Improvement of Key Feature Mining Algorithm for Sports Injury Data Based on LOF Enhanced K-Means and Sparse PCA

Tanwei Shang
Basic Science Department of Wuchang Shouyi University Wuhan 430000, China
E-mail: Shangtanwei58643@163.com

Keywords: sports injury, key characteristics, feature mining, LOF algorithm, principal component analysis algorithm

Received: September 24, 2024

Sports injury not only affects the health of athletes, but also has a negative impact on their sports performance and competitive level. By mining the key features of sports injury data, we can identify the key factors that affect athletes' performance, so as to improve sports performance and competitive level. Therefore, this paper proposes an improved key feature mining algorithm for sports injury data based on LOF enhanced k-means and sparse principal component analysis. The basic probability assignment method of attribute weight is used to assign the damage data, which provides a neat and consistent data basis for the subsequent key feature mining of sports injury. The K-means algorithm improved by LOF algorithm is used to classify the assignment results and divide the sports injury data. PCA is used to reduce data dimensions, simplify redundancy, and enhance the independence of sports injury data features. Using reweighted sparse PCA to realize key feature mining of sports injury data. The experimental results show that the proposed method can accurately capture the essential differences between non sports injury data and sports injury data, and accurately divide non sports injury data and sports injury data. At the same time, the average absolute percentage error and root mean square error of the assessment accuracy of injury factor assignment are both lower than 0.1, and the DBI values of all samples are not more than 0.13, it can effectively mine the key features of sports injury data.

Povzetek: Raziskava predlaga izboljšan algoritem za iskanje ključnih značilnosti podatkov o športnih poškodbah z združevanjem dveh algoritmov.

1 Introduction

Sports injury refers to the injury of muscles, ligaments, bones and other tissues in the process of sports activities, including sprains, strains, fractures and other types. The occurrence of these injuries is often related to many factors, such as age, gender, physical condition, training level, sports events, etc. Sports injury data plays an important role in sports medicine, rehabilitation training, athlete training management and other aspects [1]. Such data includes athletes' physiological parameters, sports training records, living habits, training data, medical records, athletes' feedback, medical data and historical injury data [2]. Based on these data, we can find out the key factors related to sports injury, such as excessive training, unreasonable exercise intensity and frequency, improper technical actions, etc; And understand the development trend of sports injury of specific groups or individuals, so as to provide clues for prevention and intervention [3]. This helps to formulate more targeted preventive measures to reduce the occurrence of sports injuries. However, this kind of data has the characteristics of huge data volume, data format and data category differences, and high data dimensions, which lead to low data utilization efficiency and inability to accurately and efficiently obtain the key features of massive data.

Therefore, many scholars have carried out research on it. Abualigah et al. proposed a new feature selection model combining the sine cosine algorithm and genetic algorithm, and screened out the most informative and important features by identifying and eliminating the redundancy, noise and attributes not directly related to the task in the original data set [4]. However, the model involves multiple parameters, including crossover probability, mutation probability, population size, iteration times, etc. in genetic algorithm, as well as specific parameters that may be involved in sine cosine algorithm. Tuning these parameters is critical to model performance, but it can also be a time-consuming and complex task. Inappropriate parameter setting may lead to poor performance of the model or local optimal solution. Kalaivani et al. obtained the data set from UCI machine learning database, and screened the optimal feature subset from the original data set through multiple feature selection algorithm combined with the evaluation criteria of autocorrelation and information gain [5]. However, the information gain has the disadvantage of preferring to select data attributes with a large number of values, resulting in a large value of information gain, which does not necessarily mean that the attribute is a key feature. Shehab et al. proposed an unbalanced data mixed feature selection cloud model based on k nearest

neighbor algorithm, which uses feature subset selection preprocessing to reduce data complexity and realize data feature mining [6]. However, when the model faces high-dimensional data, the distance calculation between samples becomes complex and it is difficult to accurately reflect the similarity between samples, which may lead to the decline of the model's feature selection performance. Tan et al. proposed a rock-climbing key point detection algorithm based on an improved hourglass network. By designing a multi-channel pooling residual structure and introducing an hourglass attention structure, the algorithm solved the problems of variable target scales and feature adaptability, improved the performance of attitude estimation methods, and verified its effectiveness and generalization ability on multiple datasets [7]. Although the multi-channel pooling residual structure aims to improve the limitations of information loss and insufficient context extraction caused by multiple upsampling and downsampling in hourglass networks, the design process may face challenges such as how to balance the information fusion of different pooling paths and how to avoid information redundancy or loss.

Data mining algorithms refer to a set of heuristic methods and calculation processes for creating data mining models based on data [8]. These algorithms search and extract specific patterns, trends and statistical information from the data through in-depth analysis of the data provided, thus providing valuable insight and decision support for data users. Typical data mining algorithms include clustering analysis, association rules, principal component analysis, etc. Among them, principal component analysis is to use orthogonal transformation to transform the observation data represented by linear dependent variables into a few data represented by linear independent variables. These linear independent variables are called principal components. Specifically, all high-dimensional data points are converted to a new coordinate system by projection mapping, and the dimension of this coordinate system is less than or equal to the dimension of the original coordinate system. At the same time, the standard for finding this new coordinate system is that in the new coordinate system, the variance sum of the data on each coordinate axis corresponding to the projected data is the largest, so the saved data is the most complete. Therefore, this paper proposes an improved key feature mining algorithm for sports injury data.

In summary, the relevant research summary table is shown in Table 1.

Table 1: Research summary table

Existing research	Defective nature	Improvements in this paper
Abualigah et al. [4] proposed a hybrid feature selection method, SCAGA, which combines sine cosine algorithm (SCA) and genetic algorithm (GA). The method utilizes the UCI machine learning warehouse dataset and evaluates key performance indicators such as classification accuracy, worst fitness, average fitness, best fitness, average feature count, and standard deviation. The results show that SCAGA performs better in balancing the exploration and utilization strategy of search space, and achieves the best overall performance on the test dataset compared to basic SCA and other related methods such as ant lion optimization and particle swarm optimization.	Improper parameter settings in the sine cosine algorithm may result in poor model performance or local optima.	The attribute weight probability basic assignment method was adopted to assign values to the damage data, which provides a more concise and consistent data foundation for subsequent feature mining, thus avoiding performance problems caused by improper algorithm parameter settings. In addition, by adopting the improved K-means algorithm and principal component analysis method based on LOF, this paper further improves the accuracy and efficiency of data processing.
Kalaivani et al. [5] used data mining classification methods such as KNN, SVM, and decision trees to predict a heart disease dataset containing 282 observations and 75 attributes from the UCI machine learning warehouse. They also utilized the Multi Feature Selection Algorithm (MFSA) combined with autocorrelation and information gain for feature selection	However, the disadvantage of information gain is that it tends to select data attributes with a large number of values, resulting in a large number of values for information gain, which does not necessarily mean that the attribute is a key feature.	We adopted methods such as weighted sparse principal component analysis, combined with the actual situation of the data and the importance of features, to conduct more comprehensive and accurate feature selection. This can effectively avoid the limitations of information gain and improve the accuracy and effectiveness of feature selection.

