A Post-Aggregation Error Record Extraction Based on Naive Bayes for Statistical Survey Enumeration

Kiyomi Shirakawa, National Statistics Center, Tokyo, JAPAN
e-mail: kshirakawa@nstac.go.jp

Abstract

At the pre-aggregation stage of micro-editing in the National Statistics Center (NSC), an optimal editing of individual data has been implemented by consistency check and range check for each survey. In particular, some errors in numeric data items are logically detected based on a probability distribution of fixed parameters of the sample mean and standard deviation. However, when an acceptance region (range) for the error detection is narrow, a lot of correct data are detected. On the other hand, when the range is wide, many errors are not detected. In general, the problem associated with the range is revealed during a data review process at the post-aggregation stage. In the past, it was customary to edit manually. In order to solve this problem, it is necessary to extract the errors in the cell of the statistical table after aggregation. Therefore, we propose the introduction and systematization of naive Bayes that can treat a parameter as a random variable. In this method, the errors can be filtered by subjective probability. Thus, it is not a strict range check compared with objective probability, such as Smirnov-Grubbs' test. The filtering in subjective probability is dependent on a combination of several items and prior probability by unit of records. In addition, the validation test of the records may also include individual data with missing values. In this paper, we assess the error extraction method that focuses on sales variable in aggregated cells of the statistical table. As a result, it is possible to extract the errors in the point of view that is different from micro editing, and that the data editing process is automated.

Keywords: cell data, cross-tabulation table, filtering, inliers, subjective probability

1. Introduction

The role of the NSC in tabulation consists of a data entry, a data editing, a creation of statistical table, and that review. Of these, the most common ratio is the data editing. Therefore, the time of the view in the last process is limited. One example of the methods for the univariate data editing is the Smirnov-Grubbs test based on the probability distribution to fix the parameter of sample mean and standard deviation. The Grubbs' test is defined for the hypothesis:

H_0: There are no outliers in the data set

H_1: There is exactly one outlier in the data set

The Grubbs' test statistic is defined as:

$$G = \frac{\max|Y_i - \bar{Y}|}{s} \qquad (1)$$

with $\bar{Y}$ and s denoting the sample mean and standard deviation, respectively. The Grubbs' test statistic is the largest absolute deviation from the sample mean in units of the sample standard deviation [1][2].

Then, several graphical techniques can, and should, be used to help detect outliers. First of them is a box plot, a robust detection of the method based on order statistics such as the quartile. The calculated statistic value of this method is less susceptible to the outliers. Calculate the interquartile range (the difference between the upper and lower quartile) and call it IQ.

Calculate the following points:

C is a constant to be set on the basis of the distribution type of variable x.
In the case of normal distribution, C is 1.5 times.

$$\mathbf{C} < \frac{\mathbf{x - upper\ quartile}}{\mathbf{IQ}} \ \mathbf{or}\ \mathbf{C} < \frac{\mathbf{lower\ quartile - x}}{\mathbf{IQ}} \quad \mathbf{(2)}$$

Further, the detection of outliers can be made by creating a scatter plot of two variables. This figure provides information of the approximate line and correlation coefficient. Simple linear regression model is represented by $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, to predict and control the desired variable Y by the explanatory variable X of this model. Other, valid information is a prediction interval and a confidence interval. Furthermore, in the case of multivariate, the technical stratified analysis, such as stratified scatter plot can be also utilized. However, these techniques require experience and specialized knowledge. Further, it is customary to edit manually. That is a big burden on the editors at the work site. Therefore, we propose the introduction and systematization of naive Bayes that can treat a parameter as a random variable. As a result, it is possible to extract the errors in the point of view that is different from micro editing, and that the data editing process is automated.
In chapter 2 of this paper, we introduce a literature review of the detection of outliers in multivariate. Chapter 3 describes the data analysis, and the definition and properties of naive Bayes. Chapters 4 and 5 describe the discussion with the detection result of the outlier, and the last chapter, describes the conclusions and future work.

2. Literature review

Several studies have been reported on the effectiveness of the method of multivariate in the NSC [3] [4] [5]. The method of Wada[3] is MSD (Modified Stahel-Donoho), which is one of the recent studies. Also, studies by Okamoto[4] show more information in them. It was reported in detail M-estimator, MCD (Minimum Covariance Determinant), BACON (Blocked Adaptive Computationally efficient Outlier Nominators), Projection Pursuit, Forward Search, Epidemic Algorithm, Nearest-Neighbor Variance Estimation, Cluster identification, the methods of robust regression tree model based on order statistics. However, Okamoto pointed out some of these methods.

a. A multivariable calculation takes too much time by quantities of variables and data.
b. Data structure of sample surveys is complicated. Assumption is a normal distribution.
c. A date correction for the imputation and missing values in the data entry.

