Financial Engineering and Risk Management (2021) 4: 26-32
Clausius Scientific Press, Canada

DOI: 10.23977/ferm.2021.040606
ISSN 2523-2576

Research and Application of Statistical Method of Data Reduction Based on Empirical Distribution

Jie SUN

Jinan Jiyang District Government Service Center, Jinan, Shandong 251400, China

Keywords: Empirical distribution, Data reduction, Statistics, Cluster

Abstract: Data reduction is used to obtain the reduced representation of the data set, which is smaller than the original data, but still maintains the integrity of the original data approximately. Mining on the reduced data set will be more effective and produce the same or almost the same analysis results. A continuous multivariate coupled distribution estimation algorithm with arbitrary distribution is proposed. The distribution is estimated from samples by empirical distribution function, and new individuals are generated by sampling. Secondly, the idea of clustering is introduced into data reduction, and a time dimension reduction method based on clustering is formed. The basic idea of this method is to cluster the time dimension of time series data. In order to verify the feasibility of the two new methods proposed in this paper, a set of simulation experiments are designed in this paper, and representative data are used for data reduction respectively. Experiments show that the two data reduction methods proposed in this paper can not only effectively reduce the amount of data and achieve the purpose of data reduction, but also improve the classification accuracy and have strong practicability.

1. Introduction

Data reduction refers to simplifying the original data set while providing the same analysis results. Its main purpose is to improve the efficiency of data processing, development and utilization, and at the same time improve the accuracy and simplify the description. Data reduction strategies include dimension reduction, quantity reduction and data compression [1]. At present, the research on the evaluation index of data reduction effect in the industry mainly focuses on two aspects: the data volume and the difference degree of information before and after data reduction. At present, the indicators for measuring the difference of information amount are mainly used in supervised machine learning models [2]. At the same time, the index calculation methods are mostly suitable for data sets with decision attributes and discrete conditional attributes, while it is difficult to calculate the continuous attributes.

From a brand-new point of view, the empirical distribution function is used to establish probability model and sample, without making the assumption that random variables obey a specific distribution. After obtaining the empirical distribution function obeyed by the sample, the value range of each component of the sample is calculated by using the inverse transformation method, which is regarded as the constraint of multivariate correlation [3-4]. Then a sample can be obtained by sampling under this constraint, and the next generation population can be obtained by repeating this process.

2. Types of Data Reduction

2.1 Feature Reduction

Feature reduction refers to finding attributes related to mining tasks from data sets containing hundreds of attributes, thus improving the quality of mining patterns and reducing the time and space cost of mining.

At present, the method of attribute subset selection is changing with each passing day, but it is based on the evaluation criteria of a subset to measure whether the attribute subset is the most representative and can best represent all attribute sets to divide samples, thus achieving the best classification effect [5]. Commonly used evaluation criteria are usually expressed in the form of functions, which mainly include evaluation functions based on information entropy, distance, relevance, consistency and classification error rate.

2.2 Sample Reduction

Is to select a representative subset of samples from the data set as a new sample. When selecting the subset size, it is necessary to consider the calculation cost, storage requirements, accuracy of estimators and other factors related to algorithm and data characteristics.

Sample reduction is actually the most complex task in data reduction, because as data mining workers, they often do not participate in the actual data collection process. The so-called mining task can be regarded as secondary data analysis, and the mining process has no connection with the optimal method of collecting data and selecting the sample set of initial data [6]. Usually, larger sample size increases the probability of representativeness of samples. Using smaller samples may lose patterns or detect wrong patterns. However, large samples usually offset many benefits brought by sampling.

2.3 Time Dimension Reduction

It adopts the feature discretization technology, which can reduce the number of discrete values of known features and make them become a few intervals, and each interval is mapped to a discrete symbol. Its advantages are simplified data description and easy understanding of data and final mining results.

Time dimension reduction is to use an alternative and smaller data representation to reduce the amount of data. The parameter method uses a model to estimate data, and only needs to store parameters (or outliers) instead of actual data.

3. Statistical Method of Data Reduction Based on Empirical Distribution

3.1 Empirical Distribution Function Sampling

Given a random independent sample $X_1, X_2, \cdots, X_l$, its empirical distribution function is:

$$F_l(X) = \frac{1}{l}\sum_{i=1}^{l}\theta(X - X_i) \quad (1)$$

The θ step function is defined as:

$$\theta(v)=\begin{cases}1, & \textit{If all vectors of vector v are positive} \\ 0, & \textit{other}\end{cases} \quad (2)$$

Empirical distribution function $F_l(X)$ is the approximation of actual distribution function $F(X)$, i.e. when $l \to \infty, F_l(X)=F(X)$.

