frontiers in Neurorobotics

ORIGINAL RESEARCH
published: 14 June 2021
doi: 10.3389/fnbot.2021.680613

A Clustering Algorithm for Multi-Modal Heterogeneous Big Data With Abnormal Data

An Yan[1], *Wei Wang*[1,2], *Yi Ren*[1] *and HongWei Geng*[1]*

[1] *Xinjiang Agricultural University, Ürümqi, China,* [2] *Anyang Institute of Technology, Anyang, China*

The problems of data abnormalities and missing data are puzzling the traditional multi-modal heterogeneous big data clustering. In order to solve this issue, a multi-view heterogeneous big data clustering algorithm based on improved Kmeans clustering is established in this paper. At first, for the big data which involve heterogeneous data, based on multi view data analyzing, we propose an advanced Kmeans algorithm on the base of multi view heterogeneous system to determine the similarity detection metrics. Then, a BP neural network method is used to predict the missing attribute values, complete the missing data and restore the big data structure in heterogeneous state. Last, we ulteriorly propose a data denoising algorithm to denoise the abnormal data. Based on the above methods, we construct a framework namely BPK-means to resolve the problems of data abnormalities and missing data. Our solution approach is evaluated through rigorous performance evaluation study. Compared with the original algorithm, both theoretical verification and experimental results show that the accuracy of the proposed method is greatly improved.

Keywords: multi-view, BP neural network, missing attributes, Kmeans, noise reduction processing, data integrity

Edited by:
Peng Li,
Dalian University of Technology, China

Reviewed by:
Haifeng Li,
Dalian University of Technology, China
Daohua Liu,
Xinyang Normal University, China

***Correspondence:**
HongWei Geng
hw-geng@163.com

Received: *15 March 2021*
Accepted: *27 April 2021*
Published: *14 June 2021*

Citation:
Yan A, Wang W, Ren Y and Geng H (2021) A Clustering Algorithm for Multi-Modal Heterogeneous Big Data With Abnormal Data. Front. Neurorobot. 15:680613. doi: 10.3389/fnbot.2021.680613

INTRODUCTION

As the carrier of information, data must accurately and reliably reflect the objective things in the real world (Murtagh and Pierre, 2014; Brzezińska and Horyń, 2020). How to extract effective information on a large number of data sets for data mining, in addition to effective data analysis technology, good data quality is the basic condition of various data mining (Adnan et al., 2020). In the era of big data, data quality is a key issue that restricts the development of the data industry (Zeng et al., 2021). Therefore, how to effectively ensure the integrity and accuracy of the data and improve the data quality has become an urgent problem to be solved. As data collection and data expression methods become more and more diversified, it has become more convenient to obtain a large amount of multi-source heterogeneous data (Rashidi et al., 2020; Wu et al., 2020). The emergence of multi-source heterogeneous data and the need to mine the inherent information on such data naturally gave rise to modeling learning for multi-source heterogeneous data. Currently, there are two main forms of multi-source heterogeneous data: multi-modal data and multi-view data. Multi-view data refers to the data obtained by describing the same thing from different ways or different angles (Kaur et al., 2019). The meaning of multi-view includes multi-modality, multi-view can express a wider range of practical problems (Ma X. et al., 2021). At present, data labels are usually difficult to obtain, the manifestation of heterogeneous data itself is extremely different. In addition, the noise and outliers contained in the original data put forward higher

requirements on the robustness of the algorithm (Ma et al., 2020; Yang et al., 2020). In particular, there are often more noise and outliers in multi-source heterogeneous data, which greatly affects the performance of the algorithm in practical applications. Therefore, unsupervised learning for multi-source heterogeneous data has important theoretical research value and broader application scenarios (Li et al., 2018).

In multi-view data, the information contained in different views usually complements each other (Sang, 2020). Fully mining the data of each view to obtain more comprehensive information is the main goal of multi-source heterogeneous data learning. The earliest multi-source heterogeneous data learning model can be traced back to the two-source data learning model based on canonical correlation analysis (Ruan et al., 2020), which mines the consistent structure information of the data on the basis of the correlation between the two-source data. In addition, Bickel and Scheffer proposed a k-means-based multi-view clustering algorithm and used it to analyze data with two conditionally independent views for text clustering. Referring to the existing literature, the model proposed by (Bickel and Scheffer, 2004) in 2004 is the first literature to study multi-view clustering. (De Sa, 2005) proposed a simple and effective spectral clustering algorithm in the literature, and used the algorithm to process web page data containing two views. This method first uses the similarity matrix to fuse the feature information of the two views, and then uses the classical spectral clustering algorithm to perform clustering and obtain the final clustering result (Zhou et al., 2020).

Self-organizing map (SOM) is an algorithm that uses artificial neural networks for clustering (Ma J. et al., 2021). This method processes all the sample points one by one, and maps the cluster centers to a two-dimensional space to realize visualization. (Yu et al., 2015) proposed an intuitionistic fuzzy kernel clustering algorithm based on particle swarm optimization. (Wu and Huang, 2015) proposed a new DP-DBS can clustering algorithm based on differential privacy protection, which implements a differential protection mechanism. Deep learning is another leap-forward development of artificial neural networks (Ventura et al., 2019). At present, clustering algorithms based on deep learning have also become a hot topic of research. The above-mentioned neural network-based Kmeans algorithm improves the clustering effect in the optimization process after the clustering center is given, but it does not clearly provide a method to determine the lack of cluster numerical attributes and data abnormalities. At the same time, there are many types and attributes of structured data and data collection is becoming more and more complex, resulting in more and more data shortages and data abnormalities. However, it is not that missing data is worthless and often it may be the value of missing data.

