Principle, Methodology and Application for Data Cleaning techniques

Zhexu Jin*

Gengdan Institute of business, Beijing University of Technology Beijing, China

*Corresponding author: 161843145@masu.edu.cn

Abstract. Con-temporarily, information data has become the cornerstone of every company's decision-making. In a vast flow of information, choosing the right data is the first step in developing successful predictions. After the determinations of the requirements, analysis purpose and prediction direction, outlier processing, missing value processing and repeated value processing are usually encountered. This paper introduces the limitations, advantages and disadvantages of different methods in application in detail. At the same time, this paper introduces some interpolation methods based on mathematical statistics, such as thermal interpolation, Lagrange interpolation and Newton interpolation. At the same time, it also provides the normal distribution processing method which is better in dealing with outlier problems, and the popular K-nearest neighbor algorithm. Finally, it illustrates the logic diagram of data cleaning in the data preparation stage. Overall, these results offer a guideline for selecting the appropriate treatment in the corresponding situation during data cleaning process.

Keywords: Data cleaning; Thermal Interpolation; K nearest neighbor algorithm; Lagrange and Newton interpolation.

1. Introduction

Data Cleaning, often called data cleansing or scrubbing, is an indispensable part of statistical analysis, also being the second step for the whole process of data analysis [1]. Thus, whether it is a research report or prediction report needs to deal with data at great length. The reason for such meticulousness lies in the fact that data errors, repetitions, outliers, loss (null value) and blank will directly affect the quality and accuracy of the results. Meanwhile, as a machine learning and computer bridge, this part is a paramount foundation for entering the field of machine learning or business analysis in the future. On the other hand, Data Cleaning will spend more longer time than other step, this may be a tedious process for many data analysts. However, if the analyst wants to get an accuracy and comprehensive investigation result, one must face this hard section.

The importance of Data Cleaning also reflected in plenty of giant internet companies in modern society since everyone is immersed in the trend of big data era. On this basis, various types of data, (e.g., the kinds of purchases one makes in our daily lives, footprints, billing information, spending habits, browsing history, income) all become valuable information, which play a huge role for commercial companies, determining the enterprise's target market division, brand positioning, and marketing strategy. Generally speaking. When human beings encounter problems, it may be the data support needed by the project or the paper. The quality of data ultimately determines the accuracy of data analysis, and data cleaning is the only way to improve the quality of data, so one needs to pay more attention to the data that needs to be cleaned [2].

In order to elaborate the various methods, this paper introduces applicability and limitations of each approach. The rest part of the paper is organized as follows. Starting from the Sec. II, this paper introduces that simple data cleaning method (graphic method) is suitable for the smaller data sets, e.g., university job, or a small company. In Sec. III, this paper introduces four methods of missing value processing, and introduces the application scope and limitations of four different methods, including Lagrange interpolation method and Newton interpolation method based on mathematical statistics. In Sec. IV, this paper introduces the application scope and limitations of outlier processing in data discretization problem by using the mathematical analysis method of normal distribution and box plots respectively and introduces in detail the K-nearest Neighbor algorithm, which is very popular at

present. In Sec. V, this article looked at the impact of duplicate values on data, hence it introduced two different methods of filtering [3] and removing duplicate data from the Data-frame format and the table format. Subsequently, the logic flow chart of data cleaning is given. Eventually, a brief summary is given in Sec. VII.

Different data cleaning methods have their own advantages and disadvantages. Big data development, mining and application are becoming more popular, but dirty data is ubiquitous. Dirty data refers to the data in the source system that is not in a given scope, or is meaningless to the business, or the data format is illegal. In general, there is non-standard coding or ambiguous business logic in the source system. The data mined is basically the actual data from production, life and business. Various reasons may lead to the loss of some important data, incorrect data collected, noise, inconsistency and other problems. Data quality is an important guarantee for the effect of data mining. Therefore, the research on data cleaning method is of great significance and also faces many problems to be solved urgently.