to improve classifier performance.		
Shehab et al. [6] proposed a novel hybrid feature selection cloud model based on k-nearest neighbor algorithm for handling imbalanced data. The model combines firefly distance measurement and Euclidean distance, and exhibits good performance compared to simple weighted nearest neighbors, effectively improving classification accuracy and reducing processing time through cloud distributed models.	However, when the model faces high-dimensional data, the distance calculation between samples becomes complex and difficult to accurately reflect the similarity between samples, which may lead to a decrease in the model's feature selection performance.	This paper uses principal component analysis to reduce the dimensionality of data, simplify redundant information, and enhance the independence of data features. This method can effectively reduce the complexity of data, reduce the difficulty of calculating the distance between samples, and improve the efficiency and accuracy of feature selection. Meanwhile, by using the improved K-means algorithm with LOF for data classification and segmentation, this paper further improves the accuracy and stability of data processing.
Tan et al. [7] proposed a rock climbing keypoint detection algorithm based on an improved hourglass, which uses a multi-channel pooling residual structure and hourglass attention structure to improve keypoint detection performance. Its effectiveness was verified on MPII, COCO, and rock-climbing datasets, with key performance indicators including detection accuracy and algorithm generalization ability.	Although the multi-channel pooling residual structure aims to improve the limitations of information loss and insufficient context extraction caused by multiple upsampling and downsampling in hourglass networks, the design process may face challenges such as how to balance the information fusion of different pooling paths and how to avoid information redundancy or loss.	Similar ideas and methods were adopted. By using weighted sparse principal component analysis and other methods, this paper has achieved key feature mining of sports injury data, which can be seen to some extent as an alternative or supplement to the multi-path pooling residual structure. Meanwhile, the method proposed in this paper places greater emphasis on the integrity and consistency of data, avoiding challenges such as information fusion and redundancy.

2 Mining key features of sports injury data

In order to accurately extract key features from massive data, effectively process and classify complex sports injury data, and achieve prediction and prevention of sports injuries. Using the basic probability allocation method of attribute weights, diverse sports injury data is transformed into a unified format, and the contribution of each attribute in the injury data is quantified. Subsequently, based on the improved K-means algorithm and combined with the LOF algorithm to handle outliers, the quantified data was accurately classified with the aim of revealing potential structures and patterns in the data. Finally, principal component analysis (PCA) combined with LASSO regression model is used to reduce the dimensionality and extract key features of the classified data, in order to reduce data complexity and improve feature independence. This series of methods aims to construct an efficient sports injury data analysis and prediction model, providing scientific basis for athletes, coaches, and medical personnel to accurately assess injury risk and develop effective intervention measures, thereby improving the health level and competitive performance of athletes.

2.1 Damage data assignment

Sports injury data typically encompasses a wide range of information, including physical parameters, exercise intensity, duration, type of exercise, and environmental conditions. The diversity and complexity of these data pose challenges in directly extracting key features from them [9]. By adopting a reasonable allocation strategy, various types of data can be converted into a unified format, simplifying the complex sports injury data into a more manageable form.

In this paper, the basic probability assignment method of attribute weight is used to assign damage data. The basic probability assignment method of attribute weight can assign different weights according to the importance of attributes, so as to quantify the contribution of different attributes in the damage data. This method helps to make better use of key attributes and ignore or reduce the impact of non key attributes in the subsequent data mining process. Attribute weight is denoted by w_j, the calculation formula is:

$$w_j = \frac{\hat{w}_{ij}}{y_A + y_B} \tag{1}$$

Where: $\hat{w}_{ij}$ represents the preweight of the attribute in the j th attributeof the i th data category; y_A means that athletes become "people with injury tendency"; y_B means that athletes become the "injury prone group", and the risk of specific injuries of athletes in different sports will be greatly increased.

The injury factor data was quantified according to the w_j calculated by the above formula, $w_j > 0.51$ indicates the internal injury causing factor; $w_j < 0.5$ indicates the external injury causing factor.

After the division of internal and external damage factors, the assignment details ofthe external damage data set y_A and internal damage data y_B are shown in Table 2.

The factor assignment in Table 2 is carried out. If an athlete has no previous injury, the factor is assigned 0, the option with the smallest impact on sports injury is assigned 1, and the option with the largest impact is assigned 3. Assigning values to sports injury data can transform different types of injuries to injury factors into a unified measurement standard, effectively solve the problems such as missing values, abnormal values or inconsistent data formats contained in the original sports injury data, so that different data can be compared and analyzed, and quickly identify the key factors related to injury occurrence, it provides a basis for key feature mining of subsequent sports injury data.

Table 2: Details of damage data assignment

Weight assignment	External damage to injury factors	Internal to injury factors
0	No previous injuries	No previous injuries
1	Minor damage	Minor damage
2	Obvious damage	Obvious damage
3	Serious injury	Serious injury

2.2 Motion data classification based on improved K-Means algorithm

Although the original data has been quantified or coded, the assigned data may not be clearly divided into different categories or groups [10]. For sports injury data, different types of sports, injury degrees, recovery stages, etc. may need to be further classified to more accurately analyze data characteristics and laws [11]. K-means algorithm is a distance based clustering algorithm, which can divide the samples in the dataset into K clusters, making the samples in the same cluster more similar, while the samples between different clusters are less similar. This clustering analysis method helps to find potential structures and patterns in the data, and provides valuable information for sports injury prediction. However, K-means algorithm is vulnerable to the problem of data imbalance. When the number of samples of one class in the data set is far more than that of other classes, the clustering effect may be affected. LOF algorithm can identify outliers in the dataset. By calculating the local outlier factor (LOF) of each data point, the LOF algorithm can evaluate the degree of anomaly of the point relative to its local neighborhood. If the LOF value of a point is far greater than 1, it is considered as an outlier. Removing these outliers before clustering or carrying out special processing can reduce their impact on the K-means clustering process, thus improving the accuracy of clustering. The algorithm implementation process is described as follows:

Inputs: The original motion dataset, density threshold, and number of outlier points, respectively, are represented by, $X = [x_1, x_1, ..., x_N]$ 、 Ω 、 n ;

Output: Top ranked n larger outlier factor value object.