In particular, since the economic survey to investigate the accounting item is a monthly survey, it is reluctant to introduce new techniques. Therefore, these methods are not used for actual inspection.
However, the Economic Census for Business Activity survey in February 2012, for the first time in Japan, was conducted. Research object is six million sites in the number of establishments of all of Japan. In addition, accounting situations of office of four million or more were investigated. Therefore, it is a large number of establishments that do not compare to the sample survey. Therefore, automated detection of outliers in post-aggregation stage and data complement constant and missing values in the pre-aggregation stage of the result table were required. As a result, the detection of outliers in a different perspective from the micro-editing could be introduced. Furthermore, there have been few examinations of outlier detection in post-aggregation.

3. Methodology

3.1 Properties of naive Bayes

Naive Bayes is a simple Bayesian classification method which assumes the independence of the output probability. In addition, the naive Bayes classifier is the simplest in the Bayesian filter. Unnecessary information can be eliminated stochastically based on Bayes' theorem[6] [7].

Bayes' theorem gives the relationship between the probabilities of A and B, P (A) and P(B), and the conditional probabilities of A given B and B given A, P (A|B) and P(B|A). In its most common form, it is:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

In plain English the above equation can be written as, $\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}$

Extended form; often, for some partition $\{A_j\}$ of the event space, the event space is given or conceptualized in terms of $P(A_j)$ and $P(B|A_j)$. It is then useful to compute P(B) using the law of total probability:

$$P(B) = \sum P(B|A_j)P(A_j) \;\rightarrow\; P(A_j|B) = \frac{P(B|A_i)P(A_i)}{\sum P(B|A_j)P(A_j)}$$

3.2 Assumption of naive Bayes and definition of data

Definition of hypotheses based on naive Bayes is as follows.

Definition of assumption (H)

H_1: This data is an outlier., H_2: This data is not an outlier.

Definition of data(D)

$D=\{D_1,\ D_2\ ,\ \ldots\ D_{n-1},\ D_n\}$ (D is Extraction condition.)

Expansion formula of naive Bayes is as follows (Figure 1) [6] [7].

$$\mathbf{P(H_i|D_j) = \frac{P(D_j|H_i)P(H_i)}{P(D_j|H_1)P(H_1) + P(D_j|H_2)P(H_2)}}$$

$(j = 1, 2, 3, \cdots, n-1, n)$, likelihood is $P(D_j \mid H_i)$,

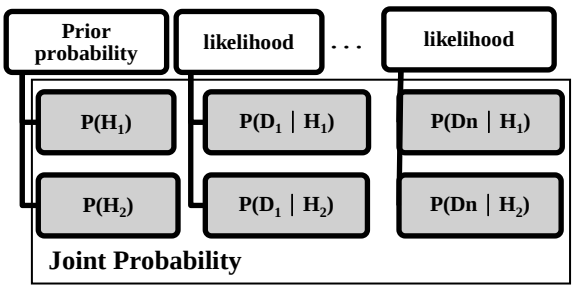
For the same data (D_j), denominator of the posterior probability $P(H_1 \mid D_j)$ and $P(H_2 \mid D_j)$ have in common. Therefore,

$P(H_1 \mid D_j) : P(H_2 \mid D_j) = P(D_j \mid H_1)\,P(H_1) : P(D_j \mid H_2)P(H_2)$

Set of prior probability based on past experience:

The probabilities of P (extraction) and P (exclusion), $P(H_1) + P(H_2) = 1$

Calculation of the ratio of the posterior probability:



$$P(H_1|D) = \frac{P(H_1)}{P(D)} \prod_j P(D_j|H_1)$$

$$P(H_2|D) = \frac{P(H_2)}{P(D)} \prod_j P(D_j|H_2)$$

It is outlier if $P(H_1|D) > P(H_2|D)$, otherwise it is not outlier.

Figure 1 **Expansion formula of naive Bayes**

In this research, we replace data by "0" or "1". Therefore, setting of boundary value to replace them is important.