2. Simple graphical Data Cleaning

2.1 Histogram analysis

Simple graphical data cleaning is suitable for small data scale (~1000), which is very simple and applicable. For example, the bar chart can clearly arrange the data [4]. In daily work and questionnaire collection, this simple method can quickly distinguish outliers in the data. It enables analyst to see the size of each data at a glance and this method easy to compare the differences between data. It clearly shows the advantages of quantity: the bar chart can clearly show the number of various quantities

2.2 Line chart analysis

Polyline analysis is also classified as a simple method of data cleaning, which is only suitable for small-scale data. Compared with the columnar analysis method at point A, the discount analysis method is easier for the analyst to see the inflection point. This means that analysts can see not only how many numbers are in the data, but also how the numbers are changing and how fast they are changing.

2.3 Scatter diagram analysis

Scatterplot is mainly the most intuitive graph to measure the relationship between two variables. The relative distance of scattered points is mainly reflected in the correlation coefficient, and its trend is reflected by the regression coefficient or slope. Although three-dimensional graph also plays a similar role, it still cannot replace the effect of two-dimensional graph in terms of intuition.

2.4 Quantile-Quantile Plot

In statistics, a Q-Q plot (Q stands for quantile) is a probability plot that graphically compares two probability distributions by putting their two quantiles together. First pick the quantile interval. The points (x, y) on the graph reflect the quantile of one of the second distributions (y coordinates) and the same quantile of the corresponding first distribution (x coordinates). Thus, the line is a curve with quantile intervals as its parameter. If the two distributions are similar, the q-q plot tends to fall on the y=x line. If the two distributions are linearly related, then the point tends to fall on a straight line on the q-q plot, but not necessarily on the y=x line. Q-q plots can be used to visually evaluate parameters in the position-scale category of the distribution.

Can be seen from the definition of Q-Q chart is mainly used to test the data distribution similarity, if you want to use the Q-Q chart to test of normal distribution of data, you can make the x axis for normal distribution quantile, y axis quantile for sample, if the two points of distribution on a straight

line, will prove the existence of the sample data with normal distribution, linear correlation It's normally distributed.

3. Missing value handling in data

In the data cleaning method, similar to the problem classification proposed in Ref. [1], lots of big companies may face the loss of values caused by various problems. Therefore, it is necessary to find the most reasonable way to deal with data loss according to different situations.

3.1 Direct deletion method

When there are not many missing values in a data set (e.g., less than 5-10%) and these data are random, normally can delete these data directly, which will not significantly affect the analysis.

3.2 Mean, mode, median substitution method

When data is missing, it will be properly to use the average, mode and median in these data to supplement the missing value. As a matter of fact, the analyst must admit that everything has two sides. To be more specific, the advantage of this method is that supplementary data do not reduce the total, and process is very simple. In this case, one only needs to find superior numbers and calculating the mean or median can be achieved. However, it also brings the disadvantage, because the data filled in is not random data, it will produce certain data deviation (according to the size of your average, towards an extension trend). Therefore, the proper method is that the mean value can be used to replace the data with normal distribution (there is not much difference between the mean value and the overall data), but if the data skew is too large (high distribution or low distribution), it is better to use the median [5].

3.3 Optimum Thermal Interpolation method

Thermal Interpolation is a very popular and effective method of data interpolation. It means to find a sample that similar to the missing value in the non-missing data set for interpolation. Whereas this method also exists merit and weakness. The advantage is that when you select samples that match the missing values, this interpolation method makes the results more accurate. The disadvantage is that when there are too many missing values and most of the missing values are distributed randomly, it is not necessarily possible to carry out one-to-one matching between non-missing data and data requiring interpolation [6].

However, a good solution here is that stratifying the data into different intervals according to the original data. When there are many random variables, one needs to match the missing values according to stratification to complete data interpolation.

3.4 Lagrange interpolation method and Newton interpolation method

Some methods in mathematics are introduced here to solve the interpolation problem, first a brief introduction is given for Lagrange interpolation method. Firstly, an interpolation data point function is generated. For each $1 \leq i \leq n$, $P(X_i)=Y_i$, and the data interpolation points are (x1, y1), (x2, y2), …, (X_n Y_n). Given n data points, the analyst wants to find an interpolation polynomial that passes through all the above points. The set of forms that satisfy all of these conditions is called the Lagrange formula.