(1) Construct g grids for detecting motion datasets based on a variable gridding strategy.

The grid space was determined as the clustering region, the dimensions of the motion data space were divided using the same size of spacing, and then similar interval segments of the same dimension were merged to obtain the space based on the grid division. Using the fast-ranking method to arrange the original motion data set of the i th dimensional data, and calculate the similarity of neighboring interval segments, after the extraction is completed, judge the similarity of neighboring interval segments in the i th dimensional space [12]. Repeatedly perform this step to get the result

of merging similar intervals in different dimensions and output this result.

(2) The number of motion data points for each grid cell is obtained:

Define the number of data points is, compare the density threshold and the number of data points size, but the number of data points is greater than the density threshold, it means that this data belongs to the sports injury data, when all the sports data to calculate the outlier factor can be terminated, and record the results of the sports injury data, the outlier factor $L(x_i)$ of the motion data to be detected is calculated as follows:

$$L(x_i)=\Omega\sum_{k=1}^{k}\frac{g_k\left(x_i^m\right)}{g_k(x_i)} \quad (2)$$

Among them, $H_k(x_i)$ denotes the set of neighborhoods of a sample of sports injury data x_i, k represents the k th set of neighbors, the local outlier is the set $H_k(x_i)$ in the sample, the mean of the localized accessible density ratios of the sample x_i; x_i^m represents the m near-neighbor samples, m represents the m neighborhood sets. $g_k(x_i)$ is the localized reachable density for the sample x_i. The degree of outliers for sample x can be determined by $L(x_i)$.

(3) Traversing to a sports injury data point requires elimination:

After elimination, we continue to iteratively calculate the data points for detecting sports injuries, and finally arrange the outlier values in the order of largest to smallest, and save the outlier values of the top part. After eliminating the outliers, a more accurate clustering center was determined according to the maximum and minimum distance criterion as follows:

After the Euclidean distance is calculated according to the maximum minimum distance algorithm, the sample points are divided into each cluster center according to the nearest neighbor principle. This algorithm is different from the traditional K-means clustering algorithm in determining the center point strategy, clustering categories K is not an empirical setup, but the following strategy is followed:

(a) Determine the initial cluster center:

Selecting an object x_i in the sample points of the motion data, which is used as the first clustering center, to find the Euclidean distance of all data points with that clustering center x_i, the method is shown in equation (3):

$$d(x_a,x_b)=L(x_i)\sqrt{\sum_{k=1}^{k}\left(x_{ak}-x_{bk}\right)^s} \quad (3)$$

Of which: x_a、x_b all represent samples, s denotes the spatial dimension, the Euclidean distance between two samples in that space is denoted by $d(x_a,x_b)$.

(b) Based on the result of $d(x_a,x_b)$, the new clustering centers are obtained, The data points with the largest European distance from x_i are classified in the same data set; The remaining moving data points are calculated, and the Euclidean distance between each data sample point and the initial center point x_i, and the sample point corresponding to the maximum value is still determined, which is divided into the same category.

Based on the above strategy of cyclic operation, complete the classification of all the sports data, and stop updating the clustering center when no more new clustering centers are generated, and get the effective clustering center K. The steps are as follows:

Step 1: In $(0,1)$, a value is selected in the interval and given, with ϕ denotes that the initial clustering center F_1 is generated at this moment;

Step 2: Cluster center update strategy.

Finding the Euclidean distance between different points and F_1 is expressed by d_{i1}, and the new cluster center F_2 is x_k corresponding to $d_{k1}=\max\{d_{in}\}$; Then, find the 3rd clustering center, find the distance between the first two clustering centers and different points defined as

d_{i1}、d_{i2}, the distance between the first two clustering centers is defined as d_{12}, when $d_n = \max\{\min(d_{i1}, d_{i2})\}$ and $d_n > \phi \times d_{12}$ with the situation of $(i = 1, 2, \mathrm{L}, n)$, the 3rd clustering center F_3 is x_l.

After determining the existence of a third clustering center, determine whether the current situation is consistent with $d_j = \max\{\min(d_{i1}, d_{i2}, d_{i3})\}$ and $d_n > \phi \times d_{12}$, verifying that a 4th clustering center currently exists. Determine whether the next clustering center exists according to the above derivation strategy, the conditions for terminating the updating of the new cluster centers is $d_n \le \phi \times d_{12}$.

Step 3: Summarize cluster centers:

The above algorithm organically combines the maximum-minimum clustering criterion and the local outlier detection algorithm to accurately determine the clustering center Z_i. The type of motion data is classified by finding the distance of each motion data from the center point with the following equation:

$$d_{x_i,Z} = \frac{d(x_a, x_b) d_n}{d_j} \quad (4)$$

$$Z_i = \arg\min \|Fr_i\| \quad (5)$$

Where, the distance between the two-motion data and the set of data types are denoted by $d_{x_i,Z}$、Z_i,the eigenvalues is described as r, the number of parameters is denoted by k.

After the above operation finally get a key cluster and a number of dispersed clusters, the key cluster is regarded as the core point, calculate the distance between each point and the core point, and then determine whether there is any abnormality in the current sports data; the greater the distance with the core point, the greater the chance of verifying that this type of sports data is sports injury data; the smaller the distance with the core point, the greater the chance of verifying that this type of sports data is sports injury data. The smaller the distance from the core, the smaller the chance of validating this type of sports data as sports injury data. Thus, the sports data can be classified into the sports injury data set Z_1 and non-sports injury datasets Z_2.

Key feature mining of sports injury data based on principal component analysis

Even if the motion data is grouped by clustering, each group may still contain a large number of feature variables. There may be redundant or highly correlated variables in these features, which not only increases the complexity of data analysis, but also may affect the accuracy of key feature mining. As the number of features increases, there will also be a so-called "dimension disaster", that is, in high-dimensional space, the distribution characteristics of data may become complex and difficult to deal with. Principal Component Analysis (PCA) is an effective dimensionality reduction technology, which can remove redundancy and noise in data while retaining the main information of data [13]. Through PCA processing, the original high-dimensional data can be projected into the low dimensional space to form several main components (principal components), which contain most of the information of the original data and are independent of each other. This can not only reduce the dimension of the data, but also remove the redundancy between features, and then extract the key features in the data.