3.3 Detection of outliers

3.3.1 Data sets

Required survey items in each economic survey are a sales and number of employees and capital. Therefore, the analysis uses the sales and the number of employees and capital. The results of the detection of outliers using univariate and statistics of sales in 2011 are as follows (Table 1). This table shows the large number of outliers.

Table 1 Datasets for analysis and number of outliers

Data sets	2011 EDINET 3,558 Records, 2010 EDINET 3,592 Records						
Variables	Capital, Number of Employees, Sales(Income), Industry classification(Major Groups), Ratio of Sales (Sales / Capital), Ratio of Employees (Employees / Capital)						
Statistics of Sales	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
	0	6,285	18,360	98,920	56,700	10,140,000	4
The Grubbs' test	Upper limit of outliers: 208,526 and more (287 Records)						
Boxplot	Upper limit of outliers: 132,323and more (443 Records)						

Note: EDINET is Electronic Disclosure for Investors' NETwork.
2010 EDINET 3,592 Records were used for data validation.

Theory of naive Bayes as described in the previous section is very simple. Further, there is a naive Bayes within the general-purpose of package R Language. However, in order to arbitrarily the settings for each parameter, we have developed a system that uses the EXCEL VBA of Microsoft. Consequently, it has enabled the inspection based on the conditional probability that the examiner calculated.

3.3.2 The detection of outliers for statistical table and boundary value of the conditional probability

Here, we examine the cell data in which we suspect that the value of the sales per enterprise is too high in the result table. Table 2 showed the value of a cell "Capital class 0-500 and Number of Employees class 201-300" is the highest.

Table 2 Result table for the detection of outliers

Sales per enterprise (Unit 1 million yen)		Number of Employees class(Unit 1 Person)				
		1-100	101-200	201-300	301-400	401-500
Capital class	0-500	3,178	5,824	15,466	12,780	17,494
	501-1000	4,626	9,347	12,045	14,150	14,085

We calculated the boundary value of outliers from the data in this table by box plot. The results of the calculation are shown in Table 3.

Table 3 Boundary value of outlier (500 million yen or less capital)

	Min	1st Qu	Median	Mean	Boundary value 3rd Qu.	Max.	3rd Qu + IQ * 1.5
Ratio of Sales (Sales/Capital)	32.8	641.5	1,401.5	3,004.8	2,477.1 or more	236,691.2	5,230.5
Ratio of Employees (Employees/Capital)	0.8	18.6	40.3	63.4	79.2 or more	1,209.8	170.1

3.3.3 Analysis of data

For detection of outliers, it is necessary to determine prior probability and the

likelihood, before calculating the posterior probability. Table 4 shows the probability of using the analysis. Further, it is possible to modify the probabilities at the discretion of the analyst.

Table 4 Each probability of using the determination

	Prior probability	Likelihood			Posterior Probability (The product of all probability)
		Ratio of Sales	Ratio of Employees	Industry	
Normal value	0.85	0.20	0.30	0.40	0.38
Outliers	0.15	0.75	0.60	0.50	0.62

In this case, prior probability is 0.15. Table 5 shows the pattern of outliers. Thus, the number of detected data can be controlled by changing the prior probability.

Table 5 Prior probability of outliers of each pattern (Excluding all zeros)

	Industry	Ratio of Sales	Ratio of Employees	Prior probability of outliers
Pattern of outlier	1	1	1	0.10 or more
	0	1	1	0.15 or more
Pattern of normal value	1	0	0	0.45 or more
	1	1	0	0.20 or more
	1	0	1	0.70 or more
	0	1	0	0.25 or more
	0	0	1	0.35 or more

4.Results

The number of target data in this research is 34 records in the cell. Result of detection, 4 data are detected as outliers. Ranks of each data of sales are 1st, 7th, 13th, and 20th. Therefore, this detection is random. Whether or not these are positive values, we have confirmed the contents of the Annual Securities Report of the previous year and the year. Figure 1 and Table 5 show which patterns are to contribute to the detected outliers. Knowing in advance, it is possible to control the number of data detection; thus, the introduction to the practice can be achieved.

