Acta Technica 62 (2017), No. 5A, 357–368

Data cleaning technology in agricultural monitoring information

Rui Zhao[2], Pingzeng Liu[2,3], Chao Zhang[2], Maoling Yan[2], Weijie Chen[2], Bangguo Li[2]

Abstract. Agricultural Information Monitoring requires not only clear and accurate data, information, and knowledge related to agricultural production and processing, but also insight into their efficient management. In order to improve data quality, this paper focuses on data cleaning methods, using repeated data cleaning methods based on hash and content recognition, missing value cleaning methods ignoring based on incomplete data and filling technology and noise data elimination method based on sub box and regression, and real-time monitoring data are processed effectively. The cleaning results show that the above methods can effectively deal with duplicated values, outliers and missing values in farmland data, the repetition value is reduced from 3 to 0, the outliers reduced from 4 to 0, and the missing values decreased from 1258 to 6, and the total number increased from 6028 to 7281.Through the research on the cleaning technology of agricultural monitoring data and the specific application, providing scientific and reliable data source for the next mining analysis, which is beneficial to the improvement of farmland management level.

Key words. Agricultural information monitoring, data acquisition, pretreatment, data cleaning.

1. Introduction

Agricultural monitoring provides data from agricultural production activities, including the systematic monitoring of crop growth. However, varying operational standards, user proficiency, and instrument quality during collection can induce errors in acquired information. In addition, weather and geography can introduce uncertainty in data acquisition and modify the stability of data sources. So, before using data for mining and analysis, it is necessary to clean the data. Data cleaning can improve the integrity of incomplete data, making erroneous data correct

[1] Acknowledgment - The project was funded by the Independent Innovation and Achievement Transformation of Shandong Province under Grant No. 2014ZZCX07106.

[2] Workshop 1 - Department of Information Science and Engineering, Shandong Agricultural University, 370900, Taian

[3] Corresponding author: Pingzeng Liu; e-mail: lpz8565@126.com

while removing redundant data. This reduces cost while maximally improving the effectiveness and comprehensibility of the data [1].

At present, in view of methods and applications of data cleaning, scholars at home and abroad have carried out relevant research and made some progress. Commonly used repeat records cleaning algorithms are: a series of improved algorithms based on sorting [2]. The complexity of them is low, but the accuracy depends largely on the sort key. If the key words are not selected, they may miss many duplicate records. Commonly used missing records cleaning methods are: mean value, hierarchical clustering, maximum likelihood estimation and Bias estimation [3-4]. In comparison, mean interpolation has great interference on samples, and the deviation between parameter estimation and real value is very large. Therefore, researchers prefer the latter two methods. Commonly used abnormal records cleaning methods are: the statistical algorithm, this method can select the sample randomly, and accelerate the detection speed, but the distribution of data has great influence on the recognition process. Method based on association rules are not easily affected by the data distribution, but it is difficult to detect some isolated abnormal points. Clustering method is sensitive to outliers, but it is difficulty to obtain conclusion in the larger sample[5]. Based on the study of the method, some scholars also put forward the framework of data cleaning model to provide a convenient cleaning process. In the application, many industries with high accuracy for customer data are cleaning there own data, such as banking, insurance and securities [6]. With the development of large data, data quality has become a key issue for all areas of information resources, and data cleaning is widely used in medicine, agriculture, e-government, transportation, water conservancy and other industries[7].

Using data from a Bohai granary project, this paper focuses on data cleaning techniques during data acquisition. Through the study of such data cleaning approaches, data quality can be improved, which help improve the effectiveness and accuracy of the data mining process, and promote and improve the overall accuracy of crop monitoring services.

2. Materials and methods

2.1. Research area overview

The "Bohai granary" is the essence of "the barn around Bohai low plain" .It is located in the eastern part of the North China Plain and made up by rivers. Including Shandong, Hebei, Tianjin, Liaoning, and it is the most important grain, fruit and vegetable producing areas in China.For a long time, the region has been affected by natural and economic social factors, there is a vast area of low yield area and a large area of saline alkali wasteland.The Bohai granary project has set up more than 140 networking sites in Binzhou, Dezhou, Dongying and other cities and counties. Through the scientific distribution of meteorological, soil, water and other types of sensors, using wireless sensor networks automatically transmitting data.The intelligent perception system of crop growth process environment information is constructed, and all kinds of data are collected and collected in real time, providing all-weather and three-dimensional data support for subsequent data analysis, monitoring, early warning and decision-making services. However, deficiencies in data caused by exceptions and redundancies require an urgent resolution involving data processing methods.

