Feature selection based on improved principal component analysis

Zhangyu Li
School of Economics and Management, Xiamen University of Technology, Xiamen, China
lzy1173745655@outlook.com

Yihui Qiu
School of Economics and Management, Xiamen University of Technology, Xiamen
China,qiuyihui@xmut.edu.cn

ABSTRACT

Abstract**:** The filtered feature selection method has low computational complexity and less time, and is widely used in feature selection, but the filtered method only considers the importance of features for label classification and ignores the correlation between features. For this reason, a feature selection method with improved principal component analysis is proposed. The main idea of the method is that on the basis of principal components, the loadings of each indicator on different principal components and their variance contribution ratios with that principal component are considered. A number of indicators with the largest cumulative contribution rates were selected, so that the final extracted indicators retained more information. Subsequently, comparative experiments are conducted using the UCI dataset, and the results show that the approach proposed in this paper has some superiority over other methods. Finally, the features of China's green innovation efficiency are selected using the approach proposed in this paper to demonstrate the feasibility of the method.

CCS CONCEPTS

• **Computing methodologies** → Machine learning; Machine learning algorithms; Feature selection.

KEYWORDS

PCA, Feature selection, contribution rate

ACM Reference Format:
Zhangyu Li and Yihui Qiu. 2023. Feature selection based on improved principal component analysis. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023), March 17–19, 2023, Shanghai, China.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3590003.3590036

1 INTRODUCTION

Feature selection is an important pre-processing step in machine learning. Its purpose is to select some features from the original dataset while retaining as much of the underlying information as possible, so as to form the desired dataset. In this way, it reduces the dimensionality of the dataset, preventing the curse of dimensionality, and reducing training time while improving algorithm efficiency [1]. Currently, there are two main categories of feature selection methods. From the selection perspective, feature selection methods can be divided into filter, wrapper, and embedded. From the transformation perspective [2], the main methods are principal component analysis, independent component analysis, and popular learning algorithms, etc. These methods all aim to reduce high-dimensional data to a lower dimensional space, but they cannot extract the original and optimal subset of features. Kale [3] introduced a PCA-based optimal feature subset selection, which is capable of handling weighted classification problems. Chi G T [4] proposed a feature selection method based on principal component analysis. By retaining the indicators with large absolute factor loadings under the same criterion, the significance of the indicators to the evaluation results is ensured. Then, by eliminating one of the two highly correlated indicators, it ensures that only a small amount of information is needed to obtain an indicator that accounts for 80% of the variance of the original indicator. However, selecting indicators pairwise through correlation has subjectivity, which is not suitable for reducing or identifying important indicators [5]. Given the above situation, this paper proposes an improved method of principal component analysis for extracting the selection of evaluation indicators, using cumulative contribution rate to identify the most important variables, thus avoiding subjectivity.

CACML 2023, March 17–19, 2023, Shanghai, China

ACM ISBN 978-1-4503-9944-9/23/03…$15.00
https://doi.org/10.1145/3590003.3590036

2 FEATURE SELECTION METHOD BASED ON PRINCIPAL COMPONENT ANALYSIS

PCA is an effective method in statistical data analysis, which is mainly used for dimension reduction [6]. Its main principle is to use the transformation of the feature space of the dataset to reduce the dimensionality of the dataset that has a high dimension and is correlated. After dimension reduction by using PCA, the original dataset will be transformed into a dataset consisting of several principal components, which do not have correlation. However, after dimension reduction by using PCA, it will become a new feature to be used with the original feature. This paper proposes a feature selection method based on PCA. The pseudo code of the algorithm is shown in Table 1.