Aiming at the above problems, this paper proposes a BPK-means algorithm to improve the BP neural network. The algorithm first predicts the missing attribute values based on the BP neural network, which greatly improves the integrity and reliability of the data; then demises the abnormal data, and finally clusters the same data through different views to verify the difference relevance to attributes and clustering research.

The rest of the paper is organized as follows: Section Preliminaries discusses the basic algorithm. Section Problem Formalization presents the formally define the BPK-means algorithm. Section Optimization of Kmeans clustering algorithm proposes optimization of Kmeans clustering algorithm namely BPK-means algorithm. Finally, we analyze the proposed algorithm through experimental results in Section Experiment and Result Analysis.

PRELIMINARIES

In this part, it mainly provides the basic algorithm of Kmeans algorithm and BP neural network, which are studied in this paper.

Kmeans Algorithm

The traditional Kmeans algorithm is an unsupervised learning algorithm, that is to cluster the unlabeled data set. Algorithm main idea is as follows: firstly, K initial cluster centers are randomly selected, and each cluster center represents a data set cluster. Then the data points are divided into the nearest data cluster. Finally, the data cluster center is recalculated until the clustering criterion function converges. The convergence function is defined as follows:

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - u_i\|_2^2 \tag{1}$$

Where E is the minimum square error of $C = \{C_1, C_2, \ldots, C_k\}$ obtained by K-means algorithm for $D = \{x_1, x_2, \ldots, x_k\}$, and u_i is:

$$u_i = \frac{1}{|C_i|} \sum_{x \in C_i} x \tag{2}$$

Where u_i is the mean vector of the C_i. Intuitively, the above formula describes the closeness of the samples in the cluster around the cluster mean vector to a certain extent. Generally, the smaller the E is, the higher the similarity of samples in the cluster is.

Definition 1 Euclidean distance is the linear distance between two points in Euclidean space. The Euclidean distance between sample x_i and x_j in m-dimensional space is as follows:

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2} \tag{3}$$

The original K-means algorithm determines the similarity of samples according to Euclidean distance.

BP Neural Networks

The more layers of neural network, the stronger learning ability. Generally, the learning ability of multi-layer network is better than that of single-layer network, but multi-layer network needs stronger learning algorithm. BP neural network algorithm is one of them, which is widely used.

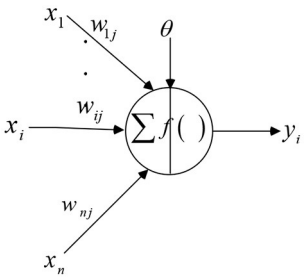

FIGURE 1 | Schematic diagram of BP neuron model.

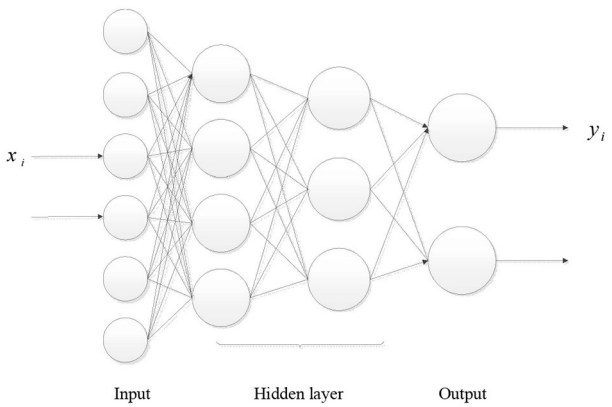

FIGURE 2 | Schematic diagram of BP neural network structure.

TABLE 1 | The main notations.

Symbols	Definitions
$E=\sum_{i=1}^{k}\sum_{x\in C_i}\|x-u_i\|_2^2$	Convergence function
$D=\{x_1,x_2,\ldots,x_k\}$	Original data samples
$C=\{C_1,C_2,\ldots,C_k\}$	Clusters
$u_i=\frac{1}{\|C_i\|}\sum_{x\in C_i}x$	The mean vector of the C_i
$d(x_i,x_j)=\sqrt{\sum_{k=1}^{m}(x_{ik}-x_{jk})^2}$	Euclidean distance
$a=\log sig(w+b)$	Transfer function
$p\cup q=\phi$	itp is missing data, and itq is noisy data
$\delta(x,y)=\begin{cases}1\ if(x=y)\\0\ otherwise\end{cases}$	Represents an indicator function
$ca=\frac{\sum_{i=1}^{n}\delta(t_i,map(r_i))}{n}$	Clustering precision
$B=\{b_1,b_2,\ldots,b_n\}$	Data samples after attribute completion
$A=\{a_1,a_2,a_3,\ldots,a_n\}$	Data samples after noise reduction processing
$R=\{r_1,r_2,r_3,\ldots,r_k\}$	The divided clusters of the cluster
$\{v_1,v_2,v_3,\ldots,v_k\}$	The initial cluster center
$d_{ij}=\|x_j-v_i\|^2$	The distance from x_j to each vector v_i
$r_j=\arg\min_{i\in\{1,2,\ldots,k\}}d_{ji}$	Mark the nearest center x_j
$f(x)=\frac{1}{1+e^{-x}}$	The S-type transfer function
$E=\frac{\sum_{l}(t_l+O_l)^2}{2}$	The error function
$l=\sqrt{m+n}+a$	The number of hidden layers
itTraingdx	The training function
itmse	The performance function
itlr	The learning rate
$I=(i_1,i_2,\ldots,i_k)$	Outlier points
$\varepsilon=\frac{\sum_{i=1}^{n}\|i_j-b_j\|}{n}$	Threshold of outliers error range

BP neural network algorithm is a kind of multilayer feedforward network. Firstly, the difference between the output values and the expected value of the network is calculated. Then the partial derivative of the difference is obtained by using the function derivative method, and the feedback processing is carried out along the opposite direction of signal transmission of the system.

