International Core Journal of Engineering Volume 6 Issue 10, 2020
ISSN: 2414-1895 DOI: 10.6919/ICJE.202010_6(10).0041

Method Analysis of Missing or Incomplete Data Population

Liping Wang

Chengdu University of Technology, Chengdu 610059, China.

Abstract

With the popularization of the Internet, more and more people choose to watch TV, go shopping and other activities through the Internet, which greatly enriches People's Daily life. The widespread use of the Internet marks that we have entered an era of information explosion, with more and more fragmented information coming into our view every day. The fragmented information is recorded in the form of information tables, which are data. With the advent of the era of information "big bang", the practical application of data mining technology is becoming more and more important. In the process of data analysis and mining, incomplete data is often found in the information table. Therefore, it is of great significance to study the corresponding data mining methods for missing data in the actual application process for the completeness and accuracy of the entire data network. Therefore, by analyzing the missing data and its causes, this paper discusses the methods and examples to fill the incomplete data in the data preprocessing, hoping that the corresponding technicians can draw lessons from it.

Keywords

Data mining, Missing data, Data population, Deep learning.

1. Introduction

In data mining ProjectThere are always some incomplete data In data set [1]. It maybe error value or missing value, this can influence the real mining effects, If we mine on incomplete data set directly. Therefore, filling the missing data is necessary;Braking, it be called as completing of incomplete data [2]. Through the research of data mining analysis shows that about 20% of the time is used to identify the target, and the remaining 60% of the time need to prepare the corresponding data, eventually to data analysis and mining work is only about 10% of the time, thus the data mining technology in the data is incomplete, and many problems such as data inconsistency and these issues will be a great influence on final results corresponding algorithm. So in the process of practical application, in order to ensure the accuracy of the data mining, data preprocessing work is needed in the staff spend a lot of time and energy, in the process of the data of compression processing need to always make sure not to make the data contains specific information produces deviation, effectively improve the overall quality of the data, thus improve the efficiency in the process of algorithm in practical application. At present, missing data processing in data preprocessing has become a key problem in data mining. In this paper, several common missing data processing methods are simply analyzed and discussed, hoping that relevant technicians can get inspiration and ideas from them.

2. Missing data and its causes

In this section, we focus on the causes of the missing data. The reasons for missing values are various, which can be mainly divided into objective reasons and human reasons. The objective reason is the lack of data caused by failure of data collection or data storage caused by equipment and technology, such as failure of data storage, insufficient accuracy, mechanical failure, etc., which leads to failure of data collection in a certain period of time (for timing data collection). The human reason is the lack

ISSN: 2414-1895 DOI: 10.6919/ICJE.202010_6(10).0041

of data caused by human subjective error, historical limitation or deliberate concealment. For example, in the process of online shopping, the purchase data of users cannot reflect the purchase preference of users due to some accidental reasons, so the collected user preference data is not accurate.

2.1 Types of missing values

Missing values can be divided into completely random missing, random missing and completely non-random missing from the distribution of missing. Missing completely at random (MCAR) means that the loss of data is random and does not depend on any incomplete variable or complete variable. Missing at random (MAR) means that the missing of data is not completely random, that is, the missing of such data depends on other complete variables. The missing not at random (MNAR) means that the missing data is dependent on the incomplete variables themselves.

In terms of the attribute belonging to the missing value, if all the missing values are the same attribute, then such a missing value becomes a single missing value. If the missing value belongs to different attributes, it is called any missing value.In addition, for the data of time series, there may be deletion with time, which is called monotone deletion.

2.2 Reasons for data loss

In the process of establishing various forms of database, the lack of data is impossible to completely eliminate, so whether the data in the database is accurate and complete enough is also uncertain. Data loss is inevitable, and there are many reasons for it, among which the most important are the following aspects.

First of all, in the actual information collection work, the information is lagging behind and cannot be obtained in a timely manner. For example, in some hospital databases, the clinical test results of each patient are difficult to be determined in a short time, which makes some specific values appear vacant.Secondly, some information is likely to be forgotten in the process of information entry, which may be caused by human factors or the failure of the corresponding storage equipment. Thirdly, when obtaining certain information, the cost is too high, which is not in line with the actual economic interests of enterprises. Moreover, some systems have high requirements on the real-time performance of information, and in many cases, conclusions are required before obtaining the corresponding information, leading to the inevitable absence of data. In addition, some information systems automatically omit information because it is not important by default. For example, there is no direct correlation between a specific attribute value and the given context. The existence of these problems will lead to data missing in the final data entered by the system.