Suppose there are m input indicators, denoted as $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_{\boldsymbol{m}})^T$; m principal components are denoted as $\boldsymbol{F} = (\boldsymbol{F}_1, \boldsymbol{F}_2, \cdots, \boldsymbol{F}_{\boldsymbol{m}})$, and $(\boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_{\boldsymbol{n}})^T$ is the relative contribution value of m input indicators, namely $\mathrm{X} = (\mathrm{X}_1, \mathrm{X}_2, \cdots, \mathrm{X}_{\mathrm{m}})^{\mathrm{T}}$. The specific steps are as follows:

Step 1: Calculate the sample correlation matrix corresponding to the standardized variable $\boldsymbol{X}^* = (\boldsymbol{X}_1^*, \boldsymbol{X}_2^*, \cdots, \boldsymbol{X}_{\boldsymbol{m}}^*)^T$.

Step 2: Obtain the eigenvalues $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \cdots, \boldsymbol{\lambda}_{\boldsymbol{m}} (\boldsymbol{\lambda}_1 \geqslant \boldsymbol{\lambda}_2 \geqslant \cdots \geqslant \boldsymbol{\lambda}_{\boldsymbol{m}} \geqslant 0)$ and unit eigenvectors

Table 1: Pseudo Code for Feature Selection Based on Principal Component Analysis

Input: sample set X=(X_1,X_2,$\cdots$,X_m)ˆT
Procedure:
1, Normalize the samples to obtain the correlation matrix Xˆ*
2, Calculate the eigenvalues λ, eigenvectors l, and variance contribution rate $\sqrt{\omega}$
3, Calculate the cumulative contribution rate a of each indicator
4, Take the indicator corresponding to the top 80% contribution rate
Output: the set Xˆ’

Table 2: Datasets Information

Datasets	Total Number of Samples	Sample Categories	Number of Features
Iris	150	3	4
Heart-h	294	13	5
Blood	748	2	4
Segment	5456	4	2
Dermatology	385	34	6

$l_1, l_2, \cdots, l_m$ from the sample correlation matrix, and calculate the contribution rate $\boldsymbol{\omega_i} = \boldsymbol{\lambda_i^*}(\sum_{j=1}^{\boldsymbol{m}} \boldsymbol{\lambda_j})^{-1}(\boldsymbol{i} = 1, \cdots, \boldsymbol{m})$.

Step 3: In order to improve the influence of each index, the square root operation with calculated contribution rate of variance was executed. Then, the weight of the feature vector (indexes on each principal component) is multiplied by the variance contribution rate corresponding to belonging the principal component, and the results of the same index are added, i.e., the corresponding contribution value of each input variable can be calculated by equation $(\boldsymbol{a_1}, \boldsymbol{a_2}, \cdots, \boldsymbol{a_m})^T = |\boldsymbol{L^*}(\sqrt{\omega}_1, \sqrt{\omega}_2, \cdots, \sqrt{\omega}_{\boldsymbol{m}})^T|$, where the absolute value symbol represents taking the absolute value of the m elements in the column vector.

Step 4: Sort the elements in vector $(a_1, a_2, \cdots, a_n)$ in descending order to obtain vector$(\mathbf{a_1^*}, \mathbf{a_2^*}, \cdots, \mathbf{a_n^*})^{\mathbf{T}}(\mathbf{a_1^*} \geqslant \mathbf{a_2^*} \geqslant \ldots \geqslant \mathbf{a_m^*} \geqslant 0)$

Step 5: According to the cumulative contribution rate criterion, select the top 80% of the k variables in terms of cumulative contribution rate to complete the feature selection.

3 COMPARISON EXPERIMENT

In order to verify the superiority and inferiority of the proposed feature selection method, this paper took the methods proposed by Singh [7] and Lim [8]as reference, and compared the proposed method with the Correlation-based Feature Selection (CFS), the Fisher Score, the Chi-square feature selection, the Relief, the minimum redundancy maximum relevance (mRMR) feature selection method and the Alpha-investing online stream feature selection algorithm. A series of performance metrics were used to evaluate feature selection, including the classification accuracy, the Normalized Mutual Information (NMI), the Jaccard coefficient, and F-value. The experiment used two classifiers, namely Support Vector Machine Classifier and Decision Tree Classifier, and In reference to Sahebi [9], five UCI data sets were adopted. The specific information of the datasets is shown in Table 1.