The basic idea of BP neural network learning algorithm is as follows (Hosseini and Azar, 2016; Li et al., 2017; Kanaan-Izquierdo et al., 2018; Wu et al., 2019):firstly, the data are input into the neural network of the selected samples. Then the results are processed and calculated in the hidden layer and the output results are taken as the input signals of the next layer, thus the error between the results of the output layers and the expected value of the neural network is obtained. Finally, the connection weights of the interconnected neurons in the neural network are adjusted to the direction of the minimum value of the error surface, and the error solving process is repeated until the output error of the whole neural network reaches the accuracy required.

The learning rule of BP neural network adopts the steepest descent method. Through the back propagation of the network, the weights and thresholds of the network are adjusted continuously to minimize the output error of the network. The topological structure of BP neural network models includes input layer, hidden layer and output layer. The BP neuron model is shown in **Figure 1**.

Let the input signal of BP neuron be P, the weight and threshold be w and b, respectively, and the processing result be y. Transfer functions are commonly used as *logsig* and *tansig*. The formula of *logsig* function is as follows:

$$a = \log sig(w + \mathbf{b}) \tag{4}$$

The structure of BP neural network is shown in **Figure 2**, including input layer, hidden layer and output layer.

PROBLEM FORMALIZATION

In this part, we introduce the models for scientific workflows. Then, we formulate the scheduling problem. To improve the readability, we sum up the main notations used throughout this paper in **Table 1**.

System Models

Scientific Model

A model of multi-modal heterogeneous big data with abnormal data in this paper is data aggregation clustering,

i.e., $D = \{x_1, x_2, \ldots, x_k\}$. Firstly, we assume that there is no missing and noise data in the initial data D, then for the multi-view clustering of D, the final clustering accuracy is evaluated according to the combination of different attributes of D data. However, for some abnormal data, such as missing and noise data, the clustering effect will be greatly reduced for the same clustering algorithm. Now we assume that for the original data D, there are p records containing missing values and q records containing noise data.

Problem Formulations

Assuming the data D, the attribute field is missing p data, and q data are noisy data. Suppose $p \cup q = \phi$, The highest clustering accuracy of the traditional Kmeans algorithm is $(M\text{-}p\text{-}q)/M$. Of course, it's basically impossible.

The goal of our proposed algorithm is that the accuracy is greater than $(M\text{-}p\text{-}q)/M$. In addition, in the multi-modal data view, the traditional Kmeans algorithm has limited fault tolerance for different modal data. The data fault tolerance of our proposed algorithm can be improved. The final result is that the proposed algorithm has a higher clustering accuracy than the Kmeans algorithm. Further, we express it through formal language $\delta(x, y)$ and ca as follows:

$$\delta(x, y) = \begin{cases} 1 \; if(x = y) \\ 0 \; otherwise \end{cases} \tag{5}$$

$$ca = \frac{\sum_{i=1}^{n} \delta(t_i, map(r_i))}{n} \tag{6}$$

where t_i is real labels, r_i is labels after clustering, n is the total number of data, $\delta(x, y)$ represents an indicator function. map function represents the optimal reproduction allocation of class labels.

$$\text{Objective1} : \max ca \tag{7}$$

$$\text{Subjectto} : p \cup q = \phi \tag{8}$$

OPTIMIZATION OF KMEANS CLUSTERING ALGORITHM

Generally, for the traditional Kmeans algorithm, the incomplete data is preprocessed and checked, and the preprocessed data is clustered. In addition, a lot of noise data appear due to the variety of acquisition methods. How to fill the missing attribute data and filter the noise data, the traditional Kmeans algorithm does not give a good solution. In view of the above problems, according to the characteristics that BP neural network can well-predict and detect unknown data, this paper proposes a BPK-means algorithm based on BP neural network to improve the Kmeans algorithm. BPK-means algorithm flow chart is described by **Figure 3**. We use the BP neural network attribute completion algorithm to detect missing attributes and by using Data denoising algorithm to process noise data.

BPK-Means Algorithm

The improvement of BPK-means algorithm and traditional Kmeans algorithm lies in the optimization of data attribute missing and denoising. Some of the missing attribute data is still meaningful. If it is discarded randomly, the result of data clustering will be inaccurate. BP neural network model can predict the missing data, and the prediction accuracy of the data can reach more than 90%, which greatly ensures the integrity of the data. In addition, outliers are used to analyze some noise data, and then BP neural network is used to judge whether the data is valid. BPK-means algorithm flow is described by Algorithm 1. Besides this, We define several functions in Algorithm 1.

$$d_{ij}(x_j, v_i) = \left\| x_j - v_i \right\|^2 \tag{9}$$

Where $d_{ij}(x_j, v_i)$ is the distance of x_j and v_i.

$$r_j = \arg\min_{i \in \{1,2,\ldots,k\}} d_{ij} \tag{10}$$

r_j is the minimum distance of $d_{ij}(x_j, v_i)$,which can determines cluster markers of x_j.

Algorithm 1: BPK-means algorithm.

