Performance Analysis of Data Mining Algorithms Based on PCA

Ruifeng Bai
Business College
Shandong University, Weihai
Weihai, 264209, China
e-mail: brf@sdu.edu.cn

Lin Yang
Scholl of Mechanical, Electrical & Information Engineering
Shandong University, Weihai
Weihai, 264209, China
e-mail: yl878604@163.com

Jie Wang
Scholl of Mechanical, Electrical & Information Engineering
Shandong University, Weihai
Weihai, 264209, China
e-mail: wangjie1990923@163.com

Jingchang Pan
Scholl of Mechanical, Electrical & Information Engineering
Shandong University, Weihai
Weihai, 264209, China
e-mail: pjc@sdu.edu.cn

Abstract—Data mining algorithms behave differently under different application context. It is an important topic to find out the characteristics of the relevant algorithms. This paper studied PCA based dimension reduction and the functional performance of data mining algorithms (ANN, Bayes, KNN, K-means) under different dimension reduction rates in finding Cataclysmic Variable Stars(CVs) in a hybrid celestial spectra dataset. The dataset was selected from SDSS(Sloan Digital Sky Survey), 1417 spectra altogether. In the dataset, there are 15 CVs, along with other type of celestial bodies. ANN, Bayes, KNN and K-means were chosen to test their performances in finding CVs and time cost under different PCA dimensions. The classification accuracy and time cost were analyzed of the four mentioned algorithms in detail under different PCA dimensions. A series of experiments were done to carry out our research. Through this study, we can understand the inherent characteristics of the four algorithms and make better choices in future data mining applications.

Keywords- PCA; Classification; Clustering; Spectrum; Cataclysmic Variable Star

I. INTRODUCTION

Celestial spectrum contains abundant physical and chemical information about the celestial body. This paper studied the spectrum of cataclysmic variable stars [1, 2, 3, 4], using data mining technology to extract the spectral characteristics from the known cataclysmic variable stars(CVs), and selecting CVs candidates, constructed the principal component spectra by the PCA[5]. Using principal component as axis, projecting the sample points directly on the sample principal component axis, and get the sample feature points of two-dimensional and three-dimensional plane, it greatly reduces the dimension of the spectral data. We also used artificial neural network (ANN) [6], K means [7], K nearest neighbor [8] and Bayes [9] in classification and clustering under more PCA dimensions to test the performances of the algorithms. The algorithms were implemented in MATLAB environment. Comprehensive analysis and comparisons were made to show the inherent performances of the above algorithms.

II. DATA PREPARATION

FITS (Flexible Image Transport System) is commonly used data format in astronomical filed. It is designed to exchange data among different platforms.

A FITS structure is composed of the following components:

- Header and Data Unit (HDU).
- 0 or several sequenced extension units (Conforming Extensions).

The FITS head includes the descriptions about the FITS file, including right ascension, declination, observation time, exposure time, etc. One can understand and access the FITS structure and information through the keywords in the FITS head.

The FITS file adopted in this paper is SDSS DR7 version (DR8 version has some different keywords). We used two keywords and their values: SPEC_CLN、SN_G. SPEC_CLN is the rough classification to SDSS spectra, most of them are correct, but some of them are not. So it is used only as a reference, not the unique criteria. SN_G recorded the signal noise ratio (SN) in g band. The information is shown in Table 1.

Experimental data were selected from the SDSS [10] DR7, total 1417 spectra, of which contains 15 CVs spectra. Preprocessing of the spectra includes wavelength selection and normalization. Each spectrum contains a range of wavelength and flux. In this paper, flux of [3800: 9000] wavelength range are extracted in order to ensure the unity of the spectra, where step is calculated according to the information stored in FITS head. Each spectrum was finally sampled to 3522 points (wave length and flux).