The paper measured the above indicators of the model through ten-fold cross validation. The principle of this method is: all training samples are divided into ten parts as randomly as possible. Nine parts are selected as the training set and the remaining one as the test set for ten times. Then the average of these ten tests is taken as the result of the experiment. In order to prove the practicability of the method proposed in this paper, 10-fold cross-validation was carried out for 30 times according to the central limit theorem to eliminate the effect of randomness on the results. Therefore, the results listed in this paper are the average after 30 times of 10-fold cross-validation.

4 RESULTS AND DISCUSSION OF THE EXPERIMENT

In Table 3, Table4, Table 5, Table 6, it shows the results of the evaluation metrics for the six algorithms on five datasets under two classifiers, namely SVM and decision tree. The bold text in the table represents the best results achieved by the proposed method compared to other methods. As shown inTable 3, Table4, Table 5, Table 6 , the method proposed in this paper achieves better accuracy on five data sets. The accuracy of the proposed method is superior to other methods in the decision tree (DT) classifier, and it also performs well in SVM classifier. In comparison the value of F, the method proposed in this paper is better than other methods: there are four data sets and two data sets for DT classifier and SVM classifier to obtain the optimality respectively. And it also achieves better results on other evaluation metrics. Although the performance on the SVM classifier is a bit worse than that of the decision tree classifier, it still has some advantages compared with other algorithms.

5 APPLICATION TO CHINA GREEN INNOVATION EFFICIENCY

As today’s world evolves, green innovation is becoming an increasingly important topic. In order to better improve green innovation efficiency, it is necessary to determine which characteristics have more significant effects on green innovation efficiency in China.

Table 3: Accuracy comparison with other algorithms in different classifiers.

				Accuracy			
DT	This article's	CFS	fisher	alpha-investing	chi square	MRMR	reliefF
iris	**0.95**	0.9292	0.9292	0.9283	0.9383	0.8917	0.933
segment	**0.9654**	0.9364	0.8889	0.9642	0.964	0.9581	0.9562
heart-h	**0.642**	0.5569	0.5882	0.5537	0.635	0.5402	0.5568
dermatology	**0.959**	0.9505	0.9285	0.6742	0.8531	0.9382	0.9004
blood data	**0.7609**	0.7136	0.7274	0.6836	0.753	0.7442	0.7071
SVM	This article's	CFS	fisher	alpha-investing	chi square	MRMR	reliefF
iris	**0.955**	0.95	0.95	0.9523	0.95	0.9417	0.9417
segment	0.9242	0.945	0.7602	0.9118	0.9264	0.9264	0.822
heart-h	0.6594	0.6161	0.6558	0.6936	0.6344	0.6637	0.6344
dermatology	0.9216	0.976	0.93172	0.6828	0.8533	0.9624	0.9523
blood data	**0.7742**	0.7658	0.7525	0.7709	0.7575	0.7575	0.7575

Table 4: Comparison of NMI with other algorithms in different classifiers.

				NMI			
DT	This article's	CFS	fisher	alpha-investing	chi square	MRMR	reliefF
iris	**0.8939**	0.8483	0.8546	0.8615	0.8835	0.8079	0.8631
segment	**0.9346**	0.6707	0.8323	0.933	0.9304	0.9255	0.9201
heart-h	0.411	0.424	0.3743	0.4147	0.4579	0.2799	0.3159
dermatology	**0.9488**	0.9416	0.9139	0.7564	0.8917	0.9224	0.8616
blood data	0.0321	0.0384	0.0277	0.0271	0.0699	0.0098	0.0307
SVM	This article's	CFS	fisher	alpha-investing	chi square	MRMR	reliefF
iris	**0.9106**	0.8919	0.8914	0.9077	0.9047	0.886	0.8853
segment	0.8827	0.708	0.7419	0.86	0.8865	0.8896	0.818
heart-h	0.1929	0.4182	0.3822	0.43066	0.3533	0	0
dermatology	0.9034	0.9647	0.9081	0.7546	0.9237	0.9524	0.9323
blood data	0.013	0	0	0	0.0239	0	0.0015

Table 5: Comparison of Jaccard coefficients with other algorithms in different classifiers.