Input: Sample set $D = \{x_1, x_2, \ldots, x_k\}$; Number of clusters k;
Output: The divided clusters of the cluster $R = \{r_1, r_2, r_3, \ldots, r_k\}$;
1:Use BP neural network to complete the missing attributes of data set D;
2: Using outliers and BP neural network to denoise the data, the processed data set sample is:

$$\text{A}=\{\text{a}_1,\text{a}_2,\text{a}_3,\ldots,\text{a}_n\};$$

3: Randomly select k samples from A as the initial vector, that is, the initial cluster center is recorded as the vector:

$$\{v_1,v_2,v_3,\ldots,v_k\};$$

4: Order $C_i = \varnothing \left(1 \le i \le k\right)$;
5: Loop $j = 1, 2, \ldots, n$;
6: Calculate $a_j$the distance to each vector $v_i\left(1 \le i \le k\right)$ and record it as$d_{ij}(x_j, v_i)$;
7: The cluster mark determined according to the nearest center x_j point r_j;
8: Group the samples x_j into corresponding clusters:

$$C_{r_j} = C_{r_j} \cup x_j;$$

9: Circulation order $i = 1, 2, 3, \ldots, k$;
10: Calculate the new cluster vectorv_i';
11: If$v_i' \neq v_i$, at this time, it is necessary to update the cluster class vector tov_i';
12: Otherwise, keep the current cluster vectorv_i unchanged;
13: End the loop until the cluster vector is not changed.

The BPK-means algorithm adds the integrity recovery of the data set and the detection of noise, which cannot guarantee the integrity of the data, and prevents the loss of important attributes of the data and low clustering accuracy. In order

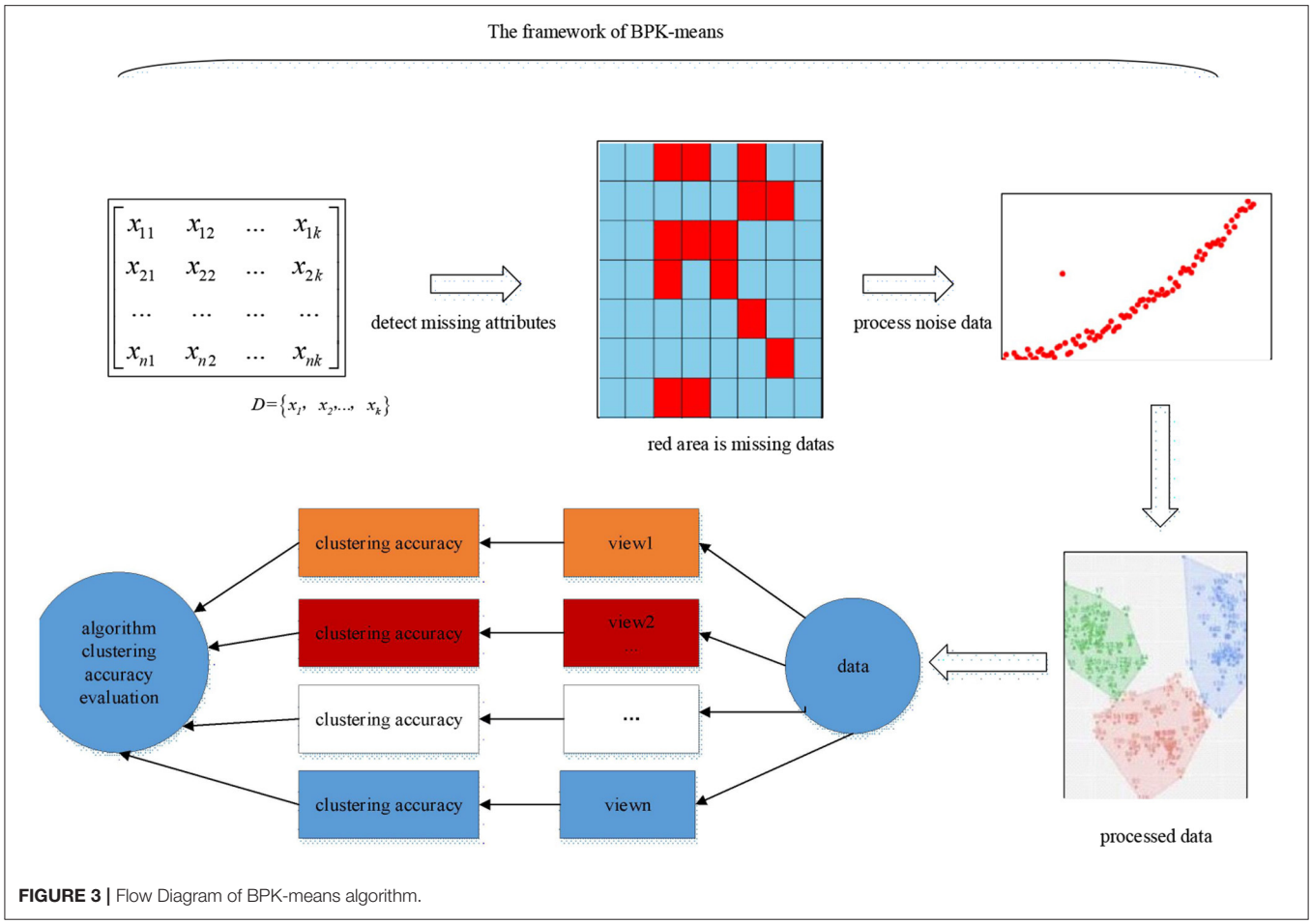

FIGURE 3 | Flow Diagram of BPK-means algorithm.

to show the execution process of the BPK-means algorithm more intuitively.

BP Neural Network Attribute Completion Algorithm

In BPK-means algorithm, BP neural network is used to complete the missing attributes of data set D in the first step. Next, the implementation process of the algorithm is described in detail. BP neural network attribute completion algorithm flow is described by Algorithm 2. What's more, We have made some functions to illustrate the application in Algorithm 2.

$$f(x)=\frac{1}{1+e^{-x}} \tag{11}$$

The function $f(x)$ is the S-type transfer function, we use it in BP net.

$$E=\frac{\sum_{i}(t_i+O_i)^2}{2} \tag{12}$$

Where E is the error function. $t_i$is the expected output and O_i is the computational output of the network.

$$l=\sqrt{m+n}+a \tag{13}$$

l is the number of hidden layers, Where m is the number of neurons in the input layer, n is the number of neurons in the output layer, and a is a constant between (Murtagh and Pierre, 2014; Ma et al., 2020). According to a large number of experimental data, this algorithm sets $a=3$.