TABLE I. SPEC_CLN SEGMENTS IN FITS

Class	Value
0	UNKNOWN
1	STAR
2	GALAXY
3	QSO
4	HIZ_QSO
5	SKY
6	STAR_LATE
7	GAL_EM

Suppose M pieces of spectral data is stored in the matrix Pij, where the size of P is M×3522, each row represents a spectrum. After getting the flux of the spectrum, it is needed to normalize the flux into a unified space in order to remove the effect caused by different flux intensity of different spectra. Normalization equation is as follows:

$$P_{ij} = \frac{P_{ij}}{\sqrt{\sum_{j=1}^{N} P_{ij}^{2}}} \qquad (1)$$

III. PCA REDUCTION

Principal components analysis (PCA) is also known as Karhunen-Loève transform, is an efficient data compression method. Its purpose is to search a low-dimensional orthogonal vectors from the total space of the data to make them the most representative of the features of the source data.

The basic process of dimensionality reduction using PCA is as follows:

1) Select M CVs spectra, denoted as Pi (i =1, 2, …, M), each spectrum is N-dimensional, forming a matrix of [M×N].

2) Input data are normalized so that each attribute falls in the same range, which helps to ensure that attributes with large range do not dominate the attributes with smaller range. The spectral matrix $P_{M\times N}$ is obtained, where M represents the number of spectra, N represents the number of components of each spectrum, each component representing the flux at the corresponding wavelength after the normalization.

3) Construct correlation matrix $C_{ij}=P\times P^{T}$, where P^{T} is the transpose of P, $i \in [1,N]$, $j \in [1,M]$, C_{ij} is a matrix of [N×N].

4) Seek the eigenvalues and eigenvectors of matrix C_{ij}, then diagonalize the matrix C_{ij}, the diagonalization equation is as follows:

$$C = R\Lambda R^{T},\ \Lambda = \begin{pmatrix} \lambda_1 & 0 & 0\cdots \\ 0 & \lambda_2 & 0\cdots \\ 0 & 0 & \lambda_3\cdots \\ \cdots & \cdots & \cdots\ddots \end{pmatrix} \qquad (2)$$

Wherein each column vector R_i of R is the eigenvector of C, and the diagonal elements $\lambda_i (i \in [1,M])$ of Λ are the eigenvalues of C, arranged in order of size.

5) Select L standard orthonormal vectors as the base of standardization input data. They are unit vectors, called the main components. The input data is a linear combination of the main components. Construct the space transformation matrix: select the feature vectors that the corresponding variance contribution rate μ is greater than a certain percentage to construct the feature matrix E. The variance contribution rate μ is defined as follows:

$$\mu = \sum_{i=1}^{L} \lambda_i \Big/ \sum_{i=1}^{N} \lambda_i,\ L < N \qquad (3)$$

If the first L eigenvalues are selected, the feature matrix E= [N×L], is the first L columns of Λ. The principal component space transformation matrix of CVs is H= P×E^{T}.

6) Principal components are arranged in descending order according to the “importance”. The size of data can be reduced by removing the weaker component (i.e., a smaller variance). Use the strongest principal component should be able to reconstruct a good approximation of the original data.

In this paper, the spectral data is reduced by PCA before classification and clustering. It is found that when the spectra is reduce to three dimension, the variance contribution rate can reach 99%. The contribution rate of the first two dimension is also up to 98%. Fig .1 and Fig .2 shows the 1417 training template on a 2-dimensional and 3-dimensional projection. It can be clearly seen that the 2-dimensional projection of training data has been apparently separate to two types, and the 3-dimensional projection is divided into two parts quite obvious.

IV. EXPERIMENTS

The experiments in this paper were carried out under MATLAB R2010a. The computation environment is as follows:

CPU：AMD PHenom(tm) Ⅱ X 4810
precessor 2.6GHz
Memory：4GB DDR3 1333MHz
Operating System：Windows7

First, we used 15624 spectra instead of 1417. The dataset was reduced to 43 dimensions using PCA. We screen the data with SN_G>10 in the FITS file.