In terms of Lagrangian interpolation, we first need to talk about the denominator.

$$L_k(x) = \sum_{i=0, i\neq k}^{j} \frac{x-x_i}{x_k-x_i} \quad (1)$$

It should be noted that $x_i \neq x_j$ since they are different roots. As can be seen from the above formula, i increases from 0 to j, but k will not be equal to i. If $k=i$, the denominator becomes 0, and this formula

is meaningless. After given k, *i* will increase from 0 to *j*. This LK (x) is a Lagrange interpolation function

$$L_k(x) = \frac{(x-x_1)\cdots(x-x_{k-1})(x-x_{k+1})\cdots(x-x_n)}{(x_k-x_1)\cdots(x_k-x_{k-1})(x_k-x_{k-1})\cdots(x_k-x_n)} \quad (2)$$

One can clearly know that when L_k(x) is equal to 1, X_j is equal to 0, and X_j can be represented as any data point, thus there is a new definition of n-1 polynomial.

$$P_{n-1}(x) = y_1 L_1(x) + \cdots + y_n L_n\,(x) \quad (3)$$

Based on simple substitution, the final Lagrange interpolation formula can be derived:

$$P(x) = \sum_{k=1}^{n} y_k L_K(x), \quad L_k(x) = \prod_{i=1, i\neq k}^{n} \frac{x-x_k}{x_k-x_j} \quad (4)$$

As same as the thing for formula 1, when *k* is equal to *i*, the denominator becomes 0, and this formula is meaningless. Each y value has a particular Lagrangian interpolation, and what we end up with is a polynomial that multiplies each y by its particular Lagrangian interpolation

The advantage of Lagrangian interpolation is that when you have a lot of missing data insertion points, you can use those interpolation points to calculate the results and get more accurate values [7].

The disadvantage of Lagrange interpolation is that it is difficult to determine the proper starting point because of its mathematical complexity [7]. Moreover, the core of this method is to construct a set of basis functions and let your polynomial satisfy this set of basis functions, i.e., it requires high mathematical requirements and logical ability of users.

Newton interpolation method and Lagrange interpolation method are both belong to the category of algebra interpolation, both through the given interpolation nodes are constantly calculating curve interpolation function. Newton interpolation method will be transformed into Lagrange interpolation method when it satisfies certain conditions in the certain conditions, but the advantage of Newton interpolation method is that it is suitable for growing situation by the insertion point. This is the key of the Newton interpolation method, which are function itself is unchanged, and this is difference between the Newton interpolation method and Lagrange interpolation method.

There is another advantage is that Newton interpolation method is more cases of the derivative in the discrete case and approximation, which means that it can reduce more than the amount of calculation of Lagrange interpolation method.

4. Outlier handling (Data discretization problem)

4.1 Comprehensive mathematical analysis method

Facing with a large amount of data, the analyst usually does not choose function mapping, which is difficult and very slow to draw due to overloaded data.

After obtaining huge data, the analyst can conduct a comprehensive descriptive statistical analysis of the data, such as determining the maximum and minimum values, range, variance, mean and standard deviation in the data. The most important step is that arranging the data in a normal distribution.

The Gaussian distribution, also known as the "normal distribution", is denoted as N (μ, σ^2) if the random variable X obeys a normal distribution with a mathematical expectation of μ and variance of σ^2. The expected value μ of the normal distribution determines its position, and the standard deviation σ determines the magnitude of the distribution. The normal distribution with μ=0 and σ=1 is the standard normal distribution. It can be defined as:

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \quad (5)$$

The variance, by the formula.5, is the degree to which the observed data diverges from the central trend, which is the mean, which is a generalization of the observed data.