BP neural network attribute completion algorithm makes use of the characteristics of BP neural network. It uses effective data to build sample set for model training, and makes prediction and evaluation for missing attribute data. It not only ensures the integrity of the data, but also the artificial data is more scientific and accurate to a certain extent.

Data Denoising Algorithm

BP neural network is used to denoise data set D in the second step. Next, in the third part, the implementation process of the algorithm is described in detail. Data denoising algorithm is described by Algorithm 3. We define the error range function ε in Algorithm 3.

$$\varepsilon=\frac{\sum_{j=1}^{n}\left|i_j-b_j\right|}{n} \tag{14}$$

Algorithm 2: BP neural network attribute completion algorithm.

Input: sample set $D = \{x_1, x_2, \ldots, x_k\}$;
Output: the whole data set $B = \{b_1, b_2, \ldots, b_n\}$;
1: Scan the data set once. Find out the record number of the data set and record the data set with incomplete attributes as $Q = \{q_1, q_2, q_3, \ldots, q_m\}$;
2: Judge the size of N. If N is more than 100000 records, then randomly select 20% as the training sample of neural network. If N is $\leq$100,000 records, then select 60% of the data set as the training sample set;
3: Three layers BP neural network model is constructed, which are input layer, hidden layer and output layer;
4: The S type transfer function is set $f(x)$.
5: The inverse error output is set, and the network weight and threshold are adjusted continuously to minimize the error function E.
According to all the samples selected in the second step, the network is modeled. In this model, the attribute of data set is used as input, and the number of output nodes is set to 1, the l is used in the design of hidden layer.
6: The excitation function of hidden layer and output layer is *Tansig* and *Logsig*, respectively. The training function is *Traingdx*. The performance function is *mse*. The number of iterations is 50,000. The expected error goal is 0.000000001, and the learning rate lr is 0.01.
7: According to the setting of the network model in the previous steps, the network model is constructed and trained. In this way, the missing data set in $Q = \{q_1, q_2, q_3, \ldots, q_m\}$ is predicted, and a complete data set is constructed and recorded as $B = \{b_1, b_2, \ldots, b_n\}$.

Algorithm 3: Data denoising algorithm.

Input: Complete sample set $B = \{b_1, b_2, \ldots, b_n\}$;
Output: Data set after noise reduction $A = \{a_1, a_2, a_3, \ldots, a_n\}$;
1: $B = \{b_1, b_2, \ldots, b_n\}$ the data are clustered by initial algorithm using Kmeans algorithm;
2: Finding out the points outside the cluster set. It is called outlier points $I = (i_1, i_2, \ldots, i_k)$;
3: For each outlier, BP neural network is used to predict the corresponding attribute value, which is compared with the existing values. We define an error range functionε.
If ε is greater than the given threshold, it is considered as a noise point for noise processing. Finally, a noise free data set is formed:

$$A = \{a_1, a_2, a_3, \ldots, a_n\}.$$

Data denoising can keep the smoothness of data. In this way, the data can be directly clustered after processing to improve the clustering accuracy of the data.

BPK-Means Algorithm Performance Analysis

The time complexity of traditional Kmeans algorithm depends on the number of attributes per record, the size of data, the number of iterations and the type of clustering. Its time complexity is expressed as O (I^*n^*k $^*m$), where I is the number of iterations, k is the type of clustering, m is the scale of data volume, and n is the number of attributes of each record. The BPK-means algorithm adds missing attribute field completion and attributes data prediction on the basis of Kmeans algorithm.

Suppose that p data are missing in attribute field and q data are noise data, then $(p + q)$ neural network data processing is needed. At present, BP neural network has three layers. Assuming that the number of neurons in each layer is $n1$, $n2$, and $n3$, respectively. For example, if a sample ($n1$ * 1) carries out feedforward calculation, it needs to carry out two matrix operations and two matrix multiplication. $n1^*n2$ and $n2^*n3$ calculations are performed respectively. The number of nodes ($n1$ and $n3$) in input layer and final output layer is determined and can be regarded as constant. The hidden layer $n2$ in the middle can be set by oneself. The time complexity of feedforward computation for a sample should be O($n1^*n2$ + $n2^*n3$) = O($n2$). The time complexity of back propagation is the same as that of feedforward. Suppose there are m training samples in total, and each sample is trained only once, then the time complexity of training a neural network should be O(m^*n2). Similarly, if a sample is predicted, the time complexity should be O($n2$).

The total time of calculating each data result is (m+2) O (n_2). Since the training samples m and n_2 can be determined basically, the time of each neural network is also linear. Finally, the total time complexity of BP neural network is(p+q)(m+2)O(n_2). The BPK-means algorithm focuses more on the accuracy of the algorithm, so that the time complexity is slightly higher than the original algorithm. For some scenes with high accuracy requirements, the cost is worth it.

EXPERIMENT AND RESULT ANALYSIS

Experimental Setup and Experimental Environment

In order to verify the effectiveness of the algorithm, four groups of experiments are carried out.