The main idea of the distribution estimation algorithm is to generate the individuals of the next generation population that obey their probability distribution according to the preferred individuals of the previous generation population [7]. Empirical distribution function can represent the probability distribution that a population prefers to obey. If the next generation population can be generated by sampling from this empirical distribution function, it is feasible to use it as the probability model and sampling process of distribution estimation algorithm.

Empirical distribution function can be easily obtained from samples. However, the empirical distribution function is obtained by superposition of several step functions, so the inverse transformation cannot be carried out, so the inverse transformation method cannot be directly used for sampling, but the idea of inverse transformation method can be considered for approximate sampling.

The empirical distribution function constructed from sample $X_1, X_2, \cdots, X_l$ cannot obtain a definite sample X' by random uniform sampling and then inverse transformation of the distribution function, but the formula (1) is simply transformed into the following formula:

$$lF_l(X)=\sum_{i=1}^{l}\theta(X-X_i) \quad (3)$$

X_i represents the individual vector in the population preference set, and $i \in [1, popSize \times selRate]$, each individual has n components, and n is a positive integer. $popSize$ is the population size, and $selRate$ is the selection probability.

Sampling is performed by inverse transformation method, $u \sim U(0,1)$ is randomly selected, and the left end of the equal sign in formula (3) becomes lu. According to the characteristics of formula (2), the value of this step function can only take 0,1.

3.2 Statistical Distribution Characteristic Difference Degree

The difference of data distribution characteristics mainly refers to the difference of distribution types, and the difference of distribution positions can be reflected by the central trend characteristics. Therefore, the calculation of the difference of data distribution characteristics in this paper mainly focuses on the distribution types. The measurement value of the difference degree P_1 is calculated by the following formula:

$$P_1=\frac{\sum_{i=1}^{n}I(i)}{n} \quad (4)$$

In formula (4), n is the total number of variables in the reduced data set, and $I(i)$ is the indicator function. When the distribution type of a variable (i.e. attribute) in the data set changes, the value is 1, otherwise it is 0.

Table 1 Shows the Candidate Distribution Types of Different Variable Types.

Table 1 Candidate Distribution Types of Different Variable Types

Data type	Variable type	Candidate distribution type
Structural data	Discrete variable	Binomial distribution and Poisson distribution
	Continuous variable	Normal distribution and exponential distribution

The measurement is calculated as follows:

(1)For the current variable, the K-S test is repeated for all optional distribution types in the original data, and the closest distribution is selected as the distribution type of the variable.

(2)K-S test is carried out on the same reduced variables. If the reduced data variables belong to the same type as the original data variables, the value of $I(i)$ is 1; otherwise, it is 0.

(3)When all the n variables in Angelica sinensis data set are checked, the difference degree P_1 of data distribution degree can be calculated.

3.3 Discretization Method Based on Likelihood Ratio Hypothesis Test

Average information can describe the statistical dependence between two variables, and can be used as an information measure of statistical dependence degree, which is expressed by probability as follows:

$$H(X:Y)=\sum_{i}\sum_{j}p(x_i,y_j)\log\frac{p(x_i,y_j)}{p(x_i)p(y_j)} \tag{5}$$

Type shows that if two random variables are independent, there are:

$$H(X:Y)=0 \tag{6}$$

In the contingency table of category attributes and conditional attributes, if the average mutual information is used to measure the degree of dependence between adjacent intervals and category attributes, formula (5) can be converted into:

$$H(d:I)=\sum_{i=1}^{m}\sum_{j=1}^{I}\frac{n_{ij}}{N}\ln\frac{Nn_{ij}}{n_i n_j} \tag{7}$$

In which d represents the category attribute and I represents the initial interval. By setting formula (7) to 0, it can be judged that the category attribute is independent of the initial interval. However, it should be noted that completely independent cases are rare in reality, so it is necessary to use hypothesis testing to determine to what extent the average mutual information is small, so that two adjacent intervals and category attributes can be considered to be independent.

The original hypothesis of likelihood ratio hypothesis test is H_0: two adjacent intervals are independent. Structural statistics:

$$T^2=-2N\Lambda=2\sum_{i=1}^{m}\sum_{j=1}^{l}n_{ij}\ln\frac{Nn_{ij}}{n_i n_j} \tag{8}$$

It asymptotically obeys χ^2 distribution with $(m-1)(l-1)$ degrees of freedom. T_α^2 is the critical value of χ^2 distribution at significance level α. When $T^2<T_\alpha^2$ is used, the original hypothesis is accepted, indicating that the probability distributions of two adjacent intervals and category attributes are independent, so the two intervals can be merged.

