Journal of Engineering and Applied Sciences 14 (8): 2465-2469, 2019
ISSN: 1816-949X

A Statistical Method for Big Data with Excessive Zero-Inflated Problem

Sunghae Jun
Department of Big Data and Statistics, Cheongju University, Cheongju, Korea

Abstract: In many cases, we meet the zero-inflated problem in big data analysis. This is because the value of zero is too much in the data table structured through preprocessing from collected big data. If the big data is analyzed as it is the performances of estimation and prediction of statistical models will deteriorate. To build valid models for big data analysis, we have to solve the zero-inflated problem of big data. So, we propose a statistical modeling to overcome the zero-inflated problem in big data analysis. In this study, we combine the method of data division with count data models such as Poisson, hurdle, negative binomial regressions. In order to verify the validity of the proposed approach, we carry out case study using simulated and patent big data.

Key words: Statistical model, big data, zero-inflated problem, count data analysis, patent big data analysis, validity, statistical modeling, data analysis

INTRODUCTION

In the field of statistical data analysis, the proportion of big data is getting bigger than that of small data. This is due to the remarkable development of Information and Communication Technology (ICT). Big data is huge data that cannot be processed by a single computer or a single software due to the size and heterogeneity of the data (Berman and Morgan, 2013; Ryu and Jun, 2016). In addition, the big data has been applied to diverse fields such as digital security, social science, intellectual property, etc., (Bhujade and Chandak, 2018; Hashmi and Ahmad, 2017; Ryu and Jun, 2016). To analyze big data, we use statistics or machine learning algorithms (Micheline *et al*., 2012). Traditionally, statistics have focused on samples (small data) extracted from populations (Lu and Li, 2013; Ross, 2012). The main goal of statistics is to estimate the population parameters by analyzing sample data representing population. But with the rapid development of ICT, the era of big data has arrived. In some cases, now we can get very large data close to the population. This means that we do not need to analyze sample data for parameter estimation. Therefore, we can understand the population easily in big data environment. However, many problems have occurred in the analysis of big data (Jun *et al*., 2014; Ryu and Jun, 2016). One of them is zero-inflated problem with excessive zeros in big data (Kim and Jun, 2016). This problem occurs when the collected big data is preprocessed into structured data (Kim and Jun, 2016). For example, in text document big data analysis, we make document-keyword matrix for statistical analysis and machine learning (Jun *et al*., 2012). This matrix consists of document and keyword as row and column, respectively and its element represents the occurred frequency value in each document. At this time, most frequency values are zeros, so, we need to statistical modeling to overcome this problem. In general statistics, zero-truncated and zero-inflated models are used for solving excessive zeros in big data (Hilbe, 2011). In addition, statistics provides some probability distribution to make the models to analyze excessive zero data (Hilbe, 2011). The probability distributions are Poisson, negative binomial, hurdle, Poisson inverse Gaussian (Hilbe, 2014). But the traditional statistics for excessive zero data have the limitation of zero ratio in data. They cannot analyze the zero-inflated data when the zero ratio is more than 50%. In big data analysis, the extremely excessive zeros often occur. So, we need new method to analyze the big data with zero values more than 50%. In this study, we propose a method to settle this problem. First, we divide the entire data with excessive zeros into analytic data according to the ratio of zeros, second, we apply the zero-inflated models or Generalized Linear Model (GLM) to analyze them. To illustrate how the proposed method could be applied to real domain, we make two experiments using simulation data and patent big data. From the experimental results, we show the performance validity of our proposed method.

MATERIALS AND METHODS

Statistical modeling for zero-inflated big data: The zero-inflated problem is common in big data analysis. In general, we collect big data from diverse sources such as Social Network Services (SNS), patent documents or web

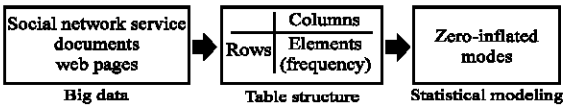

Fig. 1: Proposed process of statistical modeling to overcome zero-inflated problem in big data analysis

sites. The retrieved big data is not suitable to statistical methods or machine learning algorithms because the analytical methods based on statistics and machine learning are carried out on the table structure data with rows and columns. Figure 1 shows the proposed process of statistical modeling to solve the zero-inflated problem in big data analysis.