Experiment 1: verify the missing attribute completion algorithm through UCI data set to verify the approximation analysis of missing data and simulation prediction data, and verify the effectiveness of the algorithm.
Experiment 2: test the denoising algorithm in this paper through UCI data set to verify the denoising effect of this algorithm.
Experiment 3: the clustering accuracy of BPK-means algorithm and traditional K-means algorithm is compared and analyzed through different data sets.
Experiment 4: the clustering accuracy of BPK-means algorithm and traditional K-means algorithm is compared and analyzed through different multiple views.

The running environment of the experiment is Windows 7 operating system, 4 GB physical memory, CPU speed of 3.10 GHZ and programming language is Python3.8.

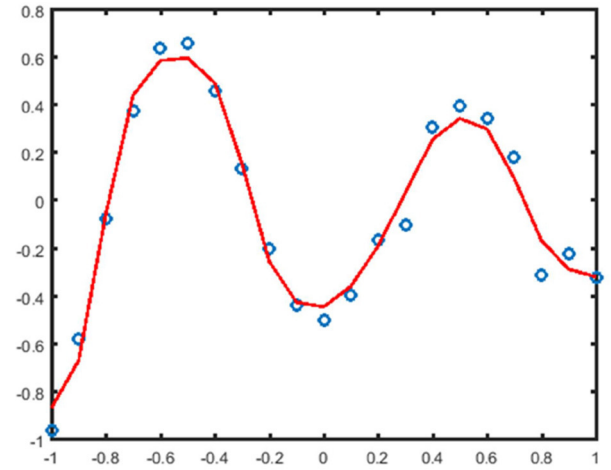

FIGURE 4 | Approximation analysis graph of missing sample attributes.

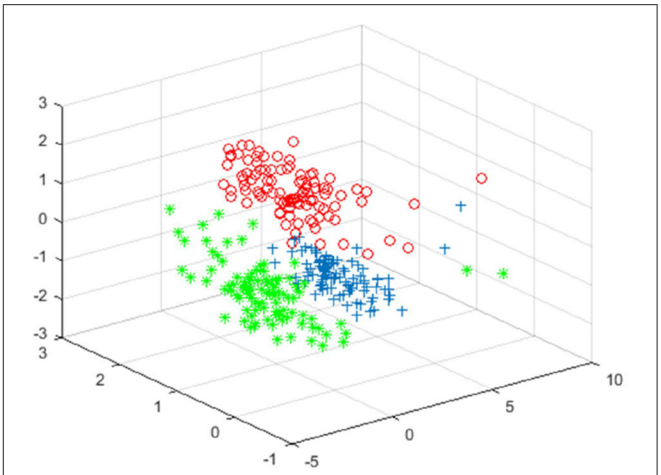

FIGURE 6 | Clustering effect map of the data after noise.

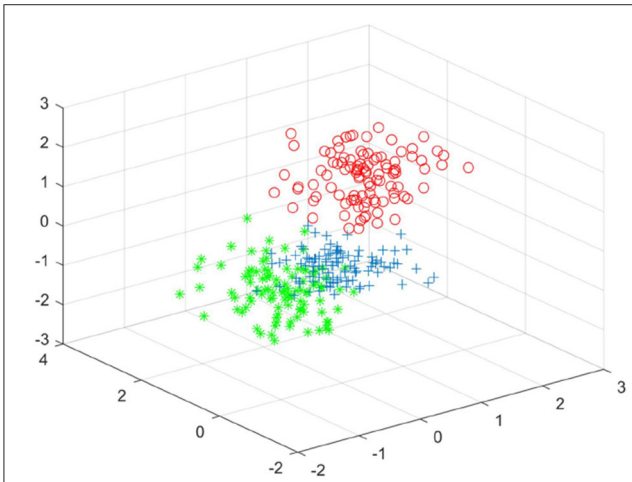

FIGURE 5 | The clustering effect diagram of the completion data of experiment one.

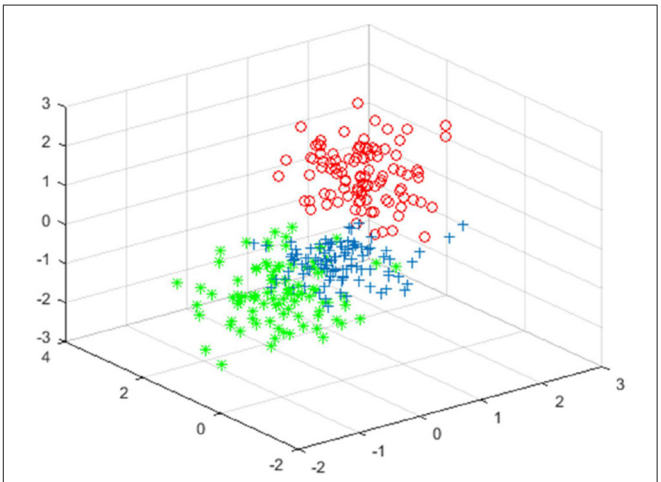

FIGURE 7 | The clustering effect map of the data after noise removal.

Experiment and Result Analysis

A Completion Algorithm for Verifying Missing Attributes

Iris data (UCI Iris, 2021) set of UCI is selected as the sample data set of the experiment. The data contains four attributes and 150 records. Randomly select 100 records as the training sample set, and then the rest of the data set, randomly select 10, making some attributes missing for the experiment. Then randomly select 10 items in the remaining data set to make some attributes missing for experiment.

The above experiments compare the approximation analysis of the predicted value and the real value of the missing attribute data. From **Figure 4**, the BP neural network can be used to predict the missing attribute data well and achieve the effect of data completion.

Validation of Data Denoising Algorithm

Select the complete data set in Experiment 1, and then add noise artificially to verify whether the denoising algorithm is feasible.

Figures 5–**7** shows the effect drawing of data completion, the effect picture after noise addition and noise removal, respectively.