				Jaccard coefficients			
DT	This article's	CFS	fisher	alpha-investing	chi square	MRMR	reliefF
iris	**0.9099**	0.8758	0.8748	0.874	0.8923	0.8167	0.8785
segment	**0.9333**	0.8796	0.801	0.9313	0.9299	0.9201	0.9171
heart-h	**0.4767**	0.3936	0.4244	0.3879	0.4685	0.3723	0.3899
dermatology	**0.9329**	0.9259	0.8725	0.5092	0.7417	0.8886	0.8192
blood data	**0.6144**	0.5594	0.572	0.5217	0.6063	0.5928	0.548
SVM	This article's	CFS	fisher	alpha-investing	chi square	MRMR	reliefF
iris	**0.9277**	0.9099	0.9077	0.9231	0.9138	0.8967	0.8967
segment	0.8596	0.8977	0.6138	0.8383	0.864	0.8633	0.6979
heart-h	0.4922	0.4478	0.4931	0.5333	0.466	0.4969	0.4649
dermatology	0.8583	0.9541	0.8762	0.5186	0.7448	0.9294	0.9106
blood data	0.6057	0.6206	0.6033	0.6273	0.6098	0.6098	0.6098

Table 6: Comparison of F-values with other algorithms in different classifiers.

			F-values				
DT	This article's	CFS	fisher	alpha-investing	chi square	MRMR	reliefF
iris	**0.9491**	0.9297	0.9312	0.9275	0.9382	0.8889	0.9306
segment	**0.9651**	0.9353	0.8887	0.9642	0.9637	0.9568	0.9565
heart-h	**0.6119**	0.5584	0.5777	0.555	0.592	0.4981	0.5293
dermatology	**0.9585**	0.9593	0.9274	0.5736	0.8222	0.9355	0.8961
blood data	0.6802	0.7074	0.6936	0.6716	0.7272	0.645	0.6972
SVM	This article's	CFS	fisher	alpha-investing	chi square	MRMR	reliefF
iris	**0.9588**	0.9496	0.949	0.9576	0.9491	0.9399	0.9402
segment	0.9227	0.9441	0.7469	0.9097	0.9258	0.9257	0.778
heart-h	0.5369	0.5609	0.5867	0.629	0.565	0.5297	0.4927
dermatology	0.9208	0.9754	0.9328	0.58	0.8074	0.9602	0.9518
blood data	**0.6675**	0.6643	0.6463	0.6712	0.6589	0.6531	0.6545

Table 7: Green Innovation Efficiency Evaluation Index System.

innovation input	Full-time equivalent of R&D staff($\boldsymbol{X}_1$)
	Internal expenditure on R&D expenses(X_2)
	Environmental pollution treatment investment amount(X_3)
	Energy saving and environmental protection expenditure(X_4)
	New product development expenses(X_5)
	Technology introduction and renovation expenses(X_6)
	Total energy consumption(X_7)
expected output	Number of patents granted(Y_1)
	Technology Market Turnover(Y_2)
	New product sales revenue(Y_3)
	Industrial value added(Y_4)

Therefore, it is necessary to make a scientific selection of indicators of green innovation efficiency. Through the typical literature, 14 indicators were selected, which contain 7 input indicators, output indicators, as shown in the Table 7. However, an excessive number of indicators will result in a lack of discriminatory efficiency in the final result. It is necessary to make a selection of indicators for subsequent work.

The empirical sample is 30 Chinese provincial-level cities excluding Hong Kong, Macao, Taiwan and Tibet, for a total of ten years from 2011 to 2020.The above input variables as well as expected output variables were screened separately using the method proposed in this paper, and the cumulative contribution rate of the five variablesX_1, X_5, X_2, X_6, X_4reached 86.8%, which can be used to replace the original seven variables; the cumulative contribution rate of the variables Y_4, Y_3, Y_1 has reached 99% of the original four expected output indicators. The results show that the full-time equivalent of R&D personnel and new product development expenses have a greater impact on the efficiency of green innovation, which is the same as the previous studies by scholars [10, 11]. The experiment shows the practicality of the method proposed in this paper.