Heterogeneity and volume are important characteristics of big data. That is big data contains diverse data types such as number, text, picture or sound. In addition, the data size is extremely large. For the texts or numbers extracted from the SNS are enormous. So for analyzing the big data, we must transform the big data into structured data such as table using text mining techniques. In this study, we use the R data language and its text mining package 'tm' (Feinerer *et al*., 2008; Feinerer and Hornik, 2018; R Core Team, 2018). A table has the form of a matrix of rows and columns. The rows and columns are documents and keywords, respectively. The value of the matrix indicates the frequency at which a particular keyword appeared in the document.

There is the zero-inflated problem in this matrix. This is because it is not uncommon for certain keywords to appear in the entire document. In general, the zero-inflated problem means that the matrix has excessively many zero values. For example, if the Poisson Model with parameter m = 5 and we extract 1,000 numbers randomly from the distribution, we can expect about 7 in the number of zeros from the 1,000 Poisson random numbers in Eq. 1:

$$1000\times\frac{e^{-5}5^{0}}{0!}=6.737947 \quad (1)$$

Using the above method, we can check whether the data set has zero-inflated problem or not. General count models such as Poisson and negative binomial regressions are not suitable to zero-inflated count data. General count models such as Poisson and negative binomial regressions are not suitable to zero-inflated count data (Hilbe, 2014). To solve the problem of zeros in data, the zero-truncated and zero-inflated models were proposed. In the zero-truncated Poisson Model, the Poisson Probability Mass Function (PMF) is in (Hilbe, 2014) Eq. 2:

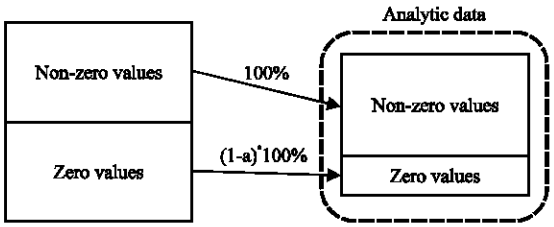

Fig. 2: Division of data with zero values for big data analysis

$$1\rightarrow 1\text{-}P(Y=0) \text{ or } \exp(-\mu)\rightarrow 1\text{-}\exp(-\mu) \quad (2)$$

Where:

Y = Poisson response random variable

μ = The parameter of Poisson distribution (mean value)

In addition, we can use the hurdle and zero-inflated Poisson models to solve the huge zeros in big data. If more zeros are actually observed than the expected number of zeros, we use the zero-inflated Poisson or negative binomial models (Hilbe, 2011). In this case, we have to consider the observed values and excessive zeros at the same time. The basic concept of zero-inflated model is to divide the count data into two parts as follow (Hilbe, 2014) Eq. 3:

$$P(Y=y)=\begin{cases} p+(1\text{-}p)f(0) & ,y=0 \\ (1\text{-}p)f(y) & ,y>0 \end{cases} \quad (3)$$

where, the probability p is defined Eq. 4:

$$Y=\begin{cases} 0 & \text{with } p \\ f(Y=y) & \text{with } (1\text{-}p) \end{cases} \quad (4)$$

That is the whole data is analyzed by two approaches which are y = 0 and y>0. Though, we consider the traditional zero-inflated models for the big data with excessive zeros, we have to overcome the limitation of the models. The meaning of 'excessive' is that the ratio of zeros does not exceed 50% in the data. But in big data analysis, we often encounter data that contains zero or more than 50%. So, we have difficulties in solving this problem by the traditional zero-inflated models. In this study, we propose an analytical strategy that adjusts the ratio of zero values in big data to settle the problem of extremely excessive zeros. Figure 2 shows the proposed data division for the big data with extremely excessive zero values.