Validation of BPK-Means Algorithm

The standard data set is selected for clustering verification experiment. The data set is from the public UCI data (UCI Iris, 2021) set. The selected data set randomly set 3–10% missing data attributes to verify the clustering effect of BPK-means algorithm and K-means algorithm.

Through the comparison of clustering effect of different algorithms and the analysis of UCI data set clustering results from **Figure 8**, we can see that it is verified that this algorithm has the advantage of high clustering accuracy in the process of clustering analysis for a small number of data with missing attributes about **Figure 9**.

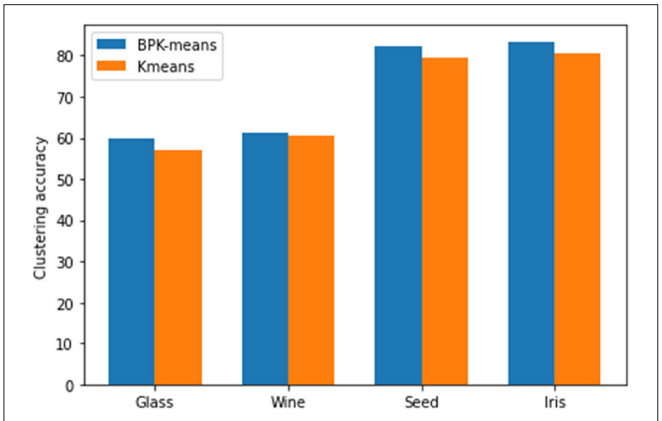

FIGURE 8 | Comparison chart of clustering effect of different data sets.

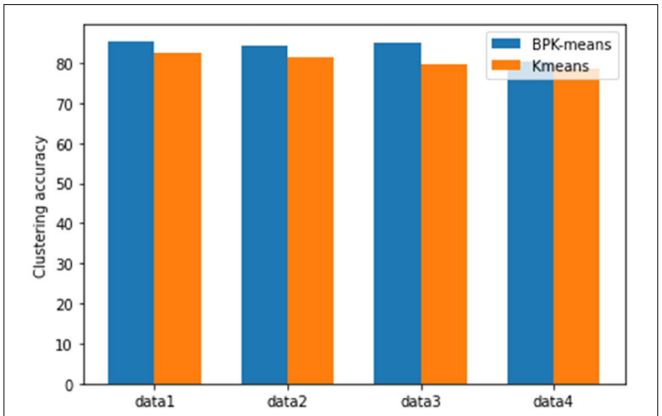

FIGURE 9 | View clustering effect diagram.

Validation the BPK-Means Algorithm in Multiple Views

The IRIS data set is selected as the experimental data set. The IRIS data set contains 4 columns of data sets and 1 column of label data. The data is randomly added with noise and missing. Sepal length and petal width are selected as data1, petal length and petal width are data2, sepal length and petal length are data3, and sepal width and petal width are data 4 for the accuracy of the multi-view verification algorithm.

By comparing the BPK-means algorithm under different views in **Figure 9**, the accuracy of the algorithm is improved in different modes, which further shows that the BPK-means algorithm proposed in this paper can be more accurate for some multi-view clustering problems than traditional algorithms.

Through the above four experiments, the traditional algorithm clustering accuracy is performed from attribute missing, data noise, BPK-means algorithm and Kmeans algorithm, and the clustering accuracy of BPK-means algorithm and Kmeans algorithm is verified against Algorithm to verify.

CONCLUSION

This paper proposes an improved Kmeans algorithm of BP neural network. The algorithm uses the characteristics of BP neural network to predict and analyze the missing data attribute values. After the completion of the data set, perform data noise reduction processing to remove suspected noise Points are eliminated using BP neural network to ensure the smoothness of the clustering effect, so that the problem of missing data attribute values due to inconsistencies in the form and type of collected data can be solved. Four sets of experiments verify that the BPK-means algorithm proposed in this article can solve the problem of missing attributes, and can solve the clustering accuracy of the data, especially in the multi-view scenario, which is more accurate than the traditional algorithm. However, when the algorithm proposed in this paper is used to process larger data sets, a problem of high time complexity will be generated. Therefore, it is necessary to study how to better improve the accuracy while reducing the time complexity. At the same time dealing with missing data and noisy data reasonably and effectively is also the future improvement and research direction of the algorithm in this paper. This is also the future improvement and research direction of the algorithm in this paper.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found at: http://archive.ics.uci.edu/ml/datasets/Iris.

AUTHOR CONTRIBUTIONS

WW and AY designed the research, performed the experiments, analyzed the data, and wrote the manuscript. YR gave a lot of suggestions in the design of the experiment and the modification of the article. HG performed the experiments. All authors contributed to the article and approved the submitted version.

REFERENCES

Adnan, R. M., Khosravinia, P., Karimi, B., and Kisi, O. (2020). Prediction of hydraulics performance in drain envelopes using Kmeans based multivariate adaptive regression spline. *Appl. Soft Comput.* 100:107008. doi: 10.1016/j.asoc.2020.107008

Bickel, S., and Scheffer, T. (2004). "Multi-view clustering," in *Proceedings of the IEEE International Conference on Data Mining*, 19–26.

Brzezińska, A. N., and Horyń, C. (2020). Outliers in rules - the comparision of LOF, COF and KMEANS algorithms. *Proc. Comput. Sci.* 176, 1420–1429. doi: 10.1016/j.procs.2020.09.152

Chen, J., and Huang, G. B. (2021). Dual distance adaptive multiview clustering. *Neurocomputing* 441, 311–322. doi: 10.1016/j.neucom.2021.01.132

De Sa, V. R. (2005). "Spectral clustering with two views," in *Proceedings of the ICML Workshop on Learning With Multiple Views* (Bonn), 20–27.

