DSI DATA SCIENCE AND INFORMETRICS

RESEARCH ARTICLE

A novel features selection method based on improved clustering algorithm

Yifan Zuo, Yongxiang Xia*

School of Communication Engineering, Hangzhou Dianzi University, Hangzhou, Zhejiang, China

Abstract

Features dimensionality reduction technology has always played an important role in data mining. This paper makes a comparative study of features dimensionality reduction techniques, and proposes a new features selection method based on improved partial priority clustering algorithm (IPPCA). Firstly, selection method of the cluster center of the partial priority clustering algorithm (PPCA) is improved, so that the operation efficiency of the algorithm is improved, and the range of input data is expanded. Then, the clustering results are applied to features selection, so that the key feature set selected can retain the characteristics of the original dataset to a large extent. Finally, the above methods are simulated on four different data sets. The experiment shows that IPPCA not only has a high efficiency, but also the clustering effect is improved. Compared with principal component analysis (PCA) algorithm and independent component analysis (ICA) algorithm, the accuracy and precision of the key feature set obtained by the proposed features selection algorithm can reach more than 90% in data classification prediction.

Keywords

Cluster; Features selection; Partial priority clustering algorithm; Improved partial priority clustering algorithm; Big data

1 Introduction

With the development of information technology, people obtain more and more data everyday and the dimensions of data are also increasing. How to mine important information from large-scale data has important research significance (Jie et al., 2020). Clustering algorithms play an important role in data mining, but there are still some problems to be solved (Madhulatha, 2012). For example, scholars have proposed many improved algorithms for the scalability problem of clustering algorithms (Huang et al., 2012; Das et al., 2021). The BRICH clustering algorithm was proposed for data clustering, which has high operating rate and is suitable for large-scale data clustering (Nirmala & Thyagharajan, 2019). However, the BRICH algorithm is sensitive to the selection of initial data points, which leads to different clustering results. Though K-means algorithm has high running efficiency, the convergence of this algorithm depends heavily on the

*Corresponding author: xiayx@hdu.edu.cn

selection of initial clustering centers. So K-means++ algorithm was proposed to solve the initial clustering center selection problem of K-means algorithm, but the number of clusters cannot be automatically controlled and needs to be determined manually (Jie et al., 2020; Cen et al., 2013). The DBSCAN algorithm based on density clustering has poor clustering performance when the density is uneven and the clustering distance varies greatly. To address this shortcoming, Tang et al. (2021) proposed the OPTICS algorithm to provide an effective method to handle datasets with uneven density and reduces the sensitivity of clustering algorithms to input parameters. However, this algorithm is not suitable for processing large datasets. Guha et al. (2019) and Gong et al. (2022) proposed partial priority clustering algorithm (PPCA). This algorithm can ignore a small portion of data when large-scale datasets are clustered, and it greatly improves operational efficiency while sacrificing some accuracy to meet the speed requirements of big data clustering. However, this algorithm has strict requirements on the shape of clusters and can only process spherical or convex data structures.

In features dimensionality reduction algorithms, there are mainly two methods: features selection and features extraction (Benkessirat & Benblidia, 2019). The main steps of features selection methods are subset generation, subset evaluation, stopping criteria and result verification (Benkessirat & Benblidia, 2019). Asnaoui et al. (2021) firstly proposed the features clustering and then perform distributed features selection. The random forest algorithm (Ning et al., 2021) was a supervised learning algorithm, which improved the problem of overfitting compared to decision tree algorithms. The principle of features selection in this algorithm is to select the optimal features partition from the features subset through classification and regression. However, in many real-world cases, the collected data are not classified, so the random forest algorithm is not suitable for features selection in such data. Gong et al. (2022) proposed an algorithm based on partial priority clustering and clustering fusion. This algorithm aims to perform features selection for unclassified data. However, it only randomly selects key features from feature clusters based on difference in data distance without considering the proportion of normal data and abnormal data in the key feature set. In order to facilitate the expression, this phenomenon is named as the physical distribution characteristics of data in this paper. Features extraction algorithms refer to combining different features to obtain new features, which can achieve a mapping from the original feature space to a new feature space (Ayesha et al., 2020). A set of key features can be obtained by this way. Traditional features extraction algorithms include principal component analysis (PCA) algorithm, independent component analysis (ICA) algorithm, and linear discriminant analysis (LDA) algorithm. Both PCA (Wang et al., 2024) and ICA (Man et al., 2023) are linear features extraction algorithms that belong to unsupervised learning algorithms. They have good performance and obvious advantages in specific scenarios of features reduction, but PCA algorithm can only handle Gaussian distributed data and ICA algorithm is an extension of PCA algorithm. Hence, these two classic features extraction algorithms are only suitable for linear data features extraction. LDA algorithm is a supervised and classification-based features extraction method based on the principle of minimizing within-class variance and maximizing between-class variance after features projection (Karg et al., 2009). However, like the PCA algorithm, it can only handle Gaussian distributed data.

Due to the problem of PPCA in the cluster center selection mechanism, the running efficiency of

