IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 16, 2019, accepted June 29, 2019, date of publication July 8, 2019, date of current version August 15, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2927251

DoPS: A Double-Peaked Profiles Search Method Based on the RS and SVM

CAIXIA QU, HAIFENG YANG, JIANGHUI CAI, JIFU ZHANG, AND YONGXIANG ZHOU
School of Computer Science and Technology, Taiyuan University of Science and Technology, Taiyuan 030024, China

Corresponding authors: Haifeng Yang (hfyang@tyust.edu.cn) and Jianghui Cai (jianghui@tyust.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant U1731126, in part by the Shanxi Province Key Research and Development Program under Grant 201803D121059, and in part by the Scientific and Technological Innovation Team of Shanxi Province under Grant 201805D131007.

ABSTRACT The double-peaked profiles in spectral data are very rare and valuable for the astronomers. Their recognitions are largely depended on visually inspect. The main reasons for auto-search of such spectra are the complex and distinct characteristics of astronomical data. In this paper, we address the problems by a double-peaked profiles search algorithm (called DoPS) based on relevant subspace (RS) and support vector machine (SVM). First, characteristics subspace is extracted by using the relevant subspace mining algorithm, in which the local density factor λ is particularly defined to measure the data sparsity. The characteristics of double-peaked profiles are represented by using the locations, the interval spaces, and strength ratio of double peaks. Second, the characteristics set is analyzed and grouped into three subsets according to the correlations among the characteristics based on the frequent patterns and rough set theory. Third, the double-peaked profiles search algorithm is proposed by using the support vectors trained from the labeled samples as thresholds. Finally, several spectral data sets from the LAMOST survey are employed to test the DoPS. The experimental results indicate that DoPS presents high performance than other similar algorithms in terms of time efficiency, noise immunity and recall, and reduced rates.

INDEX TERMS DoPS, characteristics extraction, support vector machine, double-peaked profiles.

I. INTRODUCTION

The detection and recognition of special data with rare characteristics, particular from the big dataset, are key steps for the further relative study. For example, the targets harbored with double-peaked profiles in their spectra are extremely rare and very valuable proofs for the astronomical researches. Existing methods to mine out these data are largely depended on visual inspect. In this paper, we address the problem of auto-select the double-peaked profiles from the galaxy spectra. We show that feature extraction scheme can dramatically reduce the data dimensions by obtaining the relevant subspace in the high-dimensional dataset. We also show how to design the algorithm *DoPS*- a detection and recognition of double-peaked profiles on the basis of the validated relevant subspace. And the correctness and efficiency are verified by experiments carried out on the dataset of LAMOST survey.

The associate editor coordinating the review of this manuscript and approving it for publication was Alberto Cano.

A. MOTIVATIONS

Double-peaked profiles recognition of extragalactic galaxy spectra addressed in this study is motivated by the following observations.

- Their is no one useful method provided for recognition of the rare samples from the complex and diverse astronomical data.
- It is hard to build complete and usable templates of double-peaked profiles in spectra. So, it would be meaningful to search and recognize such objects automatically for the astronomers.
- SVM is always a useful binary classification technique, however, it is not suitable for solve the above problems directly.

Motivation 1: There are a lot of new techniques which can be used to search outlier data from uniform dataset, such as pattern recognition algorithms [1], outlier mining algorithms [2], classification algorithms [3], etc. These methods show good performance and applications on the data set from more and more fields [4]–[6]. While, for the specific

background dataset of astronomy, no methods handled very well so far because of the various and distinctive data characteristics. So, to consider these characteristics and develop the methods particularly for the special data will be of significance.

Motivation 2: The spectra with double-peaked profiles are extremely rare and valuable relative to the astronomical big dataset [7], and they would be the import evidence of double black holes [8]. To mine out such spectra from the massive data is the first step in the astronomical research. The double-peaked profiles display various characteristics which greatly improved the search difficulty. For example, the spectra we collected are composition of signals and noises. The weak double-peaked profiles could not be distinguish from the spectra. Other features of the astronomical data can be found in Sec.III-A.

Further, one of the import points is the sparsity of the double-peaked profiles we detect, which would be a useful factor in improving the search quality and efficiency. Meanwhile, they might be correlations among these profiles. So, it is necessary to extract and represent these characteristics in the RS (relevant subspace) [9].

Motivation 3: The search can be categorized to binary classification problems. SVM (Support Vector Machine) is a useful supervised learning method for small and high dimensional samples. For the condition of extremely unbalanced in the ratio of positive and negative samples, the support vectors trained by the small samples can also be particularly used as thresholds for the search of rare objects from the big data set.

B. CONTRIBUTIONS

Searching the spectra with double-peaked profiles from astronomical dataset is important for the further study of universe formation and evolution. And the characteristics of double-peaked profiles show very complex and diverse. In this paper, we focus on this search problem and implement a double-peaked search algorithm based on RS(relevant subspace) and SVM.