6 CONCLUSIONS

In this paper, we propose a PCA-based feature selection method, which first uses the loadings of each indicator on different principal components and their variance contributions with that principal component to calculate a number of indicators with the largest cumulative contributions for feature selection. Then, based on the data in the UCI dataset, the proposed method is compared with other feature selection methods. Finally, the method proposed in this paper is used to select the characteristics of China's green innovation efficiency, which verified the practicability of the method proposed in this paper. The experiments show that the method performs better in the public dataset, and it may be possible to try to use other PCA (For example kernel PCA, probabilistic PCA) for subsequent studies. Finally, the features of green innovation efficiency in China were selected using the method proposed in this paper, and the feasibility of the method was verified.

REFERENCES

[1] Zhou H, Zhang Y, Zhang Y and Liu H. Feature selection based on conditional mutual information: minimum conditional relevance and minimum conditional redundancy. *Applied Intelligence*, 2019, 49(3) : 883-896. https://doi.org/10.1007/s10489-018-1305-0.

[2] Ye X L, Lan J L and Guo T. (2014) Network traffic feature selection algorithm based on PCA and taboo search. Computer

Science,41(01):187-191. https://kns.cnki.net/kcms2/article/abstract?v$=$3uoqIhG8C44YLTlOAiTRKgchrJ08w1e7M8Tu7YZds89NyEjIjuMbEOVYc9b1HujUP_e_JuVGMXjzV5RWNdGXPIQoT5NIm_MK&uniplatform$=$NZKPT.

[3] Kale A P and Sonavane S. (2018) PF-FELM: A robust PCA feature selection for fuzzy extreme learning machine. IEEE Journal of Selected Topics in Signal Processing, 12(6): 1303-1312.https://doi.org/10.1109/JSTSP.2018.2873988.

[4] Chi G T and Zhao Z C. (2018) The construction of an evaluation index system of science and technology innovation with enterprises as the main body. Scientific Research Management,39(S1):1-10.http://www.cqvip.com/qk/95604x/2018s1/7589717650484956834948484849.html.

[5] Dariush K, Wade D C and Joe Z. (2019) Number of performance measures versus number of decision making units in DEA. Annals of Operations Research. 303(1-2):529-562. https://doi.org/10.1007/s10479-019-03411-y.

[6] Li M, Wang H, Yang L, Liang, Y, Shang Z and Wan, H. (2020). Fast hybrid dimensionality reduction method for classification based on feature selection and grouped feature extraction. Expert Systems with Applications, 150, 113277. https://doi.org/10.1016/j.eswa.2020.113277.

[7] Singh N and Singh P. (2021) A hybrid ensemble-filter wrapper feature selection approach for medical data classification. Chemometrics and Intelligent Laboratory Systems, 217: 104396. https://doi.org/10.1016/j.chemolab.2021.104396.

[8] Lim H and Kim D W. (2021) Pairwise dependence-based unsupervised feature selection. Pattern Recognition, 111: 107663. https://doi.org/10.1016/j.patcog.2020.107663.

[9] Sahebi G, Movahedi P, Ebrahimi M, *et al.* GeFeS: A generalized wrapper feature selection approach for optimizing classification performance[J]. Computers in biology and medicine, 2020, 125: 103974. https://doi.org/10.1016/j.compbiomed.2020.103974

[10] Li Y, Huang N and Zhao Y. (2022) The Impact of Green Innovation on Enterprise Green Economic Efficiency. International Journal of Environmental Research and Public Health,19(24): 16464. https://doi.org/10.3390/ijerph192416464.

[11] Chen X, Liu X, Gong Z and Xie J. (2021) Three-stage super-efficiency DEA models based on the cooperative game and its application on the R&D green innovation of the Chinese high-tech industry. Computers & Industrial Engineering, 156: 107234. https://doi.org/10.1016/j.cie.2021.107234.