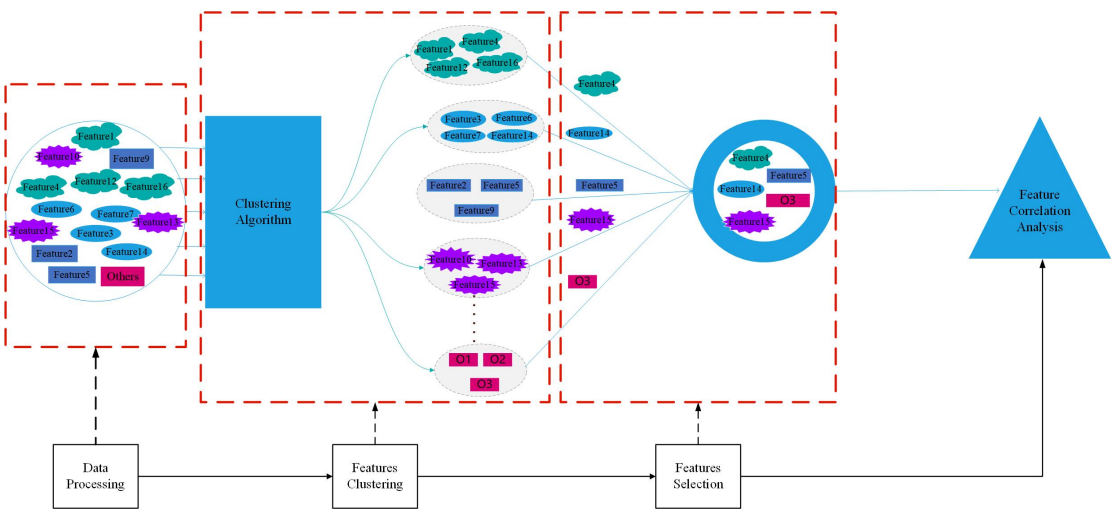
the algorithm is slow and the clustering effect may be poor. Therefore, this paper proposes IPPCA. IPPCA solves above problems, making it more suitable for big data clustering, and further solves the scalability problem of clustering algorithm. To solve the problem of insufficient accuracy and precision of data classification and prediction in features dimensionality reduction, this paper proposes a novel feature dimensionality reduction algorithm based on IPPCA. The key feature set selected by this algorithm can largely retain the characteristics of the original data set, and improve the accuracy and precision of data classification and prediction. The simulation experiments on four different types of data sets show that IPPCA not only has a high algorithm efficiency, but also significantly improves the clustering effect. Meanwhile, the key features selected by proposed algorithm have higher accuracy and precision than PCA and ICA algorithms in data classification prediction, and the experimental results are more than 90%.

2 Proposed method

In this section, we introduce the theoretical framework of the proposed algorithm, data processing, PPCA and its improvements, features selection algorithm based on clustering results, and features correlation analysis.

2.1 The framework

As shown in Figure 1, the algorithm mainly consists of four parts: data processing, features clustering, features selection, and correlation analysis. Firstly, each feature data is treated as an independent basic data unit and the dataset is processed to meet the effective form required for features clustering. Then, the dataset is clustered by the PPCA to obtain different feature clusters. The outliers of each feature data are marked by box plot method and the data anomaly is strictly defined from the perspective of the most beneficial features selection. Finally, correlation analysis based on the accuracy and precision of data classification prediction proves the effectiveness and reliability of the algorithm proposed in this paper. The effectiveness refers to the high operating efficiency of the algorithm, which is suitable for big data clustering. The reliability refers to the key features obtained through the algorithm can take into account both mathematical and physical distribution characteristics of the data.



Note: “Others” refers to other features, and “O1”, “O2” and “O3” represent a specific feature in “Others”.

Figure 1 Theoretical framework of the proposed algorithm

2.2 Data processing

With the rapid development of sensor technology and intelligent devices, automatic data collection may encounter data missing and anomalies due to sensor errors, communication transmission errors, and storage device failures. Therefore, it is necessary to observe and process the dataset to correct these data non-idealities caused by devices and transmission. Because the processing of these data in this paper is to facilitate the clustering algorithm, the data processing is relatively simple. And the specific processing methods are as follows. When encountering a sea of missing data within a moment dataset, it is advisable to remove that dataset to prevent its inclusion during clustering. This exclusion ensures that the clustering outcomes are not compromised by the incomplete data. When the data in the moment dataset is missing in a small range, the corresponding data can be deleted directly. There are two basic formats for traditional data storage. One data storage format is the measurement value of a certain feature, which denoted as $X = \{X_{Ti}\},\ i = 1, 2, \cdots n$. Where X_{Ti} represents the data corresponding to the feature Ti and the final dataset is composed of all features data. The other is that the measurement value of all features about a certain object, which denoted as $X = \{X_{T1}, X_{T2}, \cdots X_{Tn}\}$, and the final dataset is composed of these data. The algorithm proposed in this paper requires the input data to be in the first data storage format, so it needs to be converted into the first data storage format if the data are stored in the second data storage format. Finally, the data is normalized to enhance the comparability. The normalization formula is as follows

$$X^* = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{1}$$

where X^* represents the normalized data, X shows the original data, $X_{\min}$ indicates the minimum value of the data and $X_{\max}$ demonstrates the maximum value of the data.

2.3 Improved partial priority clustering algorithm (IPPCA)

This algorithm is an improvement of PPCA (Gong et al., 2022), which improves the operating rate of the original algorithm and further expands the scope of datasets that can be processed. The IPPCA is more suitable for processing large scale datasets. The pseudo-code of the algorithm is as follows:

Algorithm 1 Improved Partial Priority Clustering Algorithm

Input: The processed dataset D and initialize variables cs, ρ , r, k, r1, t.

Output: Get features clusters

```
for k=1 to cs by 1 do
    The median of the dataset is searched and used as the clustering center for clustering to obtain cluster C0.
    if isempty(C0) then
        The features data are divided into positive and negative groups and the mean values are calculated respectively, which are recorded as P and Q.
        if abs(P-Q) < r1 || P←0 || Q←0 then
            Cluster with P or Q as the clustering center to obtain cluster C.
```

```
7:              else
8:                  Cluster with P and Q as the clustering center respectively to
                    obtain cluster C1 and cluster C2.
9:              end if
10:         end if
11:         if size(C,1) < t then
12:             Perform PPCA to get cluster C3.
13:         end if
14:         if size(C1,1) < t and size(C2,1) < t then
15:             Perform PPCA to get cluster C4.
16:         end if
17:         if size(D,1) < ρ then
18:             break
19:         end if
20:     end for
```

In the above pseudo-code, ρ represents the lower limit of data volume in the dataset D. If size(D,1) < ρ , it indicates that most of the data have completed clustering and only the remaining small amount of data deviates from the overall trend. Hence, there is no need to continue clustering for these data. The parameter k is the number of clusters, cs represents the maximum number of clusters and r is the distance parameter. When $|X - C_i| \le r$, the data X is classified into the cluster of the clustering center C_i . The parameter t is the minimum data volume requirement of the effective cluster and r1 is the critical value of the absolute value of the difference between the positive and negative clustering centers.