3. Several methods of data filling

Due to the universality and non-resistance of data missing, how to fill data and improve the integrity of data filling have high research value. Data filling technology has been applied in many fields. Therefore, with the development of science and technology, people gradually put forward many methods of data filling, such as neglect method, direct deletion method, fixed value filling method and special value filling method. For different types of missing values, the filling method is different. Since the direct deletion method and the fixed value filling method have been gradually abandoned, this paper mainly introduces the special value filling method. For the special value filling method, the commonly used pretreatment methods at present include KNN algorithm and expected maximization filling method (EM), data filling algorithm based on rough set, compressed sensing method, etc.

3.1 KNN algorithm

K nearest neighbor algorithm (KNN) is the simplest basic classification algorithm proposed by Cover and Hart in 1953 [3].The basic principle of the algorithm is as follows: if the distance between each data point in the sample data set is closer and closer, then they are more similar.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202010_6(10).0041

Therefore, the category of the target sample can be predicted according to the category of K most similar samples. The method of filling based on K nearest neighbor idea is to find similar samples with real samples, and the data can be filled according to the values of similar samples. KNN method of filling the concrete steps are: in the training data and the condition of known labels, enter test data, the characteristics of the test data and training focused on to compare the characteristics of the corresponding, and find the most similar of training focus and former K data, then the test data, the corresponding category is K a classification of the data in the most times.

KNN filling method is simple in operation, and has high applicability to both classified data and discrete data. It has a good performance on data sets with strong correlation.However, if the data set is large with many missing values and there is no significant correlation between the data, the whole data set needs to be searched to find K similar records for each missing record, so its computational complexity will be high and its accuracy will be reduced. In addition, it remains to be solved how to find the most suitable similarity independent way for different data sets and how to set the optimal K value. This is also the main shortcoming of THE KNN filling method.

3.2 Expected Maximization filling method (EM)

Expected maximization filling method (EM), proposed by Dempster et al. In 1977, is a method for maximum likelihood estimation of probabilistic parameter models with hidden variables [4], which is rated as one of the top ten data mining algorithms. There are two types of so-called hidden variables: (1) data is not missing. Due to the complex distribution of variables, the corresponding likelihood function cannot be described by the expression that can be calculated. For example, the simplest mixed Gaussian model must be decomposed into several normal distributions and the implicit variable "category" must be introduced to transform the originally complex maximum likelihood estimation process into a simple version, so as to calculate the parameters of the probability model. (2) Lack of data. For example, missing data during data set collection is the case studied in this article.

The EM algorithm can be generally interpreted as estimating the value of the hidden variable based on the assumption that the parameter value of the model is known, then using the estimated value of the hidden variable to modify the model parameters, and in this way, the iterative estimation is continued until the convergence. The biggest advantage of EM algorithm lies in its simplicity and stability. Each iteration can ensure that the maximum likelihood function value increases and converges to a local optimal value. The more close the sample distribution is to the population distribution, the more accurate the estimation results will be. However, EM algorithm also has many disadvantages that cannot be ignored. First, the algorithm can only ensure the convergence to a stable point, but cannot guarantee the global optimal stability point. Second, the calculation process of EM requires the given likelihood function, that is, the distribution function of a given variable under different parameter values to be estimated, which is a very harsh condition. Therefore, in practical application, it is often assumed that the sample data obey multivariate normal distribution. When there is a big difference between the real distribution of data and the assumed conditions, the filling effect will be greatly affected. Thirdly, if there are many missing data in the data set, the convergence rate is slow.

3.3 Data filling algorithm based on rough set

In practical application, an information system is usually a two-dimensional information table, in which the rows correspond to objects in the field of argument, the columns correspond to attributes, and the contents of the table are function values. An important problem in rough set theory is the problem of computational attribute reduction. Through attribute reduction, redundant and useless components in data can be removed, and the relationship between each attribute can be revealed. The principle of attribute reduction is actually very simple. All attributes can be divided into two types: one is a conditional attribute, and the other is a decision attribute. We will set the decision attribute in the last column of the data column.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202010_6(10).0041

The discernibility matrix is a matrix introduced by Skowron, which plays a major role in rough sets. The content of matrix element is attribute set, which indicates that two different decision attribute objects have different values on this attribute set. The discernibility matrix can be divided into heterogeneous discernibility matrix and homogeneous discernibility matrix. Firstly, the importance of heterogeneous discernibility matrix and each conditional attribute is calculated, and then the homogeneous discernibility matrix is calculated. It is found that the sum of the importance of each attribute in the attribute set is the minimum value from the homogeneous discernibility matrix. Traverse the entire matrix. If some attribute values contain null values, change the null value to the corresponding attribute value of another object. Repeat this process until the matrix contains no null values. The algorithm takes advantage of the similarity of attribute values of similar objects.