Hosseini, M., and Azar, F. T. (2016). A new eigenvector selection strategy applied to develop spectral clustering. *Multidimens. Syst. Signal Process.* 28, 1227–1248. doi: 10.1007/s11045-016-0391-6

Kanaan-Izquierdo, S., Ziyatdinov, A., and Perera-Lluna, A. (2018). Multiview and multifeature spectral clustering using common eigenvectors. *Pattern Recognit. Lett.* 102, 30–36. doi: 10.1016/j.patrec.2017.12.011

Kang, Z., Zhao, X., Peng, C., Zhu, H., Zhou, J. T., Peng, X., et al. (2020). Partition level multiview subspace clustering. *Neural Netw.* 122, 279–288. doi: 10.1016/j.neunet.2019.10.010

Kaur, A., Pal, S. K., and Singh, A. P. (2019). Hybridization of chaos and flower pollination algorithm over K-means for data clustering. *Appl. Soft Comput.* 97:105523. doi: 10.1016/j.asoc.2019.105523

Li, P., Chen, Z., Yang, L. T., Gao, J., Zhang, Q., and Deen, M. J. (2018). An incremental deep convolutional computation model for feature learning on industrial big data. *IEEE Trans. Indust. Inform.* 15, 1341–1349. doi: 10.1109/TII.2018.2871084

Li, P., Chen, Z., Yang, L. T., Zhao, L., and Zhang, Q. (2017). A privacy-preserving high-order neuro-fuzzy c-means algorithm with cloud computing. *Neurocomputing* 256, 82–89. doi: 10.1016/j.neucom.2016.08.135

Ma, J., Zhang, Y., and Zhang, L. (2021). Discriminative subspace matrix factorization for multiview data clustering. *Pattern Recognit.* 111:107676. doi: 10.1016/j.patcog.2020.107676

Ma, W., Li, Q., Li, J., Ding, L., and Yu, Q. (2020). A method for weighing broiler chickens using improved amplitude-limiting filtering algorithm and BP neural networks. *Informat. Proc. Agric.* doi: 10.1016/j.inpa.2020.07.001 Available online at: https://www.sciencedirect.com/science/article/pii/S2214317320301888

Ma, X., Guan, Y., Mao, R., Zheng, S., and Wei, Q. (2021). Modeling of lead removal by living *Scenedesmus obliquus* using backpropagation (BP) neural network algorithm. *Environ. Technol. Innovat.* 22:101410. doi: 10.1016/j.eti.2021.101410

Murtagh, F., and Pierre, L. (2014). Wards hierarchical agglomerative clustering method: which algorithms implement wards criterion? *J. Classicat.* 31, 274–295. doi: 10.1007/s00357-014-9161-z

Rashidi, R., Khamforoosh, K., and Sheikhahmadi, A. (2020). An analytic approach to separate users by introducing new combinations of initial centers of clustering. *Phys. A Stat. Mech. Applic.* 551:124185. doi: 10.1016/j.physa.2020.124185

Ruan, X., Zhu, Y., Li, J., and Cheng, Y. (2020). Predicting the citation counts of individual papers via a BP neural network. *J. Informet.* 14:101039. doi: 10.1016/j.joi.2020.101039

Sang, B. (2020). Application of genetic algorithm and BP neural network in supply chain finance under information sharing. *J. Comput. Appl. Math.* 384:113170. doi: 10.1016/j.cam.2020.113170

UCI Iris (2021). Available online at: http://archive.ics.uci.edu/ml/datasets/Iris

Ventura, C., Varas, D., Vilaplana, V., Giro-i-Nieto, X., and Marques, F. (2019). Multiresolution co-clustering for uncalibrated multiview segmentation. *Signal Process. Image Commun.* 76:151–166. doi: 10.1016/j.image.2019.04.010

Wu, H., Jin, Q., Zhang, C., and Guo, H. (2019). A selective mirrored task based fault tolerance mechanism for big data application using cloud. *Wirel. Commun. Mobile Comput.* 2019:4807502. doi: 10.1155/2019/4807502

Wu, L., Peng, Y., Fan, J., Wang, Y., and Huang, G. (2020). A novel kernel extreme learning machine model coupled with K-means clustering and firefly algorithm for estimating monthly reference evapotranspiration in parallel computation. *Agric. Water Manage.* 245:106624. doi: 10.1016/j.agwat.2020.106624

Wu, W., Huang, H. (2015). Research on DP-DB Scan clustering algorithm based on differential privacy protection. *Comput. Eng. Sci.* 37, 830–834.

Yang, Z., Mao, L., Yan, B., Wang, J., and Gao, W. (2020). Performance analysis and prediction of asymmetric two-level priority polling system based on BP neural network. *Appl. Soft Comput.* 99:106880. doi: 10.1016/j.asoc.2020.106880

Yu, X., Lei, Y., Yue, S., Wang, R. Research on intuitionistic fuzzy kernel clustering algorithm based on particle swarm optimization. *J. Commun.* (2015) 36, 78–84.

Zeng, P., Sun, F., Liu, Y., Wang, Y., Li, G., and Che, Y. (2021). Mapping future droughts under global warming across China: a combined multi-timescale meteorological drought index and SOM-Kmeans approach. *Weather Clim. Extrem.* 31:100304. doi: 10.1016/j.wace.2021.100304

Zhou, H., Yin, H., Li, Y., and Chai, Y. (2020). Multiview clustering via exclusive non-negative subspace learning and constraint propagation. *Informat. Sci.* 552:102–117. doi: 10.1016/j.ins.2020.11.037

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.