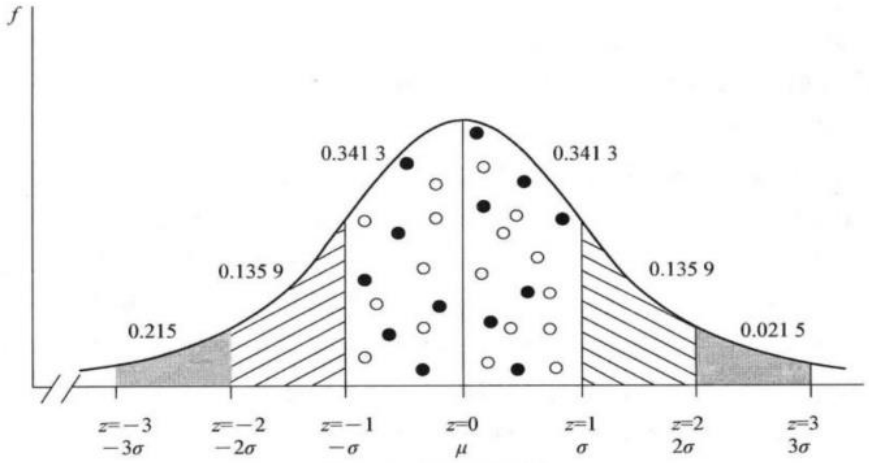

Figure 1. A sketch of Normal distribution diagram, i.e., z satisfies N (0,1) [8], where y axis represents the probability density and x axis denotes the normalization value of z (minus the mean value μ and divides the standard error σ).

The Y-axis of the normal distribution is the probability that the random variable x is equal to some number. The area of a certain interval on the X-axis under the normal curve reflects the percentage of the number of cases in this interval to the total number of cases, or the probability of the variable value falling in this interval (probability distribution). The area under the normal curve in different ranges can be calculated by Eq.5

If the data follows a normal distribution, mathematically the analyst would know that a minimum probability event is 0.003 when a set of measurements deviates more than three standard deviations from the mean. In this way, the analyst can distinguish the set of measurements from normal data by defining them as outliers

To sum up, the advantage of mathematical analysis (using the normal distribution method) is that it is very convenient to identify outliers by determining multiples of standard deviations. Nevertheless, its disadvantage is that when the data does not follow the normal distribution, though the analyst can use a method to describe it many times away from the average value, this method may fail when the data volume is too large [8, 9].

4.2 Box diagram analysis method

According to the Fig. 2, This is an example of a box plot. the method to judge outliers in box plots is based on quartile and quartile spacing. If a value is less than QL-1.5IQR or greater than QL+1.5IQR, it is said to be an outlier. QL is the lower quartile, indicating that one quarter of the observed values are smaller than QL. QU is the upper quartile, indicating that one quarter of all observed values are greater than QU. IQR is the interquartile spacing, the difference between the upper quartile QU and the lower quartile QL, and contains half of all observed values.

To sum up, the advantages of the box plot are that it provides a standard for identifying outliers and that quartiles are robust (25% of the data can be varied at will without interfering with the quartiles). Therefore, the identification of outliers is objective [10,11].

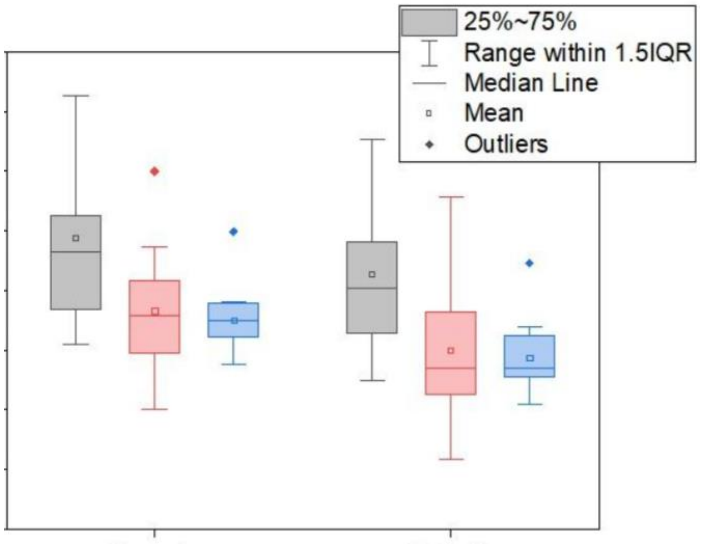

Figure 2. A sketch of box diagram [10].