In practice, due to the existence of vector features, the value of data can be positive or negative. If the number of positive and negative data is similar, most of the features data distribution will be significantly deviated from the mean line. Therefore, the average value of the sample is directly used as the clustering center to cluster in PPCA. Based on this setting, each cluster will only attract limited data, resulting in clustering failure. In order to deal with this problem, this paper proposes an improved algorithm based on PPCA. Firstly, the median is used as the clustering center for clustering. If the amount of data obtained in the cluster is too small, the average value of the dataset is calculated as the clustering center for clustering. The specific steps are as follows. To begin with, positive and negative data in features data are divided into two groups, and their mean values are calculated respectively, which denoted as P and Q. If the value of |P-Q| is greater than the set value, then P and Q are used as clustering centers for clustering separately. If the value of |P-Q| is within the set value range, then P or Q can be selected randomly as the clustering center for clustering. After clustering is completed, the data volume in the cluster is counted but whether to retain this cluster depends on the proportion of the data volume in the size of the dataset. If the data volume in the cluster is relatively large, the experimental results are

retained, otherwise the cluster is discarded and the clustering center is determined by the priority clustering algorithm to perform clustering again.

Experiments show that this method can improve the clustering effect limited by the data distribution structure in PPCA. PPCA has randomness in the selection of clustering centers, which may lead to a high complexity for clustering. Therefore, the IPPCA proposed in this paper improves the PPCA in the selection of clustering centers. Before randomly selecting the clustering center, the median or the mean of samples is used as the clustering center in turn to reduce the random clustering center selection problem in PPCA. Only if both the median and the mean of samples fail to be the clustering center, PPCA is enabled.

2.4 Features Selection Algorithm

Gong et al., (2022) mentioned that, in PPCA, the key feature set was obtained based on similar distance among the data of same feature. However, when there is a certain difference in the distance among data of the same feature, feature clusters cannot be obtained by analyzing the similarity of the proportion of data volume of features. Therefore, this method is largely limited by the influence of data distribution structure. The following improvements are made in this paper. From the clustering results of several simulation experiments, the variance of each effective cluster is calculated, then the cluster with the larger variance and data volume as the research object is selected. Because the deviation as an indicator represents the difference between data volume of each feature, the deviation of data volume of every feature in this cluster is calculated to form feature clusters. The calculation formula is as follows

$$IR = \frac{X - X'}{X'} \tag{2}$$

Where, *IR* represents the deviation degree, *X* is the feature data and X' demonstrates the mean value of the cluster. Through formula (2), the deviation degree between data volume of each feature and the mean value of this cluster is observed. The difference among data volume of features can be analyzed based on the mean value of the cluster as the similarity standard. According to the similarity principle of data volume of features, feature clusters are formed.

PPCA randomly selects a feature from each feature cluster as a key feature to form a key feature set. However, this features selection method only considers the difference of the mathematical characteristic among key features, ignoring the similarity of the physical distribution characteristics between key features and original dataset. In actual case, a sea of datasets are not directly classified after collection, so it is impossible to accurately discover important physical distribution characteristics of data from the datasets themselves. For unclassified datasets, this paper proposes a features selection algorithm that takes into account both the mathematical and physical distribution characteristics of the data. In order to grasp the actual abnormal data more accurately, the definition of abnormal data is strictly defined. In a data, as long as there is outlier of a feature data, the data is judged as abnormal data. For marking outliers of features data, the box plot method is used. Finally, considering both mathematical and physical distribution characteristics, features that retain the most outliers from the feature clusters are selected as the key features set. The pseudo-code of this part of the algorithm is as follows:

Algorithm 2 Features Selection Algorithm

Input: Valid clusters obtained by algorithm 1

Output: Key features set

1: **if** dataset D is moment dataset **then**

2: Calculate the variance of each cluster and arrange them in descending order, and select cluster with large variance and data volume.

3: All feature clusters can be obtained by analyzing the deviation between each feature data volume and the average data volume in this cluster.

4: **else**

5: Count the proportion of each feature data volume in each effective cluster, and get all feature clusters by analysis.

6: **end if**

7: **for** i=1 to size(D,2) by 1 **do**

8: **for** j=1 to size(D,1) by 1 **do**

9: Boxplot method mark outliers and plot the graph.

10: **end for**

11: **end for**

12: According to the outliers, the features are sorted in descending order. Then the key feature set is selected with combining the feature clusters, which is recorded as F.

13: Checkout the correlation between the key feature set F and the original dataset D.

In the above pseudo-code, the correlation between the key feature set and the original dataset is characterized by the accuracy and precision of regression prediction, and its calculation formula is as follows (Sun et al., 2021).

$$Accuracy = \frac{TP}{TP + FP} \tag{3}$$

$$Precision = \frac{TP + TN}{TP + FP + TN + FN} \tag{4}$$

Where, TP is the number of positive samples actually predicted as positive samples; FP is the number of negative samples actually predicted as positive samples; TN is the number of negative samples actually predicted as negative samples; FN is the number of positive samples actually predicted as negative samples.

3 Experimental results

3.1 Simulation preparation

The information of datasets is shown in Table 1. The simulated software environment is MATLAB 2021a and Python 3.8, and the hardware environment is Intel (R) Core (TM) i5-9300H CPU @ 2.40GHz 2.40 GHz and NVIDIA GeForce GTX 1650.