4. Application of Statistical Method of Data Reduction

4.1 Time Dimension Reduction Method Based on Clustering

Cluster analysis is an important technology and one of the main tasks of data mining. It is widely used, and many fields will involve the application and research of cluster analysis methods. For example, in the commercial field, clustering can help market analysts distinguish different consumer groups from the consumer database, sum up the consumption patterns of each type of consumer, discover different types of customer groups, classify and depict the characteristics of customer groups according to their purchasing habits, so that enterprises can better serve customers and improve economic benefits [8].

In this paper, the information entropy method is applied to the information measurement of time series sampling points, which is used to measure the amount of information available at each time point. Every time series is composed of many time sampling points. In a time series composed of many samples, if they are sampled at the same time point, then each time sampling point can also be regarded as a series, which is composed of the values of different samples at the time sampling point. We call it a time point series. The relationship between the information entropy value of time point series and sampling points is shown in Figure 1:

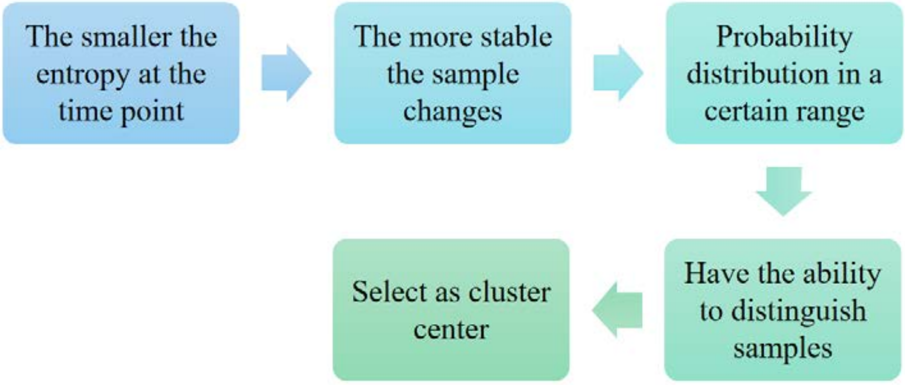
The smaller the entropy at the time point
The more stable the sample changes
Probability distribution in a certain range
Have the ability to distinguish samples
Select as cluster center

Fig.1 The Relationship between Entropy and Sampling Points

The core idea of time dimension reduction of adaptive clustering based on information entropy is: transpose the time series, so that each row represents time sampling points and each column represents samples. Calculate the entropy value of the time sampling point series, select the time point with lower entropy value as the clustering center, then cluster, and use the clustering center finally generated by clustering as the value on the new sampling point, thus achieving the purpose of data reduction.

Let the sample of time series data be $X_i, i = 1,2,\cdots,N$. For each sample $X_i = \{S_1, S_2, \cdots, S_j, \cdots\}, j = 1,2,\cdots,M$ is the sampling point of the time series, that is, the time point. S_j is the sampling value of j time points in the time series X_i. That is, the whole time series is equivalent to a matrix set with N rows and M columns.

Input: time series X of single attribute; Number of clusters K.

Output: reduced time series Y.

Algorithm description:

Transpose the data set matrix, which is a matrix set of M rows and N columns, $SUM = \{S_1^1, S_2^2, \cdots, S_j^i, \cdots, S_M^M\}$. at this time, each row table is every time point, and each column represents every time sample.

Calculate the information entropy of each line, and sort according to the entropy value from small to large.

The first K time point sequences with information entropy values are regarded as the initial center C_k of clustering.

For each time point sequence $T_j = \{S_j^1, S_j^2, \cdots, S_j^N\}$, find the center point nearest to it by the following formula.

$$k = \arg\min_{k\notin\{1,\cdots,k\}} d(c_k, X_i), k = 1,2,\cdots,K \quad (9)$$

The mean value of data points in each cluster is calculated, and the mean value vector becomes the new center of the cluster.

Repeat the above two steps until no or few time point sequences are assigned to different clusters.

The time complexity of the algorithm is $O(M + MKt)$, where M is the time dimension in the data set, that is, the number of sampling points in the time series, K is the number of expected clusters, and t is the number of iterations.

4.2 Experimental Analysis

To verify the clustering time dimension reduction method proposed in this paper, Mallat data is used in this experiment. Among them, Mallat data contains 2401 records, each record has 1024 time sampling points, and its classification task is to judge the degree of process measurement (according to the degree, it is divided into 8 grades, which are represented by numbers of 1 ~ 8 respectively).