2.2. Data acquisition

The main crop varieties of shandong project area of "Bohai granary"are wheat and corn.Prioritized monitoring information includes crop distribution, area, yield, growth of information, soil moisture, soil nutrients, tillage depth, and crop pests, which need to be monitored rapidly, accurately, and continuously. Field information collection and monitoring mainly involves four categories: cropland production information, such as the planting patterns, land use types, growth, and coverage; soil information, such as temperature, moisture, fertility, nutrient, and pH; environmental data, such as temperature, humidity, light intensity, rainfall, and crop pests; and biochemical information related to crop growth, such as plant height, leaf area, and tiller number.

2.3. Implementation method

2.3.1. The principle of data cleaning The principle of data cleaning is to use data mining techniques to transform dirty data into the clean data needed for data mining, according to designed cleaning rules shown in Figure 1.

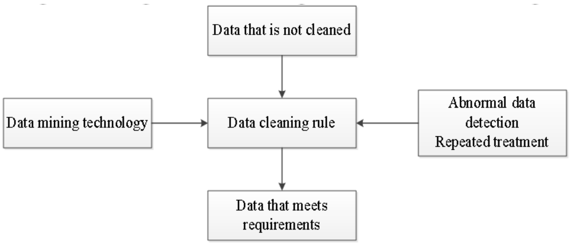
Data that is not cleaned
Data mining technology
Data cleaning rule
Abnormal data detection Repeated treatment
Data that meets requirements

Fig. 1. The principle of data cleaning

2.3.2. Cleaning of repeated records Constructing data warehouses requires the import of large amounts of data from a variety of data sources. Ideally, each real entity is associated with one corresponding record in a database or data warehouse. But because actual data may be input incorrectly, a number of redundant records may result, resulting in reduced data query efficiency and leading to erroneous data mining results. Therefore, the removal of duplicate records is necessary to improve the accuracy and speed of subsequent mining.

In order to detect and eliminate duplicate records, a determination of whether two records are repeated must be achieved. For example, if data are organized by collected time, data recorded as being collected in the same hour means that they may have been repeatedly collected.

The primary methods of data cleaning are to use a recognition based hash or using recognition-based content, which are discussed in turn.

A recognition-based hash uses a hash value to determine whether data are duplicated by generating a hash for each new data block. If the hash of a data block matches with a hash index on a storage device, then the data block is a duplicate. In order to quickly identify whether a data block has been stored, the hash index is stored in memory, so as the number of data blocks increases, the index grows. However, as the index grows more than available capacity, performance will decline rapidly, because hashes would get stored to disk, and a disk-based search is much slower than a memory-based search. As a result, most hash-based systems are independent of disk at present, and able to effectively balance internal memory and disk space needed to store data, so that the hash table will never become too large such that it decreases performance.

Recognition-based content uses metadata in the embedded file system to identify files, compares this metadata with other versions in the repository byte-by-byte, finding the difference between versions, and creates an incremental file for these different data versions. This approach avoids hash collisions, but requires application devices that support the functionality so that devices can extract metadata.

2.3.3. Cleaning missing data algorithm Cleaning missing data is a key problem in the field of data cleaning and preprocessing. Incomplete and inaccurate data affect extraction patterns and rule accuracy, leading to errors in later data mining. As a result, the decision support system may support inaccurate decision-making and information services.

Currently, many methods for cleaning missing data are available, which can be roughly divided into two categories: methods that ignore incomplete data and those that pad. The first method is the simplest to implement, often deleting incomplete attributes. The second method requires a filling algorithm, commonly used to bridge incomplete data by analyzing complete data and selecting appropriate filling values.

Deleting attributes or instances is a common method for dealing with incomplete data, and many statistical tools use this method as a default. Although the method is efficient, however, its disadvantages include removal of information and inducing bias on the data set. The deletion of attributes or instances, even when containing incomplete information, will shrink the data set so that the remaining data may lose meaning, decreasing trust. Removal of data also biases the data set, because when data processed by this method are used for classification or clustering in data mining, the resulting model will be skewed, leading to misleading results.