Table 1 Information of datasets

Dataset name	Number of dimensions	Number of samples
Transformer	23	954
ACline	7	1,155
Line	8	1,153
winequality-white	12	4,898

3.2 Comparison of experimental results for clustering algorithms

Datasets with different characteristics were used to demonstrate the clustering effect of PPAC and IPPAC. The datasets are mainly divided into two types. One is composed of multiple moment dataset including Transformer dataset and ACline dataset. The other is composed of dataset at a certain time including Line dataset and winequality-white dataset. The former is mainly used to mine key information about the entire network by studying the feature statistical value, while the latter is mainly used to mine key information at a single moment by studying the actual observed value of the entire network objects at that moment. The operating efficiency of each dataset under the PPAC and IPPAC is shown in Table 2.

Table 2 Operating efficiency of the algorithms (s)

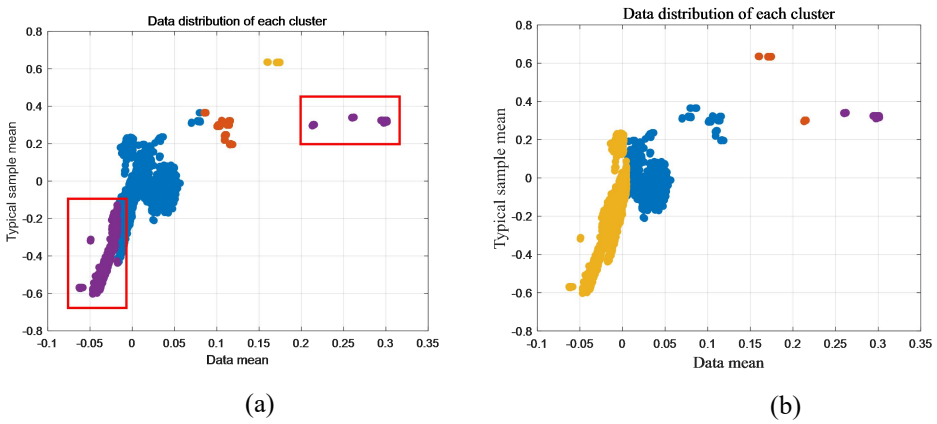
Datasets	Experiment 1		Experiment 2		Experiment 3		Experiment 4	
	PPCA	IPPCA	PPCA	IPPCA	PPCA	IPPCA	PPCA	IPPCA
Transformer	44.065	43.857	43.241	44.836	45.369	44.144	44.231	44.751
ACline	5.880	6.036	5.763	6.725	5.778	6.198	5.743	6.414
Line	9.566	8.768	8.706	8.499	8.586	8.863	9.003	8.708
winequality-white	336.191	313.856	340.450	319.062	340.518	315.009	342.597	322.672

Note: PPCA and IPPCA represent partial priority clustering algorithm and improved partial priority clustering algorithm respectively.

Transformer dataset has small differences between data of the same feature, but has significant differences among data of some different features. Some features data are vectorial, meaning that there are positive and negative values. Hence, the clustering results of Transformer dataset under the PPCA are poor, as shown in Figure 2(a). After adopting IPPCA, the clustering results are significantly improved, as shown in Figure 2(b). Similarly, the clustering results of ACline dataset under two clustering algorithms are shown in Figure 3, where Figure 3(a) represents the clustering effect of the PPCA, and Figure 3(b) represents the clustering effect of the IPPCA.

For first type of datasets, the average values of features for the selected moment datasets are calculated to generate a new dataset for study. The characteristics of this type of datasets are that the data values of the same feature are relatively similar and data volume distribution of some different features is uneven, which is suitable for mining the critical data information of the entire network. However, in the actual situation, sometimes it is necessary to study the dataset at a

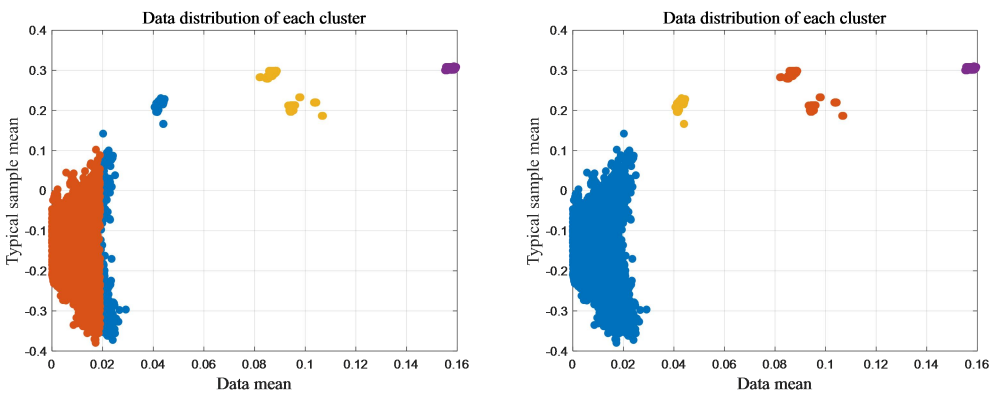
certain moment. Data with the same feature may exhibit significant differences. Therefore, it is of great research significance to verify whether clustering analysis can be used to mine key information from such datasets. Line dataset consisting of ACline data at a certain moment in the power system about a coastal city, and winequality-white dataset in the UCI (University of California, Irvine) database are studied to observe the improvement effect of the PPCA. Figure 4(a) and Figure 5(a) show the clustering effect of Line dataset and winequality-white dataset under PPCA algorithm respectively. Figure 4(b) and Figure 5(b) are the clustering effect of the above two datasets under IPPCA. Where, the typical data are data of a certain feature as reference to make the figure more visual.