We divide the data with extremely excessive zero values into data sets with non-zero and zero values. We use all data elements from the data with non-zero values and extract (1-a)*100% sample elements randomly from the data with zero-values. The ratio of zeros is manipulated by adjusting the value of (1-a) ($0 \le a \le 1$). We can consider a linear regression model in Eq. 5:

$$Y = \beta_0 + \beta_1 x + \varepsilon \quad (5)$$

where, β_0 and β_1 are regression parameters and ε is error term distributed to normal $N(0, \sigma^2)$. When the response variable Y contains excessive zero values, we divide the data set $(x_1, y_1), ..., (x_n, y_n)$ into analytic data by (1-a)*100% in Fig. 2. Using adjusting (1-a)*100%, we settle the extremely excessive zeros in big data analysis. In the next section, we illustrate the validity of our research using simulation and real domain data sets. Next, in our experiments, we consider the Akaike Information Criterion (AIC) to evaluate the performance comparison. The AIC is defined as follow (Murphy, 2012) Eq. 6:

$$AIC = -2\{\log L + K\} \quad (6)$$

Where:

logL = Maximum log-Likelihood

K = The number of estimated parameters

The smaller the AIC value is the better the performance of fitted model is.

RESULTS AND DISCUSSION

Experimental results: We used two data sets to show the performance and validity of the proposed method. First, we carried out an experiment using the simulation data generated from a linear regression model with extremely excessive zero values. Next, we collected patent documents related to Artificial Intelligence (AI) technology for performing another experiment.

Simulation data: We considered the Poisson regression model for simulation data set Eq. 7:

$$\lambda = \exp(2 + 5x_1 - 5x_2) \quad (7)$$

where, 2, 5 and -5 are b_0-b_2 for the estimated parameters of β_0-β_2, respectively. The $\lambda(>0)$ is the parameter of Poisson distribution and mean value of this distribution. In this experiment, a random variable Y is followed to the Poisson probability distribution (Akritas, 2016) Eq. 8:

Table 1: Frequency distribution of response variable Y

Count data	Frequency
0	87299
1	9237
2	1093
3	642
4	513
5	419
6	328
7	190
8	124
9	84
10	47
11	15
12	5
13	1
14	1
15	2

Table 2: Performance comparison according to zero ratio in simulation data sets

Zero ratio (%)	b_0	b_1	b_2
100	2.0313	5.1627	-5.1630
90	2.0288	5.1497	-5.1499
80	2.0261	5.1361	-5.1363
70	2.0234	5.1220	-5.1222
60	2.0205	5.1073	-5.1075
50	2.0176	5.0920	-5.0921
40	2.0144	5.0759	-5.0760
30	2.0111	5.0589	-5.0590
20	2.0076	5.0411	-5.0412
10	2.0039	5.0223	-5.0223
0	1.9999	5.0022	-5.0022

$$P(Y = y) = e^{-\lambda}\frac{\lambda^y}{y!}, \quad y = 0, 1, 2, ..., \quad (8)$$

Table 1 shows the value distribution of response variable Y in simulated data. The simulation data contains 100,000 values with 87,299 zeros. That is the percentage of zeros is 87.3%. This data set is popular zero-inflated data. In this experiment, we divided the whole data set into analytic data according to (1-a)*100% of Fig. 2. Table 2 shows the experimental results from 100 to 0% of included zero values in the analytic data set.

We found the regression parameters of b_0-b_2 are all similar to 2, 5 and -5, respectively. This means that the changes of ratios of zeros values doesn't affect parameter estimation of count regression models. Figure 3 shows the AIC values according to zero ratios.

Although, there is a slight difference, the AIC values representing the performance of the fitted and predicted model according to the ratio of zero value are almost similar. Therefore, the validity of our proposed method can be confirmed through this simulation data results.

Patent big data: In order to verify the practical performance of the proposed method, we conducted