Table 2. showed the candidate CVs, CVs and time consumptions of algorithms of ANN, Bayes, KNN and K-means, respectively. In the algorithm comparisons in Table 2, we can see that ANN，Bayes, KNN find out 10 CVs spectra and the 10 CVs found are the same in the above algorithms. But K-means found none. There were more than 9000 spectra before SN threshold for K-means. The reason for this problem may be caused by the center choice for the clustering. As for the time cost, KNN has advantage over other algorithms. But the amount of candidate spectra were relatively more than other methods.

TABLE II. EXPERIMENTAL RESULTS WITH 3

Item	**ANN**	**Bayes**	**KNN**	**Kmeans**
CVs1	11901	10134	13610	9691
CVs2	15	15	15	0
SN1	58	360	221	0
CVs3	10	10	10	0
Time	33.58 s	39.07 s	2.03 s	4.51 s

Note：

CVs1: Candidate CVs , the number of test data classified by the chosen algorithms.

CVs2: CVs in CVs1.

SN1: The number of CVs after SN screened out of CVs1.

CVs3: Left CVs after SN screening.

Time: Training time.

TABLE III. EXPERIMENTAL RESULTS UNDER 43 DIMENSIONS

Item	**ANN**	**Bayes**	**KNN**	**Kmeans**
CVs1	12763	14438	14516	6949
CVs2	15	15	15	12
SN1	41	48	43	397
CVs3	10	10	10	9
Time	10.836s	4.1527s	0.8360s	3.4919s

From the experiments above, we can see that for 2 or 3 dimensions, there were no significant differences for classification and clustering. When wiping out the data with SN<10, the algorithms can find out the same amount of CVs and the two tables showed that results. In Table2, K-means found out 9 CVs, and in Table 1, K-means found nothing. This means that K-means behaves better under low dimensions, while other algorithms have better stability for higher dimensions.

Table 3. showed that finding CVs based on SN_G , the first three algorithms find out 15 CVs and K-means also found 12 CVs. This led two threads:

1. Sometimes, it is not accurate enough to select data by using SN_G in FITS.

2. The dataset is large by SN_G selection, accounting for about 90% of all the test data, so it is not enough to show the efficiency of the alorithms.

V. PERFORMANCE ANALYSIS

For different dimensional spectra after PCA reduction, different methods including K-means, ANN, KNN and Native Bayes are used for classification and clustering of CVs. The results are shown in Fig .3. Fig .4 shows the comparison of time consumption.

As can be seen from Fig .3, K-means is a suitable method for low dimensional data. As the dimension increased, its accuracy becomes very low, having good results only within the 10-dimensions. Results of ANN are instable from the outset about forty-dimensional, while the results can basically contains all CVs before forty-dimension. KNN is a very stable method, the results have no much volatility with the change of dimension. Native Bayes is quite stable for the same training dataset, all CVs can be found before the dimension about 190. It can be seen from Fig .4 that the time consumption of ANN and Native Bayes increases significantly with the increase of dimension, while K-means and KNN algorithm cost less time, as the time consumption changes few with different dimension.

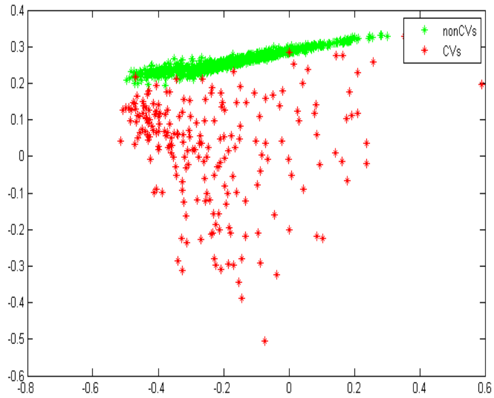

Figure 1. The 2-dimensional projection of 1417 training template.