Statistical methods that populate missing values can be used to clean data sets. Such methods obtain appropriate statistical information from the data set through some analysis, and use it to generate values that fill the gaps. In this category, the simplest and most common method is to fill missing values with the arithmetic means of the full data. Note that applying this mean-padding method affects the correlation between missing data and other data. Moreover, if a large data set is replaced by this mean-filling method, the frequency distribution of the variables may be misleading.

Artificial neural networks (ANN) may also be used to fill the missing values. ANNs were first proposed by psychologists and neuroscientists seeking to develop and test neural simulations. Neural networks require training, and thus are more suitable for applications that are able to accommodate sufficient training time [8]. They require a large number of parameters that are often determined by experience, such as network topology.

The defining characteristics of a neural network are several. A neural network is difficult to explain generally. Neural networks can deal with attribute redundancy by utilizing automatically-learned weights generated during their training stage, with redundant attributes weighted smaller. Neural network learning can be excessive, so it is necessary to employ appropriate methods such as test sets and cross-validation. This is primarily because neural networks are very flexible, allowing significant parameter variability. Training neural networks may take a considerable amount of time unless the problem is very simple, and preparing data needed to build a neural network can be very time-consuming.

The steps needed to build a neural network include 1) Determine the number of nodes in the input layer;2) Determine the number of nodes in the output layer; 3) Select a network topology; 4) Randomly initialize weights; 5) Complete the training sample, and remove missing values if found.

Clustering can also be used to populate missing values. Clustering techniques have been widely used in statistics, machine learning, pattern recognition, data mining and knowledge discovery [9]. There are many kinds of clustering methods, but their main idea is to gather similar instances into a class, with differing instances clustered into different classes. Clustering groups data into multiple classes or clusters, wherein similar data objects or records are placed in the same cluster, with significant differences between clusters.

The K-means algorithm is a clustering method based on partitioning. It randomly selects K initial clustering centers according to a chosen number of final categories. Clustering results can be obtained by minimizing an objective function. The basic K-means algorithm steps are 1) Select K initial centers; 2) Classify each object into the nearest class, forming K clusters; 3) Recalculate the center of each cluster; 4) Iterate until the cluster center no longer changes.

After the clustering results are obtained, missing values can be calculated from them. A specific description is as follows:

1. The data set D is divided into two subsets of data, with records in D1 and D2. All records in D1 are complete, and attributes do not contain missing values. The records in D2 are the missing records.

2. The data in D1 are classified using the K-means algorithm, arbitrarily selecting K objects with an initial cluster center of gravity, computing the distance from the center of gravity to each object in the dataset, and assigning each object to the nearest cluster. Then, the average of the clusters is updated, that is, the average of the objects in each cluster is computed until the objective function is minimized.

3. The records from D2 are sequentially extracting and the similarity between the record and any class in the Kth class of D1 is computed. The maximum similarity is selected, and the record is marked as Ci (i=1,2,...,k) until the data subset is empty.

4. Using the records assigned to D2, the missing values of the records are processed as follows. If missing value are numeric data, they are equal to the mean of the corresponding attribute value. If the missing values are discrete data, they are equal to the most frequent attribute value.

Figure 2 illustrates the filling of missing values by using the K-means algorithm:

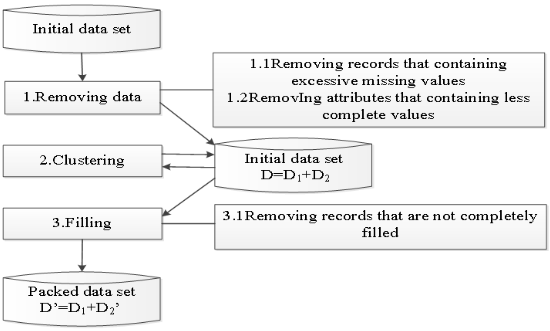
Initial data set
1.Removing data
1.1Removing records that containing excessive missing values
1.2RemovIng attributes that containing less complete values
2.Clustering
Initial data set D=D1+D2
3.Filling
3.1Removing records that are not completely filled
Packed data set D'=D1+D2'

Fig. 2. Figure 2 Flow chart of missing value filling algorithm based on K-means.