Therefore, PCA is applied to perform the dimensionality reduction on sports injury data set Z_1 to enhance the independence between the features of sports injury data [14].

Assuming that Z_1 contains n sample of sports injury data, in order to eliminate the effect of magnitude and analysis results, standardize the data in Z_1.

$$X_{norm} = \frac{X - \mu}{\sigma} \quad (6)$$

Among them, X is the data matrix of Z_1, μ is the mean vector of each feature in Z_1, σ is a vector of

standard deviations for each feature (operated element by element), X_{norm} is the normalized data matrix.

Calculate the covariance matrix C of standardized data X_{norm}:

$$C = \frac{1}{n-1} X_{norm}^T X_{norm} \quad (7)$$

For the covariance matrix C, perform the eigen-decomposition and get the eigen-values $\lambda_1, \lambda_2, \mathrm{L}\ \lambda_\alpha$ (in descending order) and the corresponding eigenvectors $v_1, v_2, \mathrm{L}\ v_\alpha$.

The first β eigenvectors corresponding to the largest eigenvalues are selected and used as principal components [15], forming the principal component matrix V:

$$V = \left[v_1, v_2, \mathrm{L}\ v_\beta \right] \quad (8)$$

Projecting it onto the principal component space, the projection of the original data onto the principal component space, i.e., the principal component score, is calculated. It can be expressed as follows:

$$A = X_{norm} V \quad (9)$$

Considering that in the application of principal component analysis (PCA), the principal component vectors obtained are not sparse enough and contain many non-zero elements, when the principal component vectors contain a large number of non-zero elements, it means that these principal components are composed of linear combinations of multiple original variables, rather than significant contributions of a few key variables [16-17]. This makes it difficult to interpret the data characteristics represented by each principal component, because each variable has a certain contribution to the principal component, but the degree of contribution may not be high, affecting the reliability of mining results. Sparse Principal Component Analysis (Sparse PCA) introduces sparsity constraints to ensure that each principal component is composed of only a few key variables. In this way, the data features represented by each principal component are clearer and easier to interpret. In sports data analysis, sparse PCA can more effectively identify key features closely related to sports injuries, providing scientific basis for developing effective intervention measures. Therefore, in order to more accurately mine key features in sports injury data, optimization framework and Least Absolute Shrinkage and Selection Operator (LASSO) regression model [18] are added on the basis of principal component analysis algorithm. The objective function formula of LASSO regression is as follows:

$$\min_{\theta,\psi} \left\| X - \theta\psi^T \right\|_s^2 + \gamma \left\| W\theta \right\|_1 \psi^T\psi = I \quad (10)$$

Among them, θ and ψ are respectively the orthogonal matrices of $n \times p$ and $p \times d$, I is the unit matrix, γ is a regularization parameter, $\|\bullet\|_s$ represents the Frobenius norm, $\|\bullet\|_1$ represents L1 norm, W by $p \times p$ the weighting matrix of order and the matrix W is a diagonal array. By introducing the L1 norm regularization term, a portion of the regression coefficients are compressed to zero, thereby achieving variable selection and feature sparsity. Therefore, based on sparse PCA, the alternating minimization method is adopted to iteratively update and solve the objective function of LASSO regression, in order to further improve the interpretability of features.

As a result, the alternating minimization method is used to iteratively update θ and ψ, solve the objective function of LASSO regression. The specific steps are as follows:

Step 1: Select any θ_0 and ψ_0 as the initial value.

Step 2: For the $t(t \geq 1)$ th iterations, updates θ_t and ψ_t, with the expression:

$$\theta_t = \arg\min_{\theta} \left\| X - \theta_{t-1}\psi^T \right\|_s^2 + \gamma \left\| W\theta \right\|_1 \theta^T\theta = I \quad (11)$$

When solving ψ_t, the following is obtained by performing a singular value decomposition of $\theta\psi_t$:

$$\theta\psi_t = \psi' W_q \theta' \quad (12)$$

Among them, ψ' is the left singular vector matrix, W_q is the diagonal matrix contains singular values, θ' is a right singular vector matrix.

Then, you can take the first χ column of ψ' as a new ψ_t. As a result of this, the ψ_t can be expressed as follows:

$$\psi_t = \arg\min_{\psi} \left\| X - \psi\theta_t^T \right\|_s^2 \psi^T\psi = I \quad (13)$$

Step 3: During the iteration process, the convergence is evaluated by checking the error condition, expressed as:

$$\left\| X - \psi_t\theta_t^T \right\|_s^2 < \xi \quad (14)$$

In the formula, ξ is a small positive number indicating the permissible margin of error.

Step 4: When $t = t_{\max}$, stop iterating. Getting the final θ and ψ, of which θ of the column vectors are sparse principal components [19-20], i.e., the key features of the sports injury data.

Through the above process, the key features in the sports injury data can be mined more accurately and provide scientific basis for the development of effective interventions.

3 Experimental analysis

3.1 Experimental setup

In verifying the application effect of the method in the paper, the test data used came from a city track and field team, and the historical sports data of 50 athletes in the dataset were selected as the test data. The amount of athletic data is 2000 items, including 25 male athletic data and 25 female athletic data, and the age of all athletes is between 18 and 25 years old. The types of sports injuries of the athletes are shown in Figure 1.

Athletes' sports injury data are obtained through Yitikang HC-901G sports monitor. The relevant parameters of the motion monitor are shown in Table 3.

During the experiment, the parameters of the proposed algorithm are set as shown in Table 4.