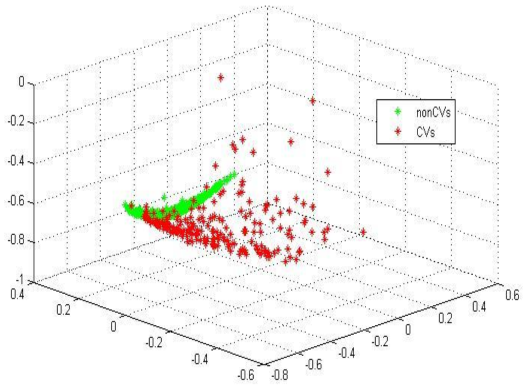

Figure 2. The 3-dimensional projection of 1417 training template

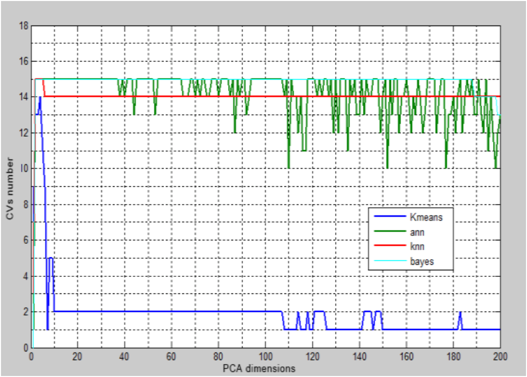

Figure 3. Results of different PCA dimension

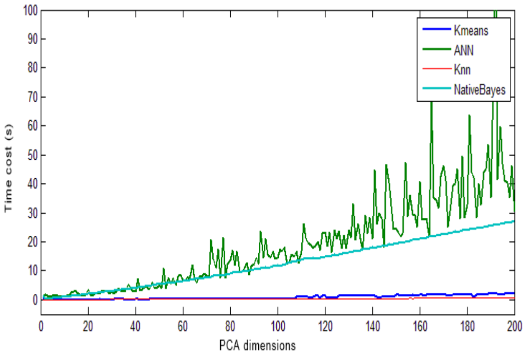

Figure 4. Comparison of time consumption of different algorithm with different PCA dimension

VI. Conclusion

We can conclude that PCA is a very powerful dimension reduction method. It can reduce the multi-dimensional data greatly and effectively. The experimental results also show that different data mining algorithms behave differently in performance under different PCA dimensions. The experimental curves can give us useful references in choosing the four famous algorithms when classifying and clustering in future applications.

Acknowledgment

This study was sponsored by The Natural Science Foundation of China (Project No. U1431102).

References

[1] Deng Shibing, Chen Jiansheng ,Multi-waveband Studies of Catacbrstoic Variables(I), PROGRESS IN ASTRONOMY,1994, Vol.12,No.1, pp42-52.

[2] Deng Shibing, Chen Jiansheng ,Multi-waveband Studies of Catacbrstoic Variables (II), PROGRESS IN ASTRONOMY,1994, Vol.12,No.3, pp229-244.

[3] Szkody P, Henden A, Fraser OJ,etc. Cataclysmic Variables from Sloan Digital Sky Survey. IV. The Fourth Year (2003), The Astronomical Journal, 2005,129: 2386-2399.

[4] Szkody P, Henden A, Agueros M, etc. Cataclysmic Variables from Sloan Digital Sky Survey. V. The Fifth Year (2004). The Astronomical Journal, 2006,131: 973-983.

[5] Qin D M,Hu z Y.Zhao Y H, “A PCA Based Efficient Stellar Spectra Classification Method”, Spectroscopy and Spectral Analysis, vol. 23, 2003, pp.182-186.

[6] Xue J Q. Neural Network and Automated Classification of Spectra.Ph.D Dissertation.National Astronomical Observatories,Chinese Academy of Sciences, Beijing, China, 1999(in Chinese).

[7] D. Arthur, S. Vassilvitskii, “How Slow is the k-means Method”, Proceedings of the 2006 Symposium on Computational Geometry (SoCG), 2006

[8] Rennie J, Shih L, Teevan J, and Karger D, “Tackling the Poor Assumptions of Naive Bayes Text Classifiers”, In Proceedings of the Twentieth International Conference on Machine Learning (ICML), 2003.

[9] Bian Zhaoqi, Zhang Xuegong. Pattern Recognition. Beijing: Tsinghua University Press, 2000

[10] SDSS online information web: http://www.sdss.org/