Mallat data are reduced by the basic TDRBC (time dimensional reduction based on cluster) method and the adaptive TDRBC method respectively, and then classified. the accuracy and root mean square error obtained by experiments are shown in fig. 2, where the values in brackets in the table are root mean square error.

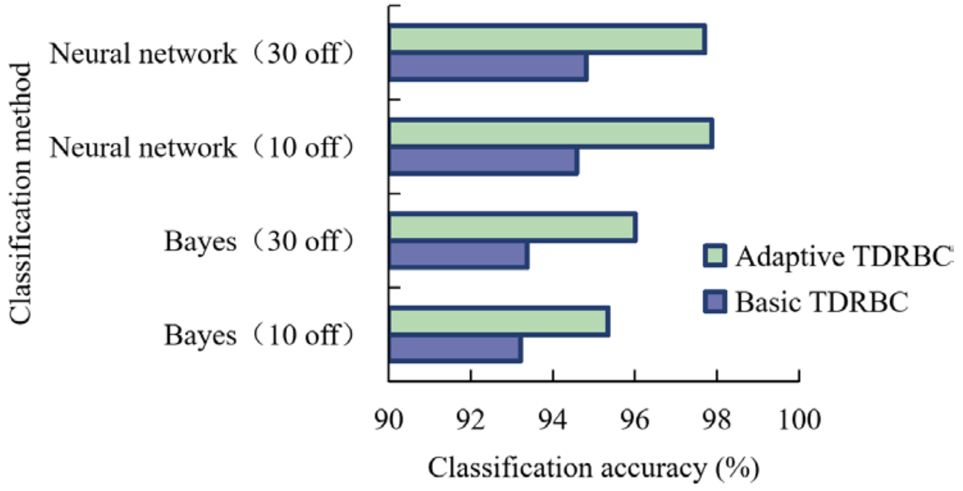

Fig.2 Comparison of Classification Accuracy of Mallat Data

In addition to the classification methods listed in figure 2, the data reduced by using the adaptive TDRBC method is also suitable for other classification methods. each group of data can find a classification method with a classification accuracy of over 95%, and the best result can be as high as 97.88%. the reason why this experiment adopts a classification method different from the previous one is mainly to show that these two methods are extensive in the selection of classification algorithms.

5. Conclusion

Data reduction is the premise of successful and efficient completion of data mining technology, and plays a key role in the process of data mining. Aiming at the deficiency that most juice estimation algorithms in continuous domain use Gaussian probability model, this paper proposes a juice estimation algorithm in continuous domain with multivariate correlation without false specific distribution. According to the characteristics of time series and the challenges of classified data mining, this paper introduces the idea of clustering into data reduction, and forms a time dimension reduction method based on clustering. Finally, the paper proves by experiments that these two reduction methods can not only reduce the dimension of time sampling points, but also improve the classification accuracy.

References

[1] Yuvaraja T, Ramya K. Statistical data analysis for harmonic reduction in 3Ø -fragmented source using novel fuzzy digital logic switching technique. Circuit World, vo. 45, no. 3, pp. 148-155, 2019.

[2] Chen M S, Hwang C P, Ho T Y, et al. Driving behaviors analysis based on feature selection and statistical approach: a preliminary study. Journal of supercomputing, vol. 75, no. 4, pp. 2007-2026, 2019.

[3] Chow C, Andrasik R, Fischer B, et al. Application of statistical techniques to proportional loss data: Evaluating the predictive accuracy of physical vulnerability to hazardous hydro-meteorological events. Journal of Environmental Management, 2019, no. 15, pp. 85-100, 246.

[4] Yang M, Shahramian S, Shakiba H, et al. Statistical BER Analysis of Wireline Links With Non-Binary Linear Block Codes Subject to DFE Error Propagation. Circuits and Systems I: Regular Papers, IEEE Transactions on, no. 99, pp. 1-14, 2019.

[5] Martin T, Drissen L, Prunet S. Data reduction and calibration accuracy of the imaging Fourier transform spectrometer SITELLE. Monthly Notices of the Royal Astronomical Society, no. 4, pp. 4, 2021.

[6] Yi xueyi. research on statistical method system and its application. no. 2016-1, pp. 18-20, 2021.

[7] Gao yuhao, Zhao Yang, cheng yingjin. analysis method of da/dn-δ k curve based on probability and statistics theory . materials development and application, vol. 34, no. 06, pp. 25-32, 2019.

[8] Liu Rui. Research on innovation of big data and statistical methods . Statistics and Consulting, no. 2, pp. 22-25, 2020.