4.3 K nearest neighbor method

KNN is a basic classification and regression method. The input of K nearest neighbor is the data set of test data and training samples, and the output is the category of test samples. K nearest neighbor algorithm does not show the training process. During the test, the distance between the test sample and all the training samples is calculated, and the prediction is made by majority voting according to the category of the nearest K training samples. The algorithm is described as follows: (1) set training sample and (2) each sample contains m characteristic dimensions

$$X = \{X^{(1)}, X^{(2)}, \cdots X^{(n)}\}, x^{(i)} = \left\{x_1^{(i)}, x_2^{(i)}, \cdots x_m^{(i)} \in R^m\right\} \tag{6}$$

Each sample contains m characteristic dimensions, the distance of L_P can be defined as:

$$L_p\left(x^{(i)}, x^{(j)}\right) = \left(\sum_{k=1}^{m} |x_k^{(i)} - x_k^{(j)}|^{\,p}\right)^{\frac{1}{p}}; p \geq 1 \tag{7}$$

When P=1, it is called Manhattan distance:

$$L_p\left(x^{(i)}, x^{(j)}\right) = \left(\sum_{k=1}^{m} |x_k^{(i)} - x_k^{(j)}|\right) \tag{8}$$

When P=2, it is known as Euclidean distance

$$L_p\left(x^{(i)}, x^{(j)}\right) = \left(\sum_{k=1}^{m} |x_k^{(i)} - x_k^{(j)}|^{\,2}\right)^{\frac{1}{2}} \tag{9}$$

When P=∞, it corresponds to the maximum in the range of coordinates

$$L_\infty\left(x^{(i)}, x^{(j)}\right) = \left(max\left|x_k^{(i)} - x_k^{(j)}\right|\right) \tag{10}$$

By roughly dividing the sample points into two or more regions distributed on the coordinate plane, two or more different colors will be selected to mark the sample category. Outlier points can be classified by calculating distance. The choice of K value will have a great impact on the results of the algorithm. If the value of K is small, it is equivalent to using the training instance in a small neighborhood for prediction. In extreme cases, K=1, the test instance is only related to the closest

sample, and the training error is very small. However, if the sample happens to be noise, the prediction will be wrong, and the test error will be very large. In other words, when the value of K is small, the phenomenon of overfitting will occur.

If the value of K is large, it is equivalent to using the training instances in a large neighborhood for prediction. In extreme case (K=n), the result of the test instance is the class with the most instances in the training data set, which will produce under-fitting.

To sum up, in the application, small K is generally chosen, and K is odd. Cross validation is usually used to select the appropriate K value. It possesses the advantages of operation simply, easy to understand, high precision, mature theory, can be used to do both classification and regression. It can be used for numerical data and discrete data [12]. However, it needs a lot of storage space specifically for large data and can be very time-costing ascribed to the calculation process (calculating the distance between the samples to be tested and all samples in the training data set). Therefore, K nearest neighbor algorithm is not effective for unbalanced data and needs to be improved [13].

5. Duplicate data handling

Taking the Data Frame data format as an example, the data contains duplicate data can set simple filtering instructions to delete duplicate data. If the data is in tabular format, the iterative method can be applied to delete duplicate data. Assigning x, the initial value of iteration; The iterative method (also known as the "rollback" method) is a process in which old values of a variable are constantly used to recursively derive new values. The opposite of the iterative method is the direct method (or first-solution method), where the problem has been solved once. Iterative algorithm is to use a computer to solve the problem of a kind of basic mode, which USES the computer's speed, suitable for repeated operation characteristics, let the computer to a set of instructions (or steps) must be in the implementation of this set of instructions executed repeatedly each time (or the steps), due to the original value of the variable is the new value, so the iteration method is divided into precise iterative and approximate iteration. Typical iterative methods (such as "dichotomy" and "Newtonian iteration") are approximation iterative methods.

$$f(x) = ax^3 + bx^2 + cx + d \tag{11}$$

Newton's iterative method is a tool for finding the zero of an arbitrary function. It's much faster than the dichotomy.