3.4 Compressed sensing

Compressed sensing (Compressed sensing), also known as compression sampling (Compressive from), Sparse sampling (Sparse from), Compressed sensing.As a new sampling theory, it develops the sparse characteristics of signals, obtains the discrete samples of signals with random sampling under the condition that the sampling rate is much lower than Nyquist sampling, and then perfectly reconstructs the signals through the nonlinear reconstruction algorithm.In recent years, compressed sensing technology has made great contributions in the field of signal processing, and the matrix filling technology derived from it (also known as matrix completion technology) has also attracted much attention from researchers. Matrix filling technology refers to the method of filling element values at the positions with incomplete values of the matrix [5]. It has been widely applied in many disciplines and engineering fields, such as machine learning, computer vision, recommendation system and social network link prediction. The popularity of matrix filling technology in the recommendation system field is mainly due to the Netflix Prize Million-dollar competition launched by Netflix in 2006. In the matrix filling technique, two or more low-dimensional matrices are used to approximate a high-dimensional matrix, which is the filling model based on matrix decomposition, and then the inner product is used to approximate the original matrix.

The main idea of matrix filling technology is to fill a matrix with missing elements by processing some of its elements. Therefore, the compressed perception theory of one-dimensional vector space is extended to the matrix filling theory of two-dimensional matrix space, and this extension has great theoretical and practical significance. Various matrix filling models have emerged in recent years. From the perspective of solving the minimization problem, the models can be divided into the matrix filling model based on kernel norm relaxation, the matrix filling model based on matrix decomposition and other optimization models.

3.4.1 Matrix filling model based on kernel norm relaxation

Matrix filling model based on kernel norm relaxation is an important relaxation technique in which the rank function of the matrix is convex and relaxed to the kernel norm, and then solved by alternating direction multiplier method or accelerated nearest neighbor gradient algorithm. The matrix filling model based on the kernel norm relaxation plays an important role in the theoretical research, but due to the continuous expansion of the data scale, it is unable to cope with the large-scale recommendation system.

3.4.2 Matrix filling model based on matrix decomposition

Matrix factorization (MF) algorithm dissolves the Matrix into several matrices, and then USES the product result to approximate the original high-dimensional Matrix, thus improving the execution efficiency of the algorithm and cleverly avoiding the Matrix singular value decomposition with high complexity. This method not only reduces the time complexity of the calculation but also improves the high sparsity of the original scoring matrix to some extent. Matrix decomposition is a powerful tool for data mining, and it can be used in different business applications, such as computer vision, pattern recognition, recommendation system and so on.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202010_6(10).0041

4. Conclusion

In this paper, through the establishment of the corresponding data model, using the specific experimental analysis of each missing data processing methods to verify the simple answer. However, there are still many problems in the actual application process of these actual missing data processing methods. Data filling should be treated with caution. It is unreasonable to merely increase the number of filling in order to achieve better filling effect. As Rubin, who pioneered the systematic approach to filling, says: "The idea of filling is very tempting and very dangerous. The temptation is to get people into such a state of euphoria that they end up so obsessed with the padded data set that they tend to ignore the bias, and that's the danger", so improve the accuracy of the data into some other processing method, which will be missing data processing method for continuous improvement and perfection, so as to effectively promote the progress and development of data mining technology in our country, and aiming at all kinds of practical problems, to pay attention to distinguish the essence of the problem, a reasonable and appropriate use of processing method is the key to solve the actual problem.

References

[1] Ian H.Witten Eibe Frank, Data Mining and Machine Learning Technology, Beijing, Mechanism Industry Press, 2006.

[2] LI Hui-min, WANG Pu, FANG Li-ying, LIU Jing-wei College of Electronic Information & Control Engineering, Beijing University of Technology, Beijing 100124. An Algorithm Based on Time Series Similarity Measurement for Missing Data Filling[C].

[3] Cover T, Hart P. Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 1967, 13(1):21-27.

[4] Wu C F J. On the Convergence Properties of the EM Algorithm. Annals ofStatistics, 1983, 11(1): 95-103.

[5] Keshavan R H, Oh S, Montanari A. Matrix completion from a few entries. IEEE International Conference on Symposium on Information Theory. IEEE Press. 2009: 324-328P.