(a) Running knee injury

(b) Muscle strain

(c) Ankle sprain

(d) Elbow sprain

Figure 1: Types of sports injuries

Table 3: Relevant parameters

Parameter	Numerical value
Size	Width 74mm * Height 12mm * Thickness 11.2mm
Weight	25g
Battery capacity	130mAh
Standby time	35 days
Display	OLEO
Temperature and humidity usage	0℃+40℃，2085%RH

Table 4: Parameter settings of the proposed algorithm

Name of the parameter	Parameter values
pain level weights	0.3
weighting of the degree of swelling	0.2
the weighting of the degree of activity limitation	0.25
time-to-injury weights	0.15
treatment time weights	0.1
K value	3
K in LOF algorithm	2
LOF outlier threshold	2.5
The main ingredient	3
sparsity parameter	0.5
the number of iterations	1000

In Table 4, the outlier threshold of the LOF algorithm is set to 2.5, which aims to ensure that the algorithm can accurately identify outliers that significantly deviate from the normal data distribution range, while avoiding misjudging too many normal points as outliers; The sparsity parameter is set to 0.5 to control the size of the neighborhood, ensuring that each point has a sufficient number of neighboring points to accurately reflect its local density while maintaining computational efficiency; In the K-means algorithm, the value of K is set to 3, which is based on a preliminary understanding of the structure of the dataset. By dividing the data into three main clusters, it helps to better understand the intrinsic structure of the data; In the LOF algorithm, the K value is set to 2, which helps to improve the robustness of the algorithm in sparse or noisy datasets.

Based on the above parameters, test the sensitivity of the proposed method and evaluate its performance with the current parameter settings. The result is shown in Figure 2.

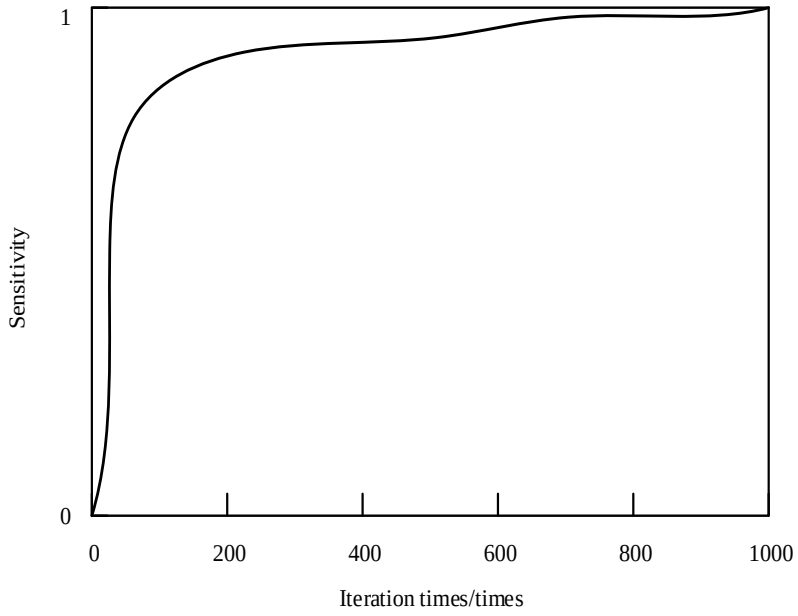
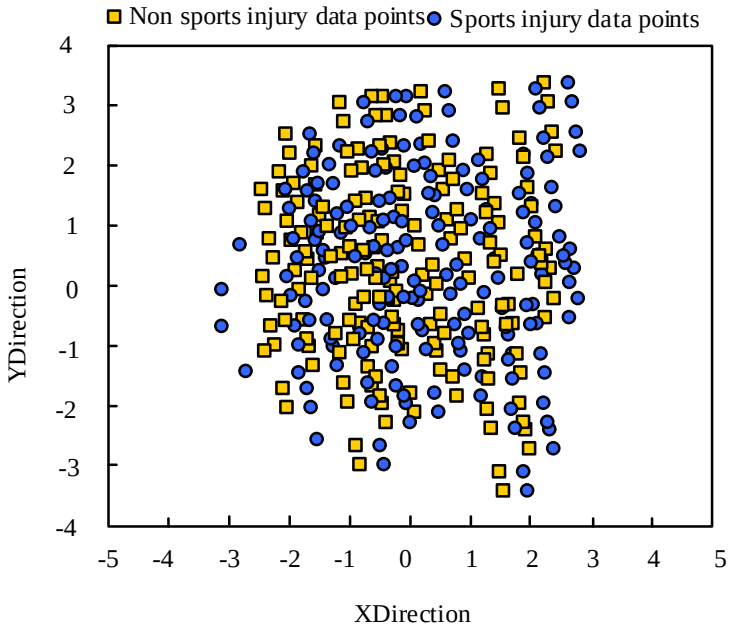


Figure 2: Sensitivity results

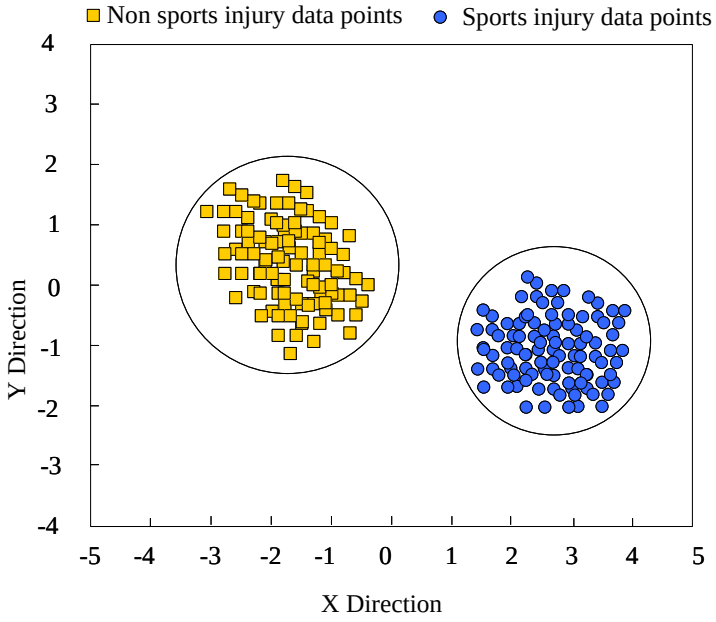
According to the analysis of Figure 2, with the existing parameter settings, the proposed method can achieve a sensitivity of 0.9 within 100 iterations. This indicates that the existing parameter settings can enable the proposed method to accurately identify more sports injury data and have high operational stability.

On this basis, clustering experiments were conducted on the motion data collected from the experiment. By clustering data points with similar features together, a clear cluster is formed. This clustering pattern not only validates the effectiveness of the clustering method proposed in this paper, but also reveals the essential differences between different categories of data. The experimental results are shown in Figure 3. Figure 3 (a) shows the raw data of the operation. By displaying the distribution of raw data points, it is possible to intuitively see the differences in features between non sports injury data and sports injury data, which provides a foundation for subsequent clustering analysis; Figure 3 (b) shows the clustering effect.