The formula is x=a-f(a)/f'(a). Where ais the guess value and x are the new guess value. keeping iterating, and $f(x)$ gets closer and closer to 0

Then one takes the derivative of that and then you take the derivative of that formula, just to make the approximation more accurate

$$f'(x) = 3ax^2 + 2bx + c \tag{12}$$

Computing the increment f(x)/f '(x), the next x: x-f(x)/f '(x); Replacing the newly generated x with x: x=x-f(x)/f '(x). If the absolute value of D is greater than 0.00001, one needs to repeat the above steps. The method is described as follows:

$$y_k = P(x_k), z_k = P(y_k), x_{k+1} = x_k - \frac{y_k^2 - 2y_k x_k + x_k^2}{z_k - 2y_k + x_k} \tag{13}$$

Duplicate values in the process needs to pay special attention to, a complete data cleaning process usually begins with duplicate values and the missing value. Although the duplicate values are generally taken delete method to deal with, but some duplicate data in some special research topics, e.g., the

order transaction details, goods such as repeat purchase rate is in the business analysis cannot be deleted [14].

5.1 Data Cleaning logic

At present, there are many data cleaning methods in the process of data pretreatment, and different data cleaning methods have advantages and limitations. Therefore, a logical framework can be roughly generated according to specific problems.

First of all, according to different fields, analysts need to determine the requirements for analysis and prepare corresponding data to start data cleaning. When starting data cleaning, the analyst should check the data set for the existence of the following three main problems: missing values, outliers, and duplicate data (for commodity categories, purchase rates, etc., which require duplicate data are not considered).

When the analyst corrects the data, fills in missing values, identifies outliers, and removes duplicates, this is clean data that represents real user information, and the results are more accurate and cleaner [15,16].

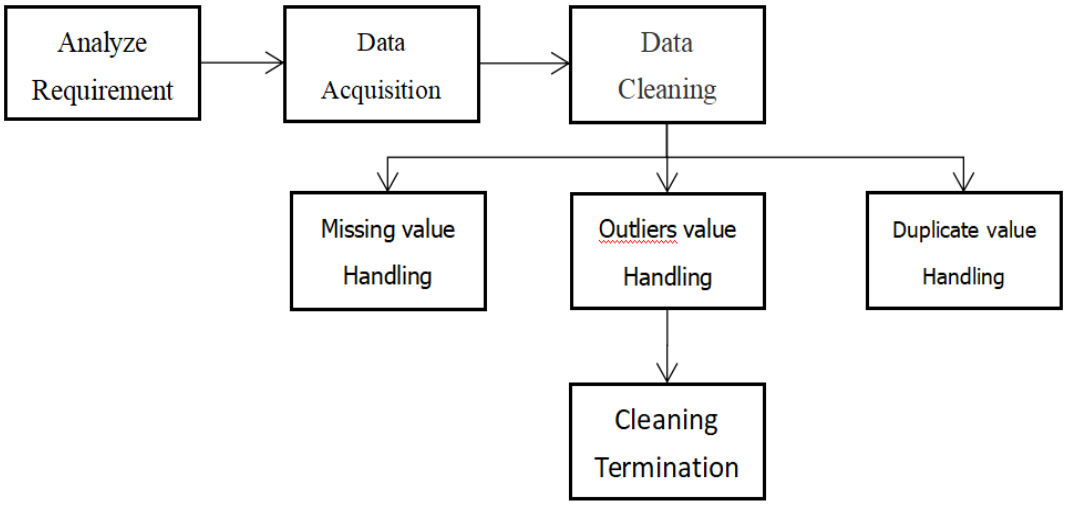

Figure 3. The framework of data cleaning.