(a) (b)

a: clustered under PPCA algorithm; b: clustered under IPPCA algorithm

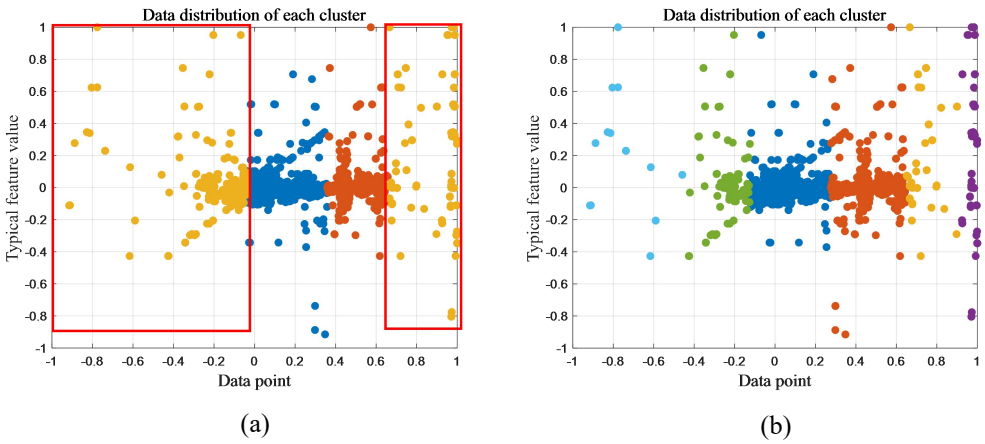
Figure 2 Comparison of clustering effect between two algorithms about Transformer dataset



(a) (b)

a: clustered under PPCA algorithm; b: clustered under IPPCA algorithm

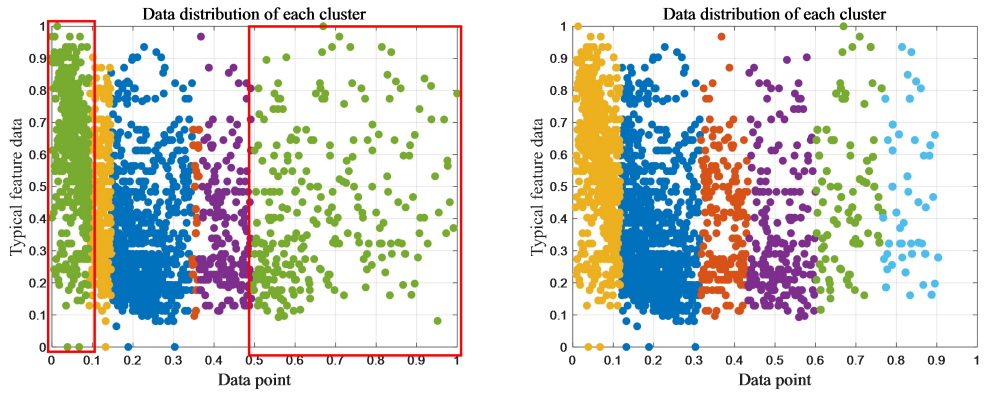
Figure 3 Comparison of clustering effect between two algorithms about ACline dataset



(a) (b)

a: clustered under PPCA algorithm; b: clustered under IPPCA algorithm

Figure 4 Comparison of clustering effect between two algorithms about Line dataset



(a) (b)

a: clustered under PPCA algorithm; b: clustered under IPPCA algorithm

Figure 5 Comparison of clustering effect between two algorithms about winequality-white dataset

3.3 Comparison of experimental results for features selection algorithms

In order to obtain the key features that retain the characteristics of the original dataset to the greatest extent, and make it more accurate and precise in data classification and prediction, this paper proposes a novel feature dimensionality reduction method based on improved clustering algorithm. Based on the above argumentation, the clustering results of IPPCA are stable, so any set of clustering results of the dataset can be taken. After clustering, it can be directly observed that there is a large gap among data volume of some features in Transformer dataset and ACline dataset. Transformer dataset has a large amount of data and many features, so Transformer dataset is taken as an example for analysis. Line dataset and winequality-white dataset have similar results after clustering. Due to the large amount of data and features in winequality-white dataset, which is taken as an example for research.

Firstly, Transformer dataset is clustered based on IPPCA, and the clustering results are shown in Table 3. For convenience of representation, CL1-CL4 represents the cluster name, CL1_P, CL2_P, CL3_P and CL4_P shows the proportion of data volume of features in the corresponding cluster, TA indicates the feature name, T1-T23 demonstrates the specific feature name, NS and OS represent the normal samples the outlier samples respectively. Then, the clustering results are

analyzed in Table 3 and the proportion of data volume of each feature in each cluster are calculated, as shown in CL1_P-CL4_P. Finally, every feature data is marked by the box plot method, and the key feature set is composed of features that can retain the physical distribution characteristics of the dataset to the greatest extent from every feature cluster based on the marking results. According to the selected key feature set, we predicted the accuracy and precision of data through formulas (3) and (4), and compared them with PCA and ICA algorithms, as shown in Table 4.

Secondly, winequality-white dataset is clustered based on IPPCA, and the clustering results are shown in Table 5, where CL1-CL8 represents the cluster name. Then, the results are analyzed to obtain the variance of each cluster, as shown in Table 5, where CL represents cluster name and VAR indicates variance of every cluster. The typical cluster are shown in bold in the table. The deviance degree of data volume of each feature in the cluster is calculated and also shown in Table 5. Finally, the key feature set is taken out through the similar operation to Transformer dataset. The accuracy and precision of the data classification prediction under different algorithms are calculated by formula (3) and formula (4), which is shown as Table 6.