As can be seen from Figure 3, this method can accurately capture the essential differences between non sports injury data and sports injury data, and effectively divide non sports injury data and sports injury data. It can be seen from Figure 3 (b) that the non sports injury data are gathered together in a circle to form a relatively concentrated and distinctive area, and the sports injury data are gathered in a circle and distributed in another area with obvious differences. This distribution pattern not only verifies the effectiveness of this method, but also provides strong data support for the subsequent sports injury prevention, early diagnosis and personalized rehabilitation program formulation.



(a) Distribution results of raw motion data



(b) Distribution results of clustered data

Figure 3: Clustering effect test results

3.2 Results and analysis

DBI is an evaluation index of clustering quality, denoted as Γ, this metric evaluates the effect of data clustering by calculating the average distance of data points within each cluster from the center of the cluster as well as the distance between the centers of different clusters, and it is used to measure the compactness and separateness of the clustering results. In the process of feature selection, by comparing the DBI values under different feature combinations, key features that have a significant impact on clustering results can be selected. The formula for Γ is as follows:

$$\Gamma = \frac{1}{k}\sum_{i=1}^{k}\max_{j\neq i}\left(\frac{\overline{X}_i + \overline{X}_j}{\left\|c_i - c_j\right\|^2}\right) \tag{15}$$

Among them, i、j denote the sum of the squares of the intraclass distances of any two classes; c_i、c_j

denote the first clustering center. Take the value of Γ between 0~1, the smaller the value means the smaller the distance within the class, and the larger the distance between the classes, the better the clustering result; on the contrary, the closer the value is to 1, it means the worse the clustering effect is.

In order to further verify the effectiveness of the method in this paper on motion data clustering, the quality of motion data clustering of the algorithm in this paper is evaluated according to formula (15), and the DBI values of the motion data clustering results of this algorithm after clustering different amounts of motion data are tested, as shown in Table 5.

It can be seen from Table 5 that after clustering the motion data with different amounts of data through the algorithm in this paper, even if the amount of data continues to increase, the DBI value of the damaged data and the non damaged data after clustering does not exceed 0.13, indicating that the similarity of data points within the cluster is high, while the data points between clusters are significantly different, which can reliably complete the clustering of the damaged data and non damaged data in the motion data.

In order to verify the reliability of the method of this paper for the mining results of key features of sports injury data, the average absolute percentage error and the root mean square error were calculated to validate the results of the assignment of sports injury to injury factors in the paper, and the average absolute percentage error is denoted by O, the root mean square error is denoted by Q. By comparing the MAPE values under different feature combinations, key features that have a significant impact on the prediction results can be selected. This helps optimize the model and improve prediction accuracy, and in sports injury data analysis, RMSE can be used to measure the accuracy of different models or algorithms in predicting the risk or degree of sports injuries. A lower RMSE value indicates that the model can better capture the intrinsic patterns of motion injury data. The formula for O and Q are as follows:

Table 5: DBI values

Number of sports injury data/piece	Sports injury data	Non sports injury data
100	0.089	0.055
200	0.072	0.049
300	0.062	0.024
400	0.051	0.056
500	0.098	0.078
600	0.078	0.069
700	0.087	0.012
800	0.104	0.099
900	0.108	0.113
1000	0.118	0.109

Table 6: Error details

Number of sports injury data/piece	Internal to injury factors		External to injury factor error	
	Mean Absolute Percent error	Root mean square error	Mean Absolute Percent error	Root mean square error
100	0.025	0.026	0.023	0.022
200	0.034	0.036	0.039	0.035
300	0.039	0.038	0.041	0.039
400	0.041	0.039	0.046	0.041
500	0.053	0.045	0.058	0.047
600	0.064	0.058	0.066	0.063
700	0.075	0.063	0.072	0.069
800	0.079	0.072	0.076	0.072
900	0.082	0.081	0.083	0.079
1000	0.088	0.087	0.089	0.084

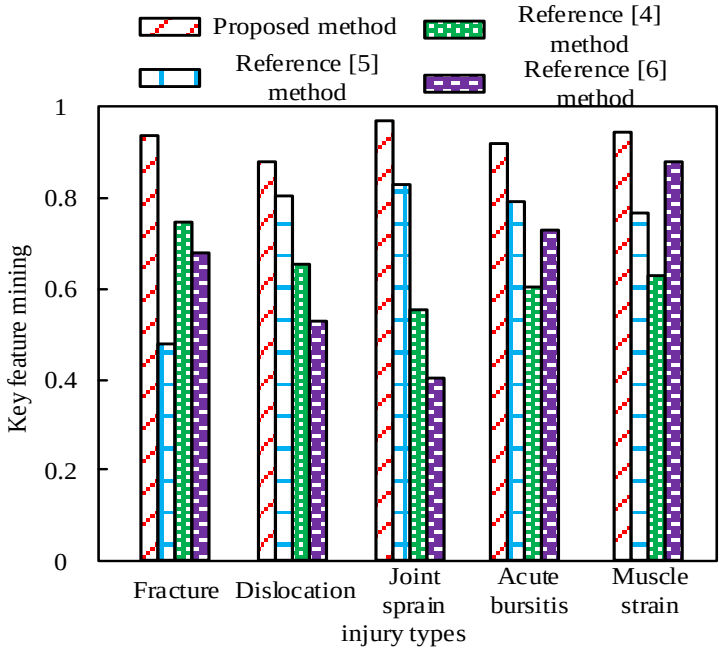
$$O = \frac{\sum_{i=1}^{n} \left| \frac{1}{p_i} \cdot \left(p_i - r_i \right) \right|}{n} \quad (16)$$