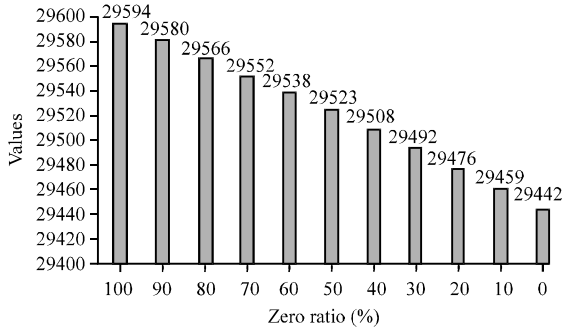

Fig. 3: AIC comparison of zero ratios; simulation data

Table 3: Top four IPC codes of AI patent big data

IPC code	Code description
G06F	Electric digital data processing
G06K	Recognition of data, presentation of data, record carriers, handling record carriers
H04N	Pictorial communication, television
G10L	Speech analysis or synthesis, speech recognition, speech or voice processing, speech or audio coding or decoding

experiment using patent big data. We retrieved the patent documents related to AI from the patent databases in the world (USPTO., 2018; Anonymous, 2018). Patent document contains diverse information about developed technology such as title, abstract, inventor's names, citations, applied and registered dates, claims, drawings and figures, International Patent Classification (IPC) codes, etc., (Hunt *et al*., 2007; Porter *et al*., 1991). We extracted the IPC codes from the collected patent big data and used top four codes (Feinerer *et al*., 2008; Feinerer and Hornik, 2018). Table 3 shows the four IPC codes and their technological description (Anonymous, 2018).

For example, the IPC code of G06F represents the technology related to electric digital data processing. We made a linear regression model for our second experiment in Eq. 9:

$$\lambda=\exp\left(b_0+b_1G06K+b_2H04N+b_3G10L\right) \quad (9)$$

where, λ the parameter of Poisson probability distribution of random variable G06F. We determined G06F as response variable because this code is first ranked code in frequency. Table 4 shows the value distribution of response variable G06F.

Of the total 13,858 data elements, the frequency of zero is 9,742, representing 70.3%. We also divided this data into analytic data sets by the zero ratios from 100 to 0%. Table 5 represents the performance comparison between the analytic data sets.

We have confirmed that there is a performance difference between 100 and 0%. In other words, although, there is a performance difference between the case where

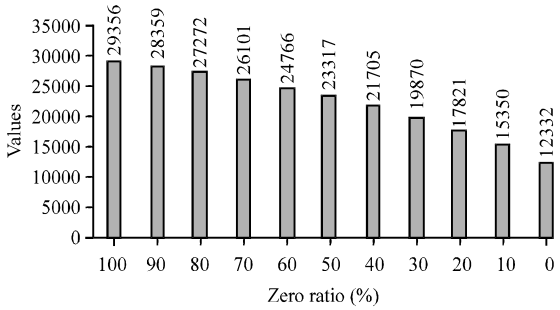

Fig. 4: AIC comparison of zero ratios: patent big data

Table 4: Frequency distribution of response variable G06F

Count data	Frequency
0	9742
1	2134
2	1137
3	507
4	202
5	30
6	12
7	9
8	4
9	3
10	78
>10	0

Table 5: Performance comparison according to zero ratio in patent big data sets

Zero ratio (%)	b_0	b_1	b_2	b_3
100	-0.32	-0.54	-0.20	-0.17
90	-0.26	-0.52	-0.19	-0.16
80	-0.20	-0.50	-0.18	-0.15
70	-0.13	-0.48	-0.17	-0.14
60	-0.05	-0.46	-0.16	-0.13
50	0.04	-0.43	-0.14	-0.12
40	0.13	-0.40	-0.13	-0.10
30	0.24	-0.37	-0.11	-0.09
20	0.35	-0.30	-0.08	-0.08
10	0.48	-0.23	-0.05	-0.05
0	0.64	-0.11	-0.02	-0.02

all zero values are used and the case where all the values are excluded it is confirmed that there is no significant difference up to 70%. Therefore, we could confirm the validity of our proposed study. Figure 4 shows the AIC values of ratio comparisons by zero values.

As shown in Table 5 it can be seen that the AIC value is not significantly different between 100 and 70%. From the experimental results of simulation data and patent big data, the performance validity of the proposed method is confirmed.