2.3.4. Noisy data elimination Noisy data have incorrect attribute values, including errors or deviations due to outliers, and can occur for several reasons. Data collection devices may be out of order; causing errors in data transmission. Because the presence of noise causes data to no longer be specified within a domain, it can affect subsequent mining results. A commonly used method of eliminating noisy data is the compartment method. The compartment method deals with the data that need to reference surrounding instances. Data are divided into boxes, and different binning techniques are used to smooth the values. Common techniques include a sliding average, which averages all values in the box, and then uses the average to replace all the data in the box, and a boundary sliding average, where the maximum and minimum values in the box are considered as "box boundaries", and each value in the box is replaced by the nearest box boundary value.

Regression is also a method of smoothing noisy data that can be fitted with a regression function. Linear regression involves finding the best-fit line using two attributes or variables, so that one property can be used to predict the other. Multiple linear regression extends linear regression, involving more than two attributes, with data fitted to a multidimensional surface [10-11]. The regression function is used to predict the trend of the variables, and then to modify the noisy data so that they are in the regressed line or curve as much as possible.

3. Specific implementation

The Bohai granary crop monitoring subsystem offers an opportunity to use technical experience and highlight practical problems from a domestic and foreign agricultural condition monitoring system, addressing improvements according to national need through research into data cleaning approaches.

3.1. Process analysis

First, the vertical and horizontal features of automatically collected data were analyzed to determine the change rules of data from this period and find missing and abnormal data. Data cleaning techniques were then applied to the data collected by this system. Through a comparison of different available methods, a series of improved solutions for each parameter were proposed to filter or repair duplicated, incomplete and erroneous data. In the end, this conversion of dirty data into clean data met quality or application requirements, thereby improving their quality.

Data cleaning includes finding and remedying missing values, removing duplicate records, and processing noisy data. The module processing flow chart is shown in Figure 3.

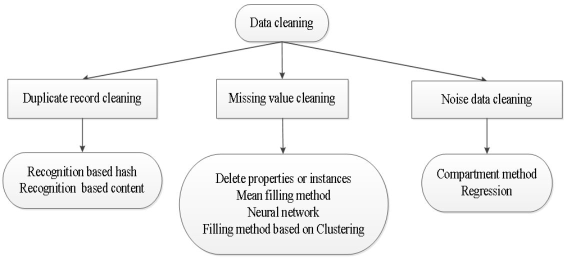

Fig. 3. Flow chart of data cleaning

Key technologies were applied to demonstrate the data visualization of the Bohai granary project, and demonstration and test work was carried out. First, data was acquired to test whether the methods of data collection have achieved a standard. Then, data cleaning was performed and data were visualized. After completing the research described above and verifying the feasibility of each technology, this paper summarizes relevant innovations and insufficiencies, and points out relevant research where applicable.

3.2. Cleaning repeated data

The basic idea of cleaning repeat values is to group all values by a certain time format in a database.If there is a duplicate record, the data with the smallest ID value is retained and the rest are deleted.The core database operation statement that deletes the repeat value is as follows:

delete from table where date_format(date,'%Y-%m-%d %H') in (select * from (select date_format(date,'%Y-%m-%d %H') from table group by date_format(date,'%Y-%m-%d %H') having count (date_format(date,'%Y-%m-%d %H')) > 1) as b) and id not in (select * from (select min(id) from table group by date_format(date,'%Y-%m-%d %H') having count(date_format(date,'%Y-%m-%d %H')) > 1) as c).

3.3. Cleaning missing data

After source data were read in, they were transformed into a matrix after processing. After cleaning had been completed, the data were also returned to the matrix format. When searching for duplicate records, the basic idea was to sort the data sets, and then compare the number of hours between adjacent records. If a duplicate record was found, the minimum of the ID record was kept and the rest were deleted. When searching missing data, an array was defined whose times were incremented by hours, so that time arrays could be compared one-by-one. If one hour existed, the mark was 1, and if it lost, the mark was 0. The core codes for querying missing values are as follows:

```
for(i=0;i<tt;i++){
try{if(time1[i].equals(
date1[k].toLocaleString())){k++;full[i]=1;
DataBean bean=pottingData(time1[i],k,1);
DataBeans.add(bean);
}else{full[i]=0;
DataBean bean=pottingData(time1[i],k,0);
DataBeans.add(bean);}}catch(Exception e){}}
```