Table 3 Clustering results of Transformer dataset

TA	CL1	CL1_P (%)	CL2	CL2_P (%)	CL3	CL3_P (%)	CL4	CL4_P (%)
T1	954	7.18	0	0.00	0	0.00	0	0.00
T2	954	7.18	0	0.00	0	0.00	0	0.00
T3	954	7.18	0	0.00	0	0.00	0	0.00
T4	0	0.00	0	0.00	954	14.67	0	0.00
T5	954	7.18	0	0.00	0	0.00	0	0.00
T6	954	7.18	0	0.00	0	0.00	0	0.00
T7	954	7.18	0	0.00	0	0.00	0	0.00
T8	952	7.16	0	0.00	2	0.03	0	0.00
T9	0	0.00	0	0.00	954	14.67	0	0.00
T10	28	0.21	0	0.00	926	14.24	0	0.00
T11	0	0.00	0	0.00	954	14.67	0	0.00
T12	147	1.11	0	0.00	807	12.41	0	0.00
T13	954	7.18	0	0.00	0	0.00	0	0.00
T14	0	0.00	0	0.00	954	14.67	0	0.00
T15	0	0.00	0	0.00	0	0.00	954	100.00
T16	954	7.18	0	0.00	0	0.00	0	0.00
T17	0	0.00	954	79.83	0	0.00	0	0.00
T18	954	7.18	0	0.00	0	0.00	0	0.00
T19	713	5.37	241	20.17	0	0.00	0	0.00
T20	954	7.18	0	0.00	0	0.00	0	0.00
T21	0	0.00	0	0.00	954	14.67	0	0.00
T22	954	7.18	0	0.00	0	0.00	0	0.00
T23	954	7.18	0	0.00	0	0.00	0	0.00
All	13,288	100.00	1,195	100	6,505	100.00	954	100.00

Note: CL1-CL4 represents the cluster name, CL1_P, CL2_P, CL3_P and CL4_P shows the proportion of data volume of features in the corresponding cluster, TA indicates the feature name, T1-T23 demonstrates the specific feature name, NS and OS represent the normal sample the outlier sample respectively.

Table 4 Analysis of the accuracy and precision of Transformer dataset under different features dimensionality reduction algorithms

Feature set obtained by methods	Normal samples			Outlier samples			True positive		
	Exp1	Exp2	Exp3	Exp1	Exp2	Exp3	Exp1	Exp2	Exp3
IPPCA	827	889	1,034	127	110	166	778	811	974
PCA	950	917	1,193	4	82	7	778	757	974
ICA	953	999	1,200	1	0	0	778	811	974

Feature set obtained by methods	True negative			Accuracy (%)			Precision (%)		
	Exp1	Exp2	Exp3	Exp1	Exp2	Exp3	Exp1	Exp2	Exp3
IPPCA	127	110	166	94.07	91.23	94.20	94.86	92.19	95.00
PCA	4	28	7	81.89	82.55	81.64	81.97	78.58	81.75
ICA	1	0	0	81.64	81.18	81.17	81.66	81.18	81.17

Note: Exp1, Exp2 and Exp3 indicates Experiment 1, Experiment 2 and Experiment 3 respectively. IPPCA, PCA and ICA represents improved partial priority clustering algorithm, principal component analysis and independent component analysis, respectively.

Table 5 Clustering results of winequality-white dataset

TA	CL1	CL2	CL3	CL4	CL5	CL6	CL7	CL8	IR(CL1)
T1	**3,054**	1,560	53	223	5	1	1	1	0.35
T2	**3,528**	329	900	115	15	4	5	2	0.56
T3	**4,289**	193	317	66	6	0	26	1	0.90
T4	**1,542**	6	3,177	2	0	0	170	1	-0.32
T5	**1,367**	43	3,431	46	6	3	1	1	-0.40
T6	**2,188**	21	2,673	4	0	0	11	1	-0.03
T7	**2,800**	1,535	80	471	7	2	2	1	0.24
T8	**2,782**	0	2,105	2	0	0	8	1	0.23
T9	**992**	1,698	14	1,717	410	50	1	16	-0.56
T10	**2,627**	1,336	171	579	147	33	1	4	0.16
T11	**1,808**	931	113	1,111	700	214	2	19	-0.20
T12	**163**	1,457	0	2,198	880	175	20	5	-0.93
All	**27,140**	9,109	13,034	6,534	2,176	482	248	53	—
VAR	**1,325,594**	516,442	1,840,531	552,943	96,354	5,516	2,279	40	—

Note: TA indicates the feature name; CL1-CL8 represents the cluster name; IR(CL1) shows deviation degree of each feature in CL1; T1-T12 demonstrates the specific feature name; VAR represents variance; the symbol ' - ' means that the value here is meaningless, so no calculation is done; bold content indicates typical cluster.

Table 6 Analysis of the accuracy and precision of winequality-white dataset under different features dimensionality reduction algorithms.

Feature set obtained by methods	Normal samples	Outlier samples	True positive
IPPCA	3,511	1,387	3,511
PCA	4,602	296	3,781
ICA	4,569	329	3,761
Feature set obtained by methods	True negative	Accuracy (%)	Precision (%)
IPPCA	1,050	100	93.12
PCA	229	82.16	81.87
ICA	242	82.32	81.73

Note: IPPCA indicates improved partial priority clustering algorithm; PCA indicates principal component analysis; ICA indicates independent component analysis respectively.

4 Discussion

4.1 Advantages of IPPCA

When adopting PPCA, there is a certain probability that the clustering performance will be poor. In our simulation, as shown in the red rectangular box in Figure 2(a), a large number of features data cannot be clustered. In this experiment, 3,665 features data cannot be clustered in the end. Further observation reveals that the actual reason why some features data in the figure cannot be clustered is that some of the residual data is positive and the other part is negative. If the typical sample is taken from the remaining features data directly and the mean value is regarded as the clustering center, the positive and negative data will be offset in the calculation process. Therefore, the clustering center will deviate from the positive and negative data, which will lead to clustering failure. Based on this reason, this paper proposes first using the median as the clustering center for clustering. For data that cannot be clustered, they are divided into positive and negative groups, and the clustering centers are obtained by calculating the mean value of groups respectively. If there are still a certain amount of features data that cannot be clustered, PPCA will be performed, which ensures that features data are maximally divided into corresponding clusters. When the parameters are the same, Transformer dataset is clustered through the IPPCA, and the clustering results obtained are significantly improved compared with that of the PPCA. There will be no case that a large number of features data cannot continue clustering.