$$Q = \sqrt{\frac{\sum_{i=1}^{n} \left(p_i - r_i \right)^2}{n}} \quad (17)$$

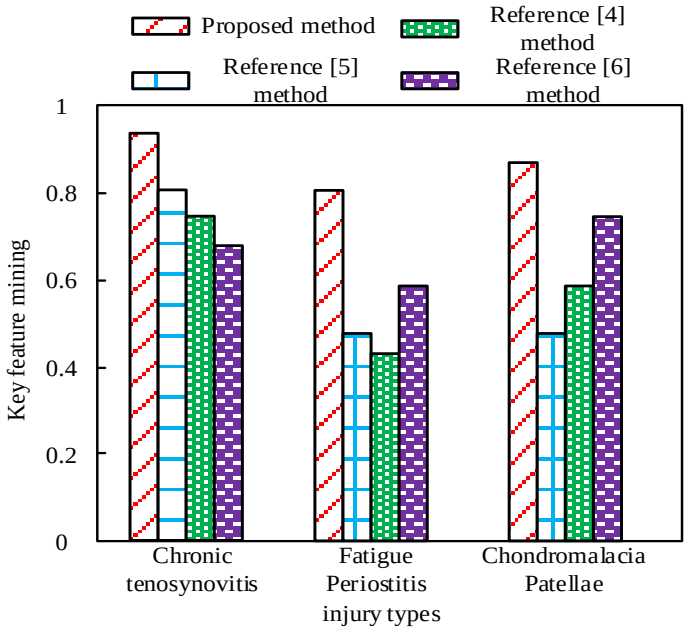
Where, the recognized and actual values of the damage-to-injury factor were denoted by p_i、r_i; total sports injury data is denoted by n; According to Eqs. (16) and (17), the characterization errors of the internal to injury factor and external to injury factor of the sports injury data are calculated respectively, and the smaller the calculation results are, the closer the recognition value is to the actual value, which proves that the method of this paper is able to effectively realize the recognition of the to injury factor of the sports injury. Table 6 shows the details of the error of the recognized to injury factor of the sports injury recorded in this experiment.

As can be seen from Table 6, the error value becomes larger with the increase of the number of sports injury data, but the growth rate is kept at a very low level. Even when the number of sports injury data surged to 1000, the recognition errors of both internal and external injury factors did not exceed 0.1, among which, the highest mean absolute percentage error of internal injury factor was 0.088, and the highest root mean square error was 0.087; the highest mean absolute percentage error of external injury factor was 0.089, and the highest root mean square error was 0.084. The small error values fully proved that the method of this paper is very good at recognizing sports injuries.

According to the urgency of sports injuries, they were categorized into acute and chronic injuries, including fractures, dislocations, joint sprains, acute bursitis, muscle strains, etc., and chronic injuries, including chronic tenosynovitis, fatigue periostitis, chondromalacia patella, etc. The proposed method was chosen as a comparative method. In order to verify the effectiveness of the proposed method more comprehensively, the method of literature [4], the method of literature [5] and the method of literature [6] were chosen as the comparison methods. The key features mining effect of different methods for different sports injuries is tested. The obtained results are shown in Figure 4.



(a) Acute injury



(b) Chronic injury

Figure 4: Mining effect of key features

From the analysis in Figure 4, it can be seen that the proposed method achieves a key feature mining effect of 0.8 or above for both acute and chronic injuries, which is relatively high. Among them, for acute injuries such as fractures, dislocations, and joint sprains, their key characteristics may be more prominent and obvious, as such injuries are usually accompanied by severe pain, swelling, and functional impairment. These features are easily identified and extracted during the data mining process, thereby improving mining accuracy. For chronic injuries such as chronic tenosynovitis, fatigue periostitis, and patellar chondromalacia, the extraction of key features may be more complex and difficult due to their longer course, relatively subtle symptoms, and tendency to recur. However, by using the reweighted sparse principal component analysis method, these features can be accurately captured, providing important references for subsequent injury prevention, diagnosis, and treatment.

4 Discussion

Compared with previous works such as literature [4], literature [5], and literature [6], the method proposed in this paper shows higher performance in mining key features of acute and chronic injuries. This advantage is mainly due to the reweighted sparse principal component analysis (PCA) method used in this paper. This method can accurately identify the most critical features for distinguishing acute and chronic injuries, providing important reference for subsequent injury prevention, diagnosis, and treatment.

Specifically, previous work may have mainly relied on traditional data mining techniques such as support vector machines (SVM), decision trees, or neural networks. Although these methods can to some extent uncover features related to injuries, they often lack precision and sensitivity for specific types of injuries, such as acute and chronic injuries. In contrast, the reweighted sparse PCA method proposed in this paper can better capture key information in the data by introducing sparsity and reweighting strategies, thereby improving the accuracy and efficiency of key feature mining.

In addition, this paper also verified the reliability of the proposed method in identifying sports injury factors by calculating indicators such as Mean Absolute Percentage Error (MAPE) and Root Mean Square Error (RMSE). The results show that even when the amount of sports injury data surges to 1000, the identification error of internal and external injury factors does not exceed 0.1, which fully proves the effectiveness of our method.

Another significant advantage of this paper compared to previous work is its comprehensiveness and systematicity. This paper not only focuses on the mining of key features, but also comprehensively evaluates the clustering results through evaluation indicators such as DBI index. This comprehensive and systematic research method enables this paper to gain a deeper understanding of the inherent patterns and characteristics of sports injury data, thereby providing more comprehensive guidance for the prevention and treatment of sports injuries.

5 Conclusion

In this paper, we propose to improve the key feature mining algorithm of sports injury data to accurately mine the key features in the sports injury data to improve the accuracy of sports injury prediction and the level of personalization of rehabilitation strategies. The specific advantages are as follows:

(1) By improving the feature mining research, the key features that are highly related to sports injuries can be extracted from the raw data more accurately. These features not only cover the basic information such as exercise intensity, frequency and mode, but also may include the athlete's physiological indicators, psychological state and other deep-level information, so as to construct a more comprehensive and accurate prediction model.

(2) PCA data dimension reduction technology can significantly reduce the dimension of data while retaining key information. This not only reduces the amount of calculation, but also improves the utilization efficiency of sports injury data, making it possible to process and analyze large-scale sports injury data.

Experiment proves that, the method of this paper can accurately classify non-sports injury data and sports injury data, and effectively mine the key features of sports injury data, which will play a more important role in the field of sports injury prevention, diagnosis and rehabilitation.

Although PCA data dimensionality reduction technology performs well in reducing data dimensions, in practical applications, as the amount of data and feature dimensions increase, the running time of the algorithm may be significantly extended. At the same time, improved feature mining algorithms may also involve more complex calculations when extracting deep level features, further increasing the computational burden. In addition, when dealing with large-scale datasets, the scalability of the algorithm becomes a key consideration factor. Although the method proposed in this paper is theoretically applicable to large-scale data, it may encounter problems such as memory limitations and tight computing resources in practical operations, which can affect the performance and scalability of the algorithm.