Missing data were divided into single point and multi-point or continuous deletion classes, and missing data were filled with the data set mean. When a single point of missing data was filled, the two points before and after the missing point were obtained, and then their average was assigned to the missing point. When continuously missing data were filled, two cases were addressed. If continuous point deletions occurred in one day, values of two days before and after the point could be used, with the average value assigned to the missing points. Alternatively, if values were missing for several days, reference values from the previous year were used to fill the data. The core codes for filling the single point missing value are as follows:

```
if(i>=1&&i!=DataBeans.size()-1
&&DataBeans.get(i-1).getStatus()==1
&&DataBeans.get(i+1).getStatus()==1
&&DataBeans.get(i).getStatus()==0){
avg(DataBeans.get(i).getK(),DataBeans.get(i).getDate(),soil_con,soil_humi,soil_temp,soil_hu
```

3.4. Cleaning noise data

Abnormal data were queried according to the prescribed conditions, and modified according to the compartment method. Taking air temperature as an example, according to historical analysis, the highest temperature in the Bohai region is 40, and the lowest temperature is < -20, and the temperature difference between adjacent hours was >5. These were regarded as abnormal data. The core codes for processing noisy data are as follows:

```
for(j=0;j<air_temp.length;j++){
if(air_temp[j]!=null&&air_temp[j+1]!=null){
if(air_temp[j]>40||air_temp[j]<-20||air_temp[j+1]-air_temp[j]>5||air_temp[j+1]-
air_temp[j]<-5){
    avg1(j,id,air_temp);c++;}}}
```

Results and analysis

In order to verify the feasibility of data cleaning techniques applied to information acquired in the field during agricultural monitoring, integration technologies were analyzed and verified using the Bohai granary project.

Using air temperature as an example, the data collected through IOT were visualized. Blanks represent missing data, and were divided into single point and continuous deletion classes, as shown in the Figure 4. In the data, the sudden ascendant or sudden descendant are noisy data that need to be cleaned. They are shown in the Figure 5.

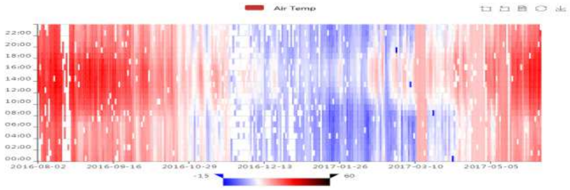

Fig. 4. Original air temperature thermal chart

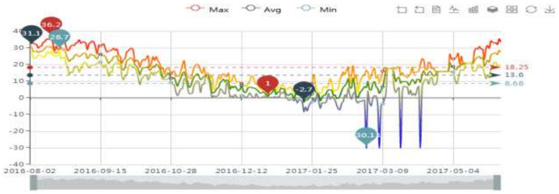

Fig. 5. Original air temperature line chart

After processing, missing and abnormal value have been reduced, as shown in Figure 6 and Figure 7.

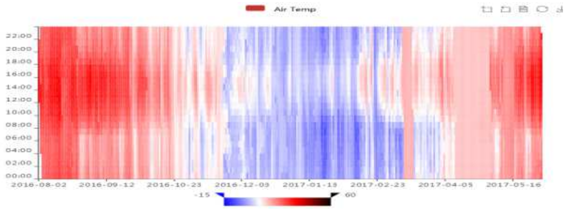

Fig. 6. Air temperature thermal chart after processing

In order to accurately demonstrate the data processing module, we compare data sets from August 2, 2016 until June 2, 2017 as shown in Table 1. Experiments show that the cleaning methods are successful.