In both Figure 3(a) and Figure 3(b), because the features data distribution in ACline dataset is relatively concentrated and a significant amount of positive and negative data in the dataset does

not exist, there is no large amount of features data that cannot be clustered. Therefore, whether it is PPCA or the IPPCA, the clustering results obtained are similar. But the difference in the selection of clustering center will lead to difference in the final clustering results.

For Line dataset and winequality-white dataset, it can be observed that there are a large number of features data cannot be clustered due to features data distribution structure. These data are shown in the red box in Figure 4(a) and Figure 5(a). However, IPPCA can improve clustering results, which are shown in Figure 4(b) and Figure 5(b).

The operating time of the four datasets under PPCA and IPPCA was counted through four simulation experiments, and time unit is seconds. The results are shown in Table 2. It can be observed that the operating time of the proposed IPPCA and PPCA is similar when the size of the data set is not large. However, as the size of the data set increases, the operating time of IPPCA is a certain degree faster than that of PPCA. Therefore, IPPCA is more suitable for information mining in big data.

4.2 The selected key features of datasets

In Table 3, data volume of all features about Transformer dataset in each cluster are summarized. It is observed that data volume of some features in every valid cluster have significant differences but others are similar. Therefore, feature clusters can be analyzed by calculating the proportion of data volume of features in the cluster. According to the proportion of data volume of each feature in Table 3, it can be analyzed that T1, T2, T3, T5, T6, T7, T8, T13, T16, T18, T19, T20, T22, T23 belong to one cluster, T4, T9, T10, T11, T12, T14, T21 belong to one cluster and T15 as well as T17 belong to a separate cluster. That is, features about this dataset are divided into four feature clusters. Where, a key feature is selected in each cluster to characterize this cluster and the final selected key feature set can take into account the characteristics of T12 and T19. Therefore, even if the data of T12 and T19 exist in two clusters at the same time, they do not need to be separately classified as a cluster. Datasets composed of data at different moment, due to data clustering based on statistical value of features, the final processed dataset exhibits similar characteristics after clustering. That is, in the same cluster, the data volume distribution of some features is relatively concentrated, and that of others is divergent. The feature clusters can be obtained based on the proportion of data volume of each feature in the cluster.

In order to take into account both the mathematical and physical distribution characteristics of datasets, the key feature set is finally selected, which has a large gap in distance and can retain the physical distribution characteristics of the original datasets to the greatest extent. Therefore, T8, T10, T15 and T17 as the key feature set are selected. Table 4 summarizes the accuracy and precision of Transformer dataset under three feature reduction algorithms in three different simulation experiments. The number of samples selected in each experiment was different. Experiment 1 randomly selects 954 samples, Experiment 2 selects 999 samples, and Experiment 3 selects 1,200 samples. When all features are involved in the determination of normal samples and abnormal samples, the number of normal samples in the three experiments is 778, 811, 974 and the number of abnormal samples is 176, 188, 226. It can be clearly observed that under the same number of key features, the accuracy and precision of the key feature set obtained by the algorithm proposed in this paper in predicting data classification are significantly higher than the that of PCA algorithm and ICA algorithm.

In Table 5, the data volume of all features in each cluster about winequality-white dataset is summarized. It is observed that the distribution characteristics of this dataset are significantly different from those in Transformer dataset. The characteristic of this dataset after clustering is that there is a certain data volume for every feature in each cluster, and the feature clusters cannot be obtained by directly analyzing the proportion of data volume of features in feature clusters, so it is particularly important to find the method to analyze and obtain feature clusters. In Section 3.3, the method is mentioned. The cluster with the largest data volume and maximum variance is the most representative, followed by the cluster with larger variance and maximum data volume is also typical. According to the above principle of screening the typical cluster and combined with the size of the variance in Table 5 and data volume in the cluster, the first cluster in Table 5 is selected as a typical cluster. Then the deviation about data volume of each feature in the cluster is calculated and results are shown in Table 5. Finally, feature clusters are obtained through DBSCAN (Luo et al., 2023) clustering algorithm or direct analysis. According to the experimental analysis, T4, T5, T6, T9, T11 are a cluster, T1, T7, T8 are a cluster and T2, T3, T12 are a cluster alone. That is, features about this dataset are divided into five feature clusters.

After experimental analysis, T1, T2, T3, T9 and T12 are finally selected as the key feature set. According to the abnormal data determination method proposed in Section 2.4, when all features are involved in the determination of normal samples and abnormal samples in winequality-white dataset, the number of normal samples is 3,848, and the number of abnormal samples is 1,050. The accuracy and precision of the data prediction results are shown in Table 6, which are obtained by repeating the steps of selecting physical features in Transformer dataset. It can be observed that the features selection algorithm proposed in this paper has higher accuracy and precision compared to PCA algorithm and ICA algorithm in winequality-white dataset in the UCI database.

Features selection algorithm proposed in this paper not only adds new content in data mining, but also has better experimental results than traditional algorithms. The analyzed Transformer dataset and winequality-white dataset are representative of the two types of datasets mentioned in section 3.2, so this algorithm is universal to a certain extent. From the analysis of the experimental results, it can conclude that the features selection method based on the IPPCA not only solves problems of the scalability problem and the input data range in the clustering algorithm, but also improves the clustering effect. At the same time, the key feature set obtained can retain the characteristics of the original data set to the greatest extent, and has higher accuracy and precision than the traditional algorithm in data classification and prediction.

