of regularization. So, Variable selection and estimation of data sets has been performed by the LASSO. A secure practise resolution could not be identified by the LASSO as mentioned earlier and using Convex Optimization Form it can be resolved efficiently. Many iterative procedures can be used to compute resolution of the LASSO for high-dimensional locale. Coordinate Descent algorithm has been used for the LASSO regression in this research. We consider simulation steps as follows:

Produce an error vector as En×1 ∼ Nn(0, In) and then calculate a retort trajectory as Y = Xβ0 + E.

Produce alternative fault trajectory as En×1 ∼ Nn(0, In) and then calculate a novel retort trajectory as Ytest= Xβ0 + E.

Calculate the OLS fit Xβ̂OLS = X (XTX) −1XTY.

Calculate the likelihood fault for OLS, as PE (OLS) = 1/ n $\|Ytest - X\hat{\beta}OLS\|_2^2$.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Describe a lattice of 50 standards of λ correspondingly spaced among 0 and 1, as λseq = [0, 0.02, ..., 1].

Calculate the LASSO calculator β^ (λ) for every λ ∈ λseq, and then calculate the LASSO fit for every LASSO calculator as Xβ^ (λ).

Calculate the prediction error for the LASSO for each λ ∈ λseq, as

PE (LAS S O, λ) =1/ n $\|Ytest - X\hat{\beta}OLS\|_2^2$.

Calculate the no of accurate 0 measurements for each LASSO calculator β^ (λ) as,

NZC (β^ (λ)) = $\sum_{i=1}^{p} 1(\beta\hat{}(\lambda))j$ = 0), where 1 is a pointer purpose.

A real world problems with high dimensional has been considered and the LASSO method has been applied for estimation and variable selection.

IV. EXPERIMENTAL SETUP

The Proposed method has been verified for the performance using Spambase dataset acquired from UCI Machine Learning Repository. There are 4601 observations and 57 attributes in this dataset, which characterize the contents of an email. The first column in given dataset has been binary variable representing spam email. Real valued explanatory variables represented the rest of the 57 columns. In these columns, 48 columns of data that contain word frequency, 6 columns of contain character frequency, and 3 columns of data contain capital letters sequence. The jackknife distance, as shown in figure 1, has been computed with estimates of the mean, standard deviation, and correlation matrix without the observation data. With an Outlier, the distances jack-knifed have been practical use. The Mahalanobis distance has been twisted in this case and covers Outliers. Also this creates other data points seem to be outlying than they supposed to be. For each value, the distance of jackknife has been computed as follows:

$$J_i = \sqrt{\frac{(n-2)n^2}{(n-1)^3} \times \frac{M_i^{\,2}}{1 - \frac{nM_i^{\,2}}{(n-1)^2}}}$$

Where:

n = observations number
p = variables number as columns
Mi = Mahalanobis distance for the ith observation

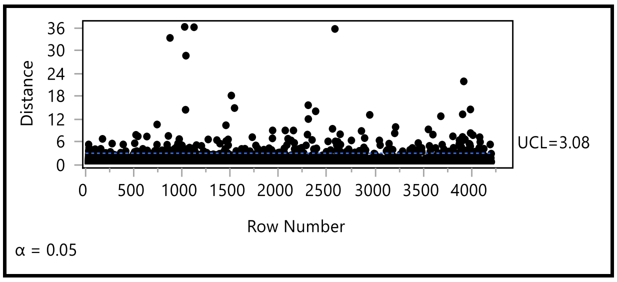

Fig. 1. Estimation of jack-knifed distances using spambase for outlier detection

The platform for Scatterplot 3D displays three-dimensional view of associated data table contains the values of numeric columns and also it is rotatable. Three variables can be displayed at one time from the data set. The 3D scatterplot can illustrate a biplot representation of the points and variables of higher dimensions picturing variation. The 3D scatterplot represents outstanding directions of data as shown in figure 2:

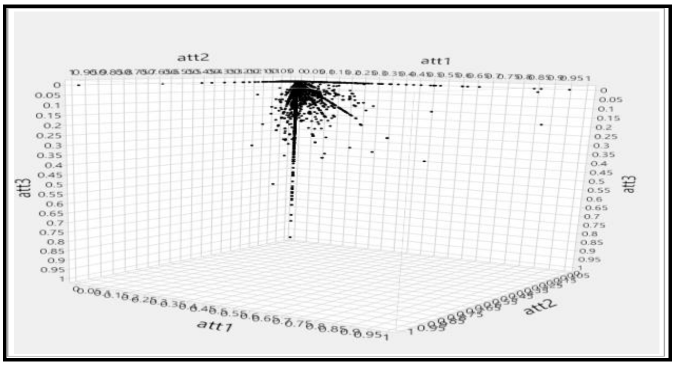

Fig.2. 3D scatter plot for LOF with outlier points

Correlation has been a measure of two variables linear association and the correlations are estimated by Row-wise method for att1, att2, att3 and att4 as shown in table 1:

Table-I: Estimation of correlations by Row-wise method for attribute4, attribute 3, attribute 2 and attribute 1

Name	Attribute1	Attribute 2	Attribute 3	Attribute 4
Attribute 1	1.0000	0.0341	0.0628	0.0054
Attribute 2	0.0341	1.0000	0.0279	-0.0089
Attribute 3	0.0628	0.0279	1.0000	-0.0199
Attribute 4	0.0054	-0.0089	-0.0199	1.0000

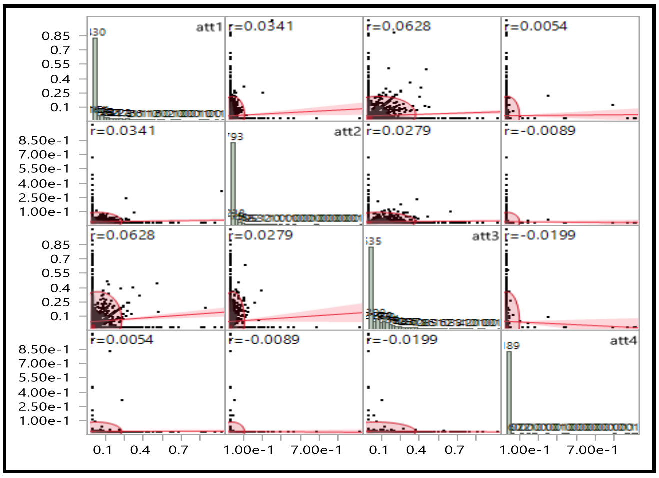

Fig. 3. Scatterplot Matric using spambase for outlier detection

Retrieval Number: A1478109119/2019©BEIESP
DOI: 10.35940/ijeat.A1478.109119
Journal Website: www.ijeat.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

A scatterplot matrix has been a collection of scatterplots organized into a grid (or matrix). The relationship between two variables of att1, att2, att3 and att4 has been shown as a scatterplot, as shown in figure 3:

The following table details of estimated parameters such Rsquare, RMSE, Mean Response etc. as shown in table 2:

Table-II: Various Parameters of Fit

Parameter	Value
Mean Response	0.007889
RSquare	0.001162
Observations	4207
Root Mean Square Error	0.031806
RSquare Adj	0.000925

The report of Analysis of Discrepancy has been provided the calculations to compare the fitted model to all predicted values as a prototype where the standards equal to the riposte mean as shown in table 3:

Table-III: Analysis of Discrepancy

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	0.0049490	0.004949	4.8921
Error	4205	4.2539077	0.001012	Prob > F
C. Total	4206	4.2588567	0.0270*	

Residual by Predicted Design illustrates the residuals designed against the projected standards of Y and residual standards distributed arbitrarily around zero as shown in figure 4:

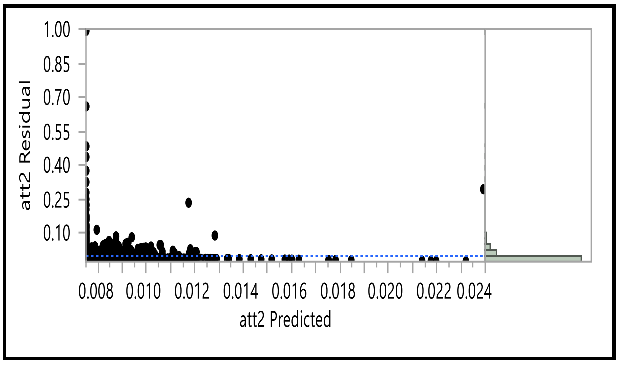

Fig. 4. Residual by Predicted Design

The Actual by Projected design seems by defaulting. It delivers a graphical evaluation of model fit has been done by this and also shows variation because of random effects. The plots shows the marginal predicted values of Y with the observed values of Y, shown in figure 5:

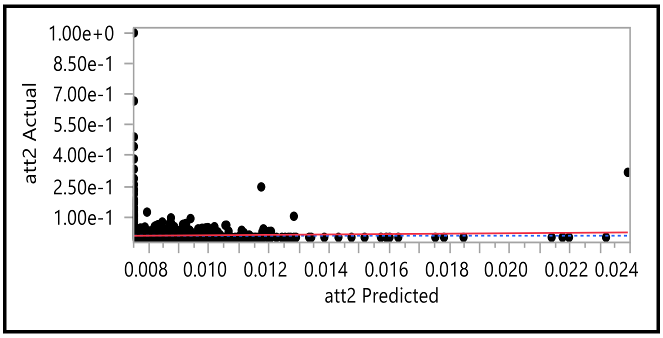

Fig. 5. Actual by Predicted Plot

Residuals have been plotted against row numbers as Residual by Row Design. This can detect the decorations that outcome from the row gathering of the interpretations as shown in figure 6:

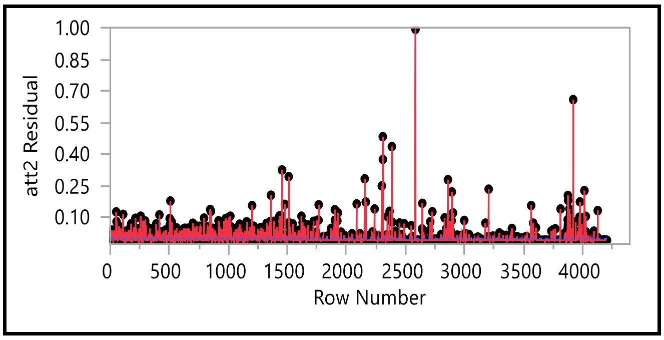

Fig. 6. Residual by Row Plot

The implementation of Lasso regression is done using spambase dataset. The parameters such as RSquare 0.001162, RMSE 0.031806 and Mean Response 0.007889 are calculated using Spambase dataset. The results from experiment show that the proposed method detects Outliers in high dimensional data with potentially higher accuracy. Several methods are used for detection of outliers from a particular dataset. Each method aims to identify the outliers and gives the best result to other. We compare these methods according their advantages and disadvantages to find best method.

Techniques to detect Outliers can be compared with different factors and dataset. Straight comparison is not possible because every method workings on its distinct collection of constraints. In this research, the comparison is made with different parameters such efficiency, time complexity, high dimensional data, complexity analysis etc. Calculations of these parameters indicate that a method to detect Outlier based on Lasso Regression was the most capable out of the studied approaches. In this comparisons, fiver algorithms have been used i.e Density, distance, deviation, clustering, and regression-based methods. It has been compared based on proposed dataset. From result simulation it has been found out that best algorithm is ridge regression with optimal time complexity.

Table-II: Comparison of various outliers' detection approaches

Factors	Outliers Detection Approaches				
	Density-based Method	Distance-Based Method	Clustering-based Method	Deviation-based Method	Regression-based Method
Efficiency	Efficient	Efficient	Very efficient	Efficient	High Efficient
Computation Time	High	Low	Low	High	High
Complexity Analysis	Extremely difficult	Moderately difficult	Less difficult	Moderately difficult	Less difficult
High Dimensional Data	Valid	Valid	Valid	Valid	Valid
Time Complexity	O (n log n)	O (KN2)	O (n2 log n)	O (n2)	Optimal time complexity

V. CONCLUSION

Detecting Outlier has been research significantly in the data mining area. Finding data objects that have been radically unlike from the respite of dataset is the process. Huge databases from real life applications have been surfacing rapidly in the past five years. Detecting Outliers has been facing greater challenges as well as opportunities for research. Many traditional techniques to detect Outlier have not been working well. So, the need for creating an up to date method to detect Outliers becomes critically vital. This research aims at method to identify anomalies in high-dimensional data based on Lasso Regression. This framework is implemented in the open source JMP software. The parameters such as RSquare 0.001162, RMSE 0.031806 and Mean Response 0.007889 are calculated using Spambase dataset. The results from the experiment show that the method proposed detects Outliers in high dimensional data with potentially higher accuracy.

ACKNOWLEDGMENT

There anonymous reviewers of this research paper have been deeply appreciated for their insights, thoughtful suggestions, and constructive comments, as part of the review process of draft.

REFERENCES

1. Jiang W., Shah S., Liu H. (2004),"On-line outlier detection and data cleaning," Computers and Chemical Engineering, 28, 1635–1647.
2. B A Turlach, B Presnell, and M R Osborne (2000), "A New Approach to Variable Selection in Least Squares Problems", IMA Journal of Numerical Analysis, 20(3):389403.
3. Alan J Lee F Seber and George A (2003), "Linear Regression Analysis", Wiley, 2003.
4. Jiawei Han Micheline Kamber, JianPei Data-Mining Concepts and Techniques-3rd-Edition-Morgan-Kaufmann-2011.
5. Cabrera, J. B. D., Lewis, L., and Mehra, R. K. (2001), "Detection and classification of intrusions and faults using sequences of system calls", SIGMOD Records 30, 4, 25 - 34.
6. M. Lichman and K. Bache, UCI machine learning repository, 2013.
7. C. Pizzuti and F. Angiulli (2002), "Fast outlier detection in high dimensional spaces", In Proceedings of the 6th European Conference on Principles of Data Mining and Knowledge Discoverys (PKDD), Helsinki, Finland, pages 15– 26.
8. V. Kumar, A. Banerjee, and V. Chandola (2012), "Anomaly detection for discrete sequences: A survey", IEEE Transactions on Knowledge and Data Engineering, 24(5):823– 839.
9. T. Newsham, T. Ptacek "Insertion, Evasion, and Denial of Service: Eluding Network Intrusion Detection", Secure Networks Inc, 1998.
10. Bishop, C (1994), "Novelty detection and neural network validation", In Proceedings of IEEE Vision, Image and Signal Processing. Vol. 141. 217 - 222.