Table 1. Data analysis before and after data processing

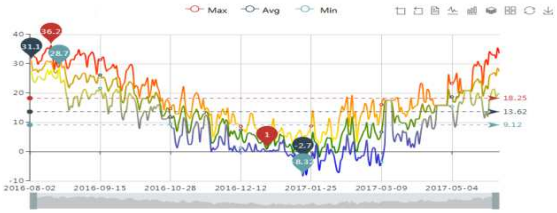

Fig. 7. Air temperature line chart after processing

Class	Before processing	After processing
Sum	6028	7281
Duplicate data	3	0
Single missing data	303	0
Multipoint missing data	956	6
Noise data	4	0

4. Summary and discussion

This paper studies and implements techniques for data cleaning. Meanwhile, related technologies are applied to verify the feasibility of the whole plan based on the research. The paper focuses on application of the data cleaning techniques in the context of agricultural monitoring, and in particular, on crop monitoring.

However, with the range and business of crop monitoring expanding, new demands and problems in monitoring processes will soon appear, as related technologies at home and abroad are undergoing change and improvement. Thus, more work on agricultural monitoring remains to be done; however, in this paper, we put forward future research prospects.

(1) Data must be trusted by the user

Credibility includes accuracy, integrity, consistency, validity, uniqueness, and other indicators [36]. Data cleaning should seek to ensure that data are accurate, missing records or missing fields are repaired, same attribute of the same entity in different systems is identical, and the data meet user-defined conditions or thresholds, such that problems can be solved successfully.

(2)Application of data mining in data cleaning

Effectively cleaning missing values despite different data mining patterns requires further analysis and research. Improving the accuracy of cleaning algorithms, reducing the computing time, and improving the efficiency of data cleaning need further research.

(3)Cleaning of unstructured data

Until recently, data cleaning has focused on structured data, but over the past several years, unstructured or semi-structured data (such as XML) have received more attention. This is partly due to the characteristics of XML itself including generality and self-description; however, more attention should be paid to such data sources in data cleaning.

(4)Application of big data in data cleaning

The core of data thinking is to make full use of good data to solve real problems; therefore in data cleaning, attention should be paid to training practitioners in data thinking. The data thinking approaches needed include data is the core; data has value; a guarantee of whole data; data improves efficiency; big data correlations; and big data predictions. In the era of big internet data, thinking in terms of big data is part of an objective existence [37]. The trend of today's information age is to think and solve problems with this new approach. In data cleaning, in order to achieve data mining successfully, we should use a variety of big data methodologies.

References

[1] S. Xu, B. Lu, M. Baldea: *Data cleaning in the process industries.* Reviews in chemical engineering *31* (2015), No. 5, 453–490.

[2] J. Z. Zhang, Z. Fang, Y. J. Xiong: *Optimization algorithm for cleaning data based on SNM.* Journal of central south university *41* (2010), No. 6, 2240–2245.

[3] M. Celton, A. Malpertuy, G. Lelandais: *Comparative analysis of missing value imputation methods to improve clustering and interpretation of microarray experiments.* Bmc genomics *11* (2010), No. 1, 1–16.

[4] M. Nakai, D. G. Chen, K. Nishimura: *Comparative study of four methods in missing value imputations under missing completely at random mechanism.* Open journal of statistics *4* (2014), No. 1, 27–37.

[5] M. C. Limas , O. Mer: *Outlier detection and data cleaning in multivariate non-normal samples: the paella, algorithm.* Data mining & knowledge discovery *9* (2004), No. 2, 171–187.

[6] G. Xiao: *Data processing model of bank credit evaluation system.* Journal of software *6* (2011) 1241–1247.

[7] X. Y. Wang, J. L. Zhang, W. U. Fang: *Research on traffic flow data cleaning rules.* Computer engineering *37* (2011), No. 20, 191–193.

[8] H. Zhang, Q. Wang, J. Zhu: *Influence of sample data preprocessing on BP neural network-based GPS elevation fitting.* Journal of geodesy & geodynamics *31* (2011), No. 2, 125–128.

[9] S. V. Dronov, E. A. Dementjeva: *A new approach to post-hoc problem in cluster analysis.* Journal of applied physiology respiratory environmental & exercise physiology *57* (2003), No. 57, 44–51.

[10] E. V. Bystritskaya, A. L. Pomerantsev, O. Y. Rodionova: *Non regression analysis: new approach to traditional implementations.* J. Chemometrics *14*, (2015), No. 5, 667–692.

[11] D. C. Montgomery, E. A. Peck: *Introduction to linear regression analysis.* Journal of the royal statistical society *170* (2001), No. 3, 856–857.

Received November 16, 2017