Future work will focus on optimizing algorithm performance and enhancing scalability to address the challenges posed by large-scale and high-dimensional data. We will explore more efficient data dimensionality reduction techniques and improve the computational process of feature mining algorithms, while utilizing parallel computing, distributed computing, and cloud computing technologies to reduce runtime and overcome resource limitations. In addition, the plan is to apply the algorithm to additional data types, such as biomechanical data, to comprehensively understand the health status of athletes. At the same time, alternative mining algorithms such as deep learning and machine learning ensemble methods will also be studied to enrich the feature selection process and improve prediction accuracy. We hope to broaden the application scope of the algorithm and improve its practicality and influence in the fields of sports injury prevention, diagnosis, and rehabilitation.

References

[1] A. Ruffault, M. Sorg, S. Martin, C. Hanon, L. Jacquet, E. Verhagen, and P. Edouard, “Determinants of the adoption of injury risk reduction programmes in athletics (track and field): an online survey of 7715 french athletes,” British Journal of Sports Medicine, vol. 56, no. 9, pp. 499-505, 2021. https://doi.org/10.1136/bjsports-2021-104593

[2] G. S. Bullock, J. Mylott, T. Hughes, K. F. Nicholson, R. D. Riley, and G. S. Collins, “Just how confident can we Be in predicting sports injuries? A systematic review of the methodological conduct and performance of existing musculoskeletal injury prediction models in sport,” Sports Medicine, vol. 52, no. 10, pp. 2469-2482, 2022. https://doi.org/10.1007/s40279-022-01698-9

[3] V. Sarlis, V. Chatziilias, C. Tjortjis, and D. Mandalidis, “A data science approach analysing the impact of injuries on Basketball Player and Team Performance,” Information Systems, vol. 99, pp.101750.1-101750.16, 2021. https://doi.org/10.1016/j.is.2021.101750

[4] L. Abualigah, and A. J. Dulaimi, “A novel feature selection method for data mining tasks using hybrid Sine Cosine Algorithm and Genetic Algorithm,” Cluster Computing, vol. 24, no. 3, pp. 2161-2176, 2021. https://doi.org/10.1007/s10586-021-03254-y

[5] K. Kalaivani, M. Priya, and P. Deepan, “Heart disease prediction system based on multiple feature selection algorithm with ensemble classifier,” ECS Transactions, vol. 107, no. 1, pp. 8049-8059, 2022. https://doi.org/10.1149/10701.8049ecst

[6] N. Shehab, M. Badawy, and H. A. Ali, “Toward feature selection in big data preprocessing based on hybrid cloud-based model,” The Journal of Supercomputing, vol. 78, no. 3, pp. 3226-3265, 2021. https://doi.org/10.1007/s11227-021-03970-7

[7] G. X. Tan, T. N. Tang, T. Yi, and H. F. Chen, “Rock climbing keypoint detection algorithm based on improved hourglass,” Modern Electronics Technique, vol. 47, no. 17, pp. 117-122, 2024. https://doi:10.16652/j.issn.1004-373x.2024.17.019.

[8] Y. S. Chen, and X. Z. Zhou, “Simulation of large data set local anomaly mining based on self encoder,” Computer Simulation, vol. 40, no. 6, pp. 495-498+508, 2023. https://doi.org/10.3969/j.issn.1006-9348.2023.06.091

[9] N. Shehab, M. Badawy, and H. A. Ali, "Toward feature selection in big data preprocessing based on hybrid cloud-based model," The Journal of Supercomputing, vol. 78, no. 3, pp. 3226-3265, 2022. https://doi.org/10.1007/s11227-021-03970-7

[10] B. Mirzaei, B. Nikpour, and H. Nezamabadi-Pour, "CDBH: A clustering and density-based hybrid approach for imbalanced data classification," Expert Systems with Applications, vol. 164, pp. 114035.1-114035.15, 2021. https://doi.org/10.1016/j.eswa.2020.114035

[11] D. Krleza, B. Vrdoljak, and M. Brcic, "Statistical hierarchical clustering algorithm for outlier detection in evolving data streams," Machine Learning, vol. 110, no. 1, pp. 139-184, 2021. https://doi.org/10.1007/s10994-020-05905-4

[12] R. D. Vaghela, and S. S. Iyer, "A comparative analysis of clustering algorithm," ECS Transactions, vol. 107, no. 1, pp. 2435-2443, 2022. https://doi.org/10.1149/10701.2435ecst

[13] E. Kepes, J. Vrabel, P. Porizka, and J. Kaiser, "Addressing the sparsity of laser-induced breakdown spectroscopy data with randomized sparse principal component analysis," Journal of Analytical Atomic Spectrometry, vol. 36, no. 7, pp. 1410-1421, 2021. https://doi.org/10.1039/d1ja00067e

[14] E. O. Omuya, G. O. Okeyo, and M. W. Kimwele, "Feature selection for classification using principal component analysis and information gain," Expert Systems with Applications, vol. 174, pp. 114765.1-114765.12, 2021. https://doi.org/10.1016/j.eswa.2021.114765

[15] E. Tsalera, A. Papadakis, and M. Samarakou, "Novel principal component analysis-based feature selection mechanism for classroom sound classification," Computational Intelligence, vol. 37, no. 4, pp. 1827-1843, 2021. https://doi.org/10.1111/coin.12468

[16] S. S. Dey, M. Molinaro, and G. Wang, "Solving sparse principal component analysis with global support," Mathematical Programming, vol. 199, no. 1-2, pp. 421-459, 2023. https://doi.org/10.1007/s10107-022-01857-w

[17] J. Kim, M. Tawarmalani, and J. P. P. Richard, "Convexification of permutation-invariant sets and an application to sparse principal component analysis," Mathematics of Operations Research, vol. 47, no. 4, pp. 2547-2584, 2022. https://doi.org/10.1287/moor.2021.1219

[18] Y. Jiang, S. P. Wu, K. Hu, and L. B. Long, "Imbalanced data classification method based on Lasso and constructive covering algorithm," Journal of Computer Applications, vol. 43, no. 4, pp. 1086-1093, 2023. https://doi.org/10.11772/j.issn.1001-9081.2022040490

[19] A. D. McRae, J. Romberg, and M. A. Davenport, "Optimal convex lifted sparse phase retrieval and PCA with an atomic matrix norm regularizer," IEEE Transactions on Information Theory, vol. 69, no. 3, pp. 1866-1882, 2023. https://doi.org/10.1109/TIT.2022.3228508

[20] R. Kawasumi, and K. Takeda, "Automatic hyperparameter tuning in sparse matrix factorization," Neural Computation, vol. 35, no. 6, pp. 1086-1099, 2023. https://doi.org/10.1162/neco_a_01581