5 Conclusions

This paper proposes a new features dimensionality reduction algorithm, namely the features selection method based on IPPCA. Compared to traditional PPCA, this algorithm improves the selection of clustering center, which is useful under the condition where a large number of features data cannot be clustered due to the influence of data distribution structure. Moreover, it can improve the operating rate of PPCA to some extent as the size of the dataset increases. When performing the key features selection based on the IPPCA, the obtained feature set retains the physical distribution characteristics of the original data set as much as possible, which greatly reduces the data redundancy and achieves better data compression effect. The simulation results

on different datasets indicate that the proposed IPPCA can improve the clustering effect of PPCA. At the same time, it also improves the operating rate of the algorithm as the size of the dataset increases. Compared to PCA algorithm and ICA algorithm, the proposed features selection algorithm exhibits higher accuracy and precision.

References

Asnaoui, Y., Akhiat, Y., & Zinedine, A. (2021). Feature selection based on attributes clustering. *2021 Fifth International Conference On Intelligent Computing in Data Sciences (ICDS)*, 1-5. doi: 10.1109/ICDS53782.2021.9626770

Ayesha, S., Hanif, M. K., & Talib, R. (2020). Overview and comparative study of dimensionality reduction techniques for high dimensional data. *Information Fusion, 59*, 44-58. doi: 10.1016/j.inffus.2020.01.005

Benkessirat, A., & Benblidia, N. (2019). Fundamentals of feature selection: An overview and comparison. *Proceedings of IEEE/ACS International Conference on Computer Systems and Applications(AICCSA)*, 1-6. doi: 10.1109/AICCSA47632.2019.9035281

Cen, Z. Y., Li, B., & Tian, L. F. (2013). High dimensional transfer function design based on k-means ++ for volume visualization. *Journal of Computer Applications, 32*(12), 3404-3407. doi: 10.3724/SP.J.1087.2012.03404

Das, J., Majumder, S., & Mali, K. (2021). Clustering techniques to improve scalability and accuracy of recommender systems. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 29*(4), 621-651. doi: 10.1142/S0218488521500276

Gong, X., Zuo, Y., Zhang, Y., Chen, M., & Tu, H. (2022). Key features selection of power system operation via improved clustering algorithm. *2022 IEEE Asia Pacific Conference on Circuits and Systems(APCCAS),* 25-29. doi: 10.1109/APCCAS55924.2022.10090368

Guha, S., Li, Y., & Zhang, Q. (2019). Distributed partial clustering. *ACM Transactions on Parallel Computing, 6*(3), 1-20. doi: 10.1145/3322808

Huang, L. , Wang, J. , & He, X. (2012). A graph clustering algorithm providing scalability. *Journal of Networks, 7(2)*, doi: 10.4304/jnw.7.2.229-235

Jie, C., Jiyue, Z., Junhui, W., Yusheng, W., Huiping, S., & Kaiyan, L. (2020). Review on the research of K-means clustering algorithm in big data. 2*020 IEEE 3rd International Conference on Electronics and Communication Engineering (ICECE)*, 107-111. doi: 10.1109/ICECE51594.2020.9353036

Karg, M., Jenke, R., Seiberl, W., Kühnlenz, K., Schwirtz, A., & Buss, M. (2009). A comparison of PCA, KPCA and LDA for feature extraction to recognize affect in gait kinematics. *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops(ACII)*, 1-6. doi: 10.1109/ACII.2009.5349438

Luo, J., Liao, X., Wang, Y., Zhang, J., Yu, Z., Wang, G., & Li, X. (2023). An entropy-based adaptive DBSCAN clustering algorithm and its application in THz wireless channels. *IEEE Transactions on Antennas and Propagation, 71*(12), 9830-9837. doi: 10.1109/TAP.2023.3326924

Madhulatha, T. S. (2012). An overview on clustering methods. *IOSR Journal of Engineering, 2*(4), 719-725. doi: 10.9790/3021-0204719725

Man, X., Wang, W., Wang, X., & Wang, Y. (2023). Risk detection system of bridge construction based on fast ICA algorithm. *Lecture Notes on Data Engineering and Communications Technologies, 169*, 36–44. doi: 10.1007/978-3-031-28893-7_5

Ning, F., Cheng, Z., Meng, D., & Wei, J. (2021). A framework combining acoustic features extraction method and random forest algorithm for gas pipeline leak detection and classification. *Applied Acoustics, 182*, 108255. doi: 10.1016/j.apacoust.2021.108255

Nirmala, G., & Thyagharajan, K. K. (2019). A modern approach for image forgery detection using brich clustering based on normalised mean and standard deviation. *2019 IEEE International Conference on Communication and Signal Processing(ICCSP)*, 441-444. doi: 10.1109/ICCSP.2019.8697951

Sun, Z. , Wang, Z. , Chen, Y. , Liu, P. , & Dorrell, D. G. (2021). Modified relative entropy based lithium-ion battery pack online short circuit detection for electric vehicle. *IEEE Transactions on Transportation Electrification, 8*(2), 2332-7782. doi: 10.1109/TTE.2021.3128048

Tang, C. H., Wang, H., Wang, Z. W., Zeng, X. K., Yan, H. R., & Xiao, Y. J. (2021). An improved OPTICS clustering algorithm for discovering clusters with uneven densities. *Intelligent Data Analysis, 25*(6), 1453-1471. doi: 10.3233/IDA-205497

Wang, T., Xie, Y., Jeong, Y., & Jeong, M. K. (2024). Dynamic sparse PCA: a dimensional reduction method for sensor data in virtual metrology. *Expert Systems with Applications, 251*, 123995. doi: 10.1016/j.eswa.2024.123995