CONCLUSION

In this study, we proposed data segmenting method to overcome the zero-inflated problem in big data analysis. We have met the extremely excessive zeros in structured big data. Of course, there were many methods to analyse

the data sets with excessive zeros, the tolerance limits of the methods were <50%. The zero-inflated Poisson, hurdle and negative binomial models are the same. So, we studied on an approach to settle the extremely excessive zeros in big data analysis. We did not depend on the statistical models such as generalized linear model based on probability distribution and link function. We divided the whole data with excessive zeros into analytic data sets according to the ratio of zero values. We have confirmed the possibility of solving the excessive zeros problem without using the statistical zero-inflated models from the results of this study. In addition, if the ratio of zeros in big data is too large, the popular zero-inflated models cannot be used. In this case, the proposed method provides effective performances. However, our research has limitations that rely on empirical methods rather than providing theoretically systematic methods. To solve this problem, we will develop a model that can provide more improved results with more theoretical framework in our future researches.

REFERENCES

Akritas, M., 2016. Probability and Statistics with R for Engineers and Scientists. Pearson Media Company, London, UK., ISBN:9780134995359, Pages: 528.

Anonymous, 2018. WIPS corporation. WIPSON, South Korea. https://www. wipson.com/service/mai/main.wips

Berman, J.J. and K. Morgan, 2013. Principles of Big Data: Preparing, Sharing and Analyzing Complex Information. Morgan Kaufmann, Burlington, Massachusetts, USA., ISBN:9780124045767, Pages: 261.

Bhujade, S.S. and M.B. Chandak, 2018. A hotel recommendation system for big data applications using a keyword aware approach. J. Eng. Appl. Sci., 13: 523-528.

Feinerer, I. and K. Hornik, 2018. Package tm Corpus. Text Mining Package, 10: 1-1.

Feinerer, I., K. Hornik and D. Meyer, 2008. Text mining infrastructure in R. J. Statistical Software, 25: 1-54.

Hashmi, A.S. and T. Ahmad, 2017. Frequency-based fast algorithm for anomaly detection in big data. J. Eng. Appl. Sci., 12: 7389-7392.

Hilbe, J.M., 2011. Negative Binomial Regression. 2nd Edn., Cambridge University Press, New York, ISBN-13: 9781139500067, Pages: 541.

Hilbe, J.M., 2014. Modeling Count Data. Cambridge University Press, New York, USA., ISBN:9781107028333, Pages: 284.

Hunt, D., L. Nguyen and M. Rodgers, 2007. Patent Searching: Tools and Techniques. Wiley Publishing Company, Hoboken, New Jersey, USA., ISBN:9780470116838, Pages: 256.

Jun, S., S.S. Park and D.S. Jang, 2014. Document clustering method using dimension reduction and support vector clustering to overcome sparseness. Expert Syst. Appl., 41: 3204-3212.

Jun, S., S.S. Park and S.D. Jang, 2012. Technology forecasting using matrix map and patent clustering. Ind. Manage. Data Syst., 112: 786-807.

Kim, J. and S. Jun, 2016. Zero-inflated poisson and negative binomial regressions for technology analysis. Intl. J. Softw. Eng. Appl., 10: 431-448.

Lu, J. and D. Li, 2013. Bias correction in a small sample from big data. IEEE. Trans. Knowl. Data Eng., 25: 2658-2663.

Micheline, K., J. Han and J. Pei, 2012. Data Mining: Concepts and Techniques. 3rd Edn., Elsevier, Amsterdam, Netherlands, ISBN:978-0-12-381479-1, Pages: 673.

Murphy, K.P., 2012. Machine Learning: A Probabilistic Perspective. MIT Press, Massachusetts, USA., ISBN:9780262304320, Pages: 1104.

Porter, A.L., A.T. Roper, T.W. Mason, F.A. Rossini and J. Banks *et al*., 1991. Forecasting and Management of Technology. Vol. 18, John Wiley & Sons, Hoboken, New Jersey, USA., Pages: 331.

R Core Team., 2018. R: A language and environment for statistical computing. Foundation for Statistical Computing, Vienna, Austria.

Ross, S.M., 2012. Introduction to Probability and Statistics for Engineers and Scientists. 4th Edn., Elsevier Publishing Company, Seoul, South Korea, Pages: 670.

Ryu, J. and S. Jun, 2016. A superpopulation model for patent big data analysis. Intl. J. Software Eng. Appl., 10: 153-162.

USPTO., 2018. A strong will. United States Patent and Trademark Office, USA. http://www.uspto.gov.