6. Conclusion

In summary, this paper introduces Data cleaning methods and analysis some limitations in practice and analyzes some data cleaning literature to find some good methods and explains them in the paper. In the Sec. 2.1, this paper discusses that graph analysis is efficient in the case of small data scale, but not applicable in the case of large data. In the Sec. 2.2, when faced with missing values, this paper provides a variety of methods to deal with missing values and discusses the situation that the median mode is sometimes better than the average in mathematical statistics, and the reasons why Lagrange interpolation is difficult to achieve in the Sec. 2.2.4 though it has a good effect. In the Sec. 3, this paper discusses the possible outlier problem and puts forward several methods to deal with outliers. For example, if the data follows the normal distribution, the normal distribution graph is a good method to distinguish outliers, and the effective boxplot is also used to analyze the advantages and limitations. There is no doubt that there is a very popular K-nearest neighbor algorithm in machine learning, which can not only classify data, but also perform outlier discrimination.

According to the limitations of the above methods, the scope of application of each method is clearly defined in this paper, such as whether the sample size of data is too large, too small, whether it follows normal distribution, and whether the difference with the mean value is too large. Therefore, the method provided in this paper should be used in combination with the specific situation.

Finally, this paper summarizes the methods in the process of data cleaning, analyzes the advantages and disadvantages, and analyzes the application situation. Each method has its limitations, it must

carry out data cleaning according to the demand, purpose and prediction direction after getting the data. Combined with the method mentioned in this paper, good results should be achieved. Overall, these results offer a user guide for data cleaning.

References

[1] Rahm, E., and Do, H. H, "Data cleaning: Problems and current approaches," IEEE Data Eng. Bull., 23(4), 3 - 13, 2000.

[2] Dasu, T., and Johnson, T., "Exploratory data mining and data cleaning," John Wiley & Sons, 479 2003.

[3] Rachael Tatman. Kaggle. Data Cleaning Challenge, 2009.

[4] The diagram is collected from: What is machine learning for data cleaning. Available at: https://blog.csdn.net/HowieXue/article/details/104270918.

[5] Ragel A, Cremilleux B. "Mvc-Areprocessing method to A study on the Influence of different parameters on the performance of the System," Knowledge - Based System Journal, 12 (5/6): 158 - 163, 1999.

[6] Shen J. J, Chang C. C., Li. Y. C, "Combined association rules for dealing with missing values," Journal of Information Science, 33 (4): 246 - 254, 2007.

[7] Xu Xiaoli, "Implementation of Lagrange interpolation algorithm in Engineering Application," Forest Teaching, 01: 17 - 19, 2010.

[8] Yu Bingjie, "Anomaly Detection Algorithm based on Gaussian Model," China University of Mining and Technology, 2017.

[9] Wang Qilong, "Research on image classification method based on Gaussian distribution modeling," Dalian University of Technology, 2018.

[10] Peng Jun, Mo Hongwei, and Yuan Guiguai, "Visualization of box chart of sample data using R software in statistical teaching in colleges and universities," China new communications, 22 (21): 186 - 187, 2020.

[11] Yang Bin, "Comparison of several methods of normality test," Statistics and Decision, (14):72 - 74, 2015.

[12] Pei Ya-chen, "Application research of k-nearest neighbor classification algorithm," Communications world, 26 (01): 286 - 287, 2019.

[13] Wang Qinghua, Liu Jiangwei, and Zhang Lanlan. Journal of Xi'an polytechnic university, 2015, 35(02):

[14] Cao Hai, Sun Jing, and Shi Xibin. "Short Text Duplicate Removal Algorithm Based on Feature Iteration," Computer Engineering, 41 (12): 54 - 57, 63, 2015.

[15] Hao Shuang, LI Guo-liang, Feng Jian-hua and Wang Ning, "Overview of Structured Data Cleansing," Journal of tsinghua university (science & technology), 58 (12): 1037 - 1050, 2018.

[16] Krishnan, S., Franklin, M. J., Goldberg, K., Wang, J., and Wu, E, "Activeclean: An interactive data cleaning framework for modern machine learning. In Proceedings of the 2016 International Conference on Management of Data," pp. 2117 - 2120, 2016.