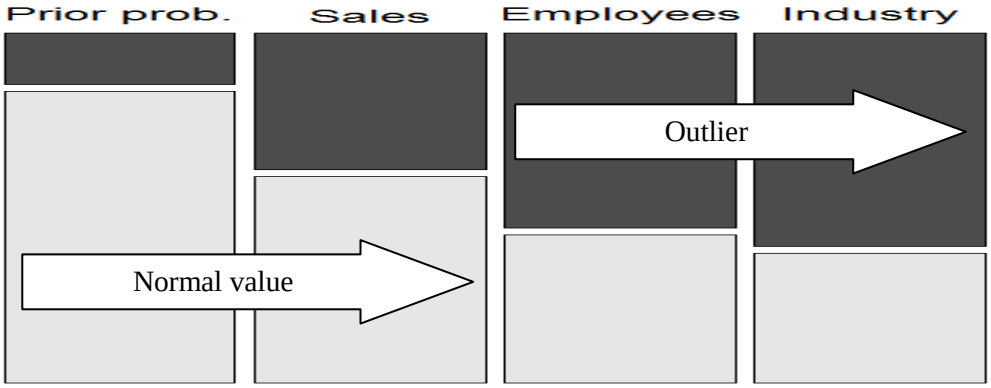


Figure 2 Case of the product of all probability

5. Discussion

We have introduced this method for the detection of outliers in the preliminary

summary of the 2012 Economic Census for Business Activity. Result table including accounting items applied is "Table 8 Enterprises, etc., Sales (Income), Expenses and Added Value by Industry (Major Groups) and Legal Organization (3 Groups) for Japan and Prefectures". As a result, we extract the data that there was an error in the sales. There was a missing value to the other items in some of the data. Thus, we have extracted as inliers data[8][9] as normal value during data editing. However, all of the data detected by this method are not the error data. In the determination of the error data, we defined as "innocent until proven guilty (beyond a reasonable doubt)". It should be noted that the work was done in a short examination period; therefore, the number of data that can be verified is small. Furthermore, since the result table number for this aggregate is small, it is not able to measure the effectiveness of this method.

6. Conclusions and future work

Outlier detection based on cell aggregation of the result table is the clues to solve some difficult problems faced by the analyst. The clue is the fragmentation of large amounts of data and the determination of the variables in the multivariate data.

In addition, by using a simple calculation of naive Bayes, we have an ambiguous classification by subjective probability and handling of missing values. This ambiguous classification is dependent on the environment of strict examination after data editing. Analysis of the cell data is to determine the target location of the examination, it is found that this does not mean that the outlier is error data. Future research is an extension to other economic surveys based on the experience of authentic information in the aggregate 2012 Economic Census for Business Activity. In particular, the economic surveys for the examination of the results table in the NSC are monthly or yearly. In each of the economic surveys, there are a lot of data that can be compared with, unlike the Economic Census, which is conducted every five years. Therefore, it is possible to analyze many aspects. A future plan is the outlier detection by pattern recognition and machine learning. In addition, we are planning to study methods for the detection of complement fixed in excess computing data directory.

References

[1] Frank E. Grubbs. (1950). ”Sample Criteria for Testing Outlying Observations,” The Annals of Mathematical Statistics vol.21, no.1: 27-58.

[2] Harry Zhang. (2004). ”The Optimality of Naive Bayes,” International Florida Artificial Intelligence Research Society Conference, FLAIRS, AAAI Press, Miami Beach, Florida, USA.

[3] Kazumi, Wada. (2010). “Detection of Multivariate Outliers: Modified Stahel-Donoho Estimaitors”, No.67 Vol.4, Ministry of Internal Affairs and Communications Statistical Research and Training Institute.(In Japanese)

[4] Masato, Okamoto (2004). “*Tahenryo Hazurechi kenshutsu no kenkyu douko oyobi Canada oroshiuri kourigyochosa ni okeru tahenryo hazurechi kenshutsuho*“, *Seihyo gijyutsu kenkyu report 1*, National Statistics Center. (Non-disclosure)(In Japanese)

[5] Franklin, S. and Brodeur M. (1997). “A practical application of a robust multivariate outlier detection method“, Proceedings of the Survey Research Methods Section, the American Statistical Association, PP. 186-191

[6] Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze (2008). “Introduction to Information Retrieval“, Cambridge University Press; 1 edition

[7] Harry Zhang. (2004). “The Optimality of Naive Bayes,” International Florida Artificial Intelligence Research Society Conference, FLAIRS, AAAI Press, Miami Beach, Florida, USA.

[8] William E. Winkler.(1997). “Problems With Inliers”, Conference of European Statisticians. Work Session on Statistical Data Editing, UNECE

[9] United Nations Geneva (2000). “Glossary of Terms on Statistical Data Editing” (Conference of European Statisticians Methodological Material)