The contributions of in this work are summarized as follows:

- The characteristics extraction, identification and representation of double-peaked profiles are provided based on the relevant subspace.
- A classification analysis of characteristics is carried out based on the frequent pattern mining and rough set theory.
- A double-peaked profiles search algorithm based on the SVM and above results is proposed.
- The proposed algorithms and techniques are integrated in *DoPS*, which is evaluated through extensive experiments using the extragalactic galaxy spectra observed by LAMOST survey.

C. ROADMAP

The rest of the paper is organized as follows. Sec.II summarizes the related works relevant to this work. Sec.III gives the theoretical basis and the introduction of astronomical data. The double-peaked profiles search method named DoPS is detailed in Sec. IV. Then, Sec.V discusses the experimental settings as well as the results. And finally, a summary and acknowledgement of this work is given in Sec.VI and Sec.VI respectively.

II. RELATED WORKS

Double-peaked profiles are composed of two or more peaks in specific range in spectra, which reflect rare universe evolution. Galaxies and QSO(quasar) are provided to investigate double-peaked profiles, which may be produced by AGNs [10], double SMBHs [11] and other machines, such as jet-cloud interaction, chance superposition, and kinematics in the narrow-line regions [12]–[14]. As telescope technique updates, data on different wavelength bands can be obtained. The newest researches for double-peaked profiles spectra are based on radio wavelength bands. High-solution radio observations is used to search double-peaked profiles spectra, which is a good way to confirm double-peaked AGNs [15]. The observed candidate samples show that one of them is dual AGNs, and ones of remnant are produced by other machines. Although people focus on radio bands, double-peaked profiles are still occurred frequently on the optical bands. Spectra from LAMOST(Large Sky Area Multi-object Fiber Spectroscopic Telescope) on optical bands provide galaxies and QSO source to research for double-peaked profiles. The spectra with double-peaked profiles in LAMOST are searched efficiently. There are 20 candidate spectra in LAMOST DR1 are found by Shi *et al.* [16]. And Wang *et al.* search 325 candidate spectra with double-peaked profiles in LAMOST DR4 [17]. With improvement of accuracy of observation instrument, size of the spectra data obtained is increasing year by year. So we have to present various ways to mining meaningful knowledge from massive data. Different viable methods are applied on searching spectra with double-peaked profiles. In order to confirm extragalactic galaxies, machine learning based on spark parallel is used to identify double-peaked profile of Hα line [18]. Double-peaked [OIII] and Balmer profiles are searched by statistic analysis to find candidate binary AGNs in SDSS(Sloan Digital Sky Survey) in [19] and [20] respectively. In addition, a cross-correlation technique also is proposed to detect complex(double-peaked or multiple components) profiles of [OIII], avoiding observation of large dataset with eyes [21]. And the method based on multi-gaussian fit of narrow profiles is applied for searching galaxies and AGNs with double-peaked narrow profiles [16]. It can be found that these methods are more based on mathematics than data mining using for searching double-peaked profiles. Some algorithms may be applied in detection of spectra with special characteristics, such as clustering, classification, outlier detection, association rules, in which classification algorithm with labelled data is supervised. And we are interested in SVM(support vector machine), which could classify data into either one class of two classes.

Pattern recognition is a newer area in computer, which can analyze and process characteristics information's. The goal of pattern recognition is to build a procession of description, analysis, classification and explanation for describing data. There are different based methods in pattern recognition, such as structure pattern recognition, statistic pattern recognition, artificial neural network pattern recognition, fuzzy pattern recognition, and so on. Pattern definition methods are applied in various industries. In [22], DNA-based neural networks is implemented to recognize molecular patterns for biological organisms. Winner-talk-all computation is suggested to enhance the capability of DNA-based neural networks avoiding less pattern numbers. The proposed method can recognize nine patterns consist of 20 different DNA molecular. Pattern recognition is also used in high performance imaging with high spectral and spatial resolution, high dynamic range, and/or at high speed [23], in which some tasks of pattern recognition are introduced, such as processing of images, depth estimation, data reconstruction, classification object detection and recognition. Different high performance images are involved, including hyper-spectral, depth field, RGB-D, and magnetic resonance. In addition, high-resolution real-space imaging of nano-particle also adopts pattern recognition algorithm, which is to simulate topography images from density functional theory using the known total structure of the Ag374 nano-cluster protected by tert-butyl benzene thiol [24]. Driving pattern recognition based on feature extracted with multi-layer perception neural network is used to build an adaptive energy management controller, which is for real time power split between fuel cells and super-capacitors [25]. After the driving pattern recognition is obtained, some factor are optimized using genetic algorithm to improve management controller. Pattern recognition is applied in science researching above these methods, and it also can be used for daily life. Malicious software detection on smart-phones based pattern recognition focuses on recognizing malicious software addressed against Android platform, which is to prevent installation software in victim systems [26]. There are three processing layers in pattern definition including monitoring, analysis and decision-making. And the experimental results demonstrate this proposed method is a great of detecting malicious software for most smart-phones on Android system. In some methods of pattern recognition, we take more attention on classifications.

Outlier detection aims to find points which are different with other points in dataset. The outlier in dataset may involve some useful information we need, which is meaningful in big data to search it. The methods for outlier detection are proposed based on two approaches: local outlier factor(LOF) and principal component analysis(PCA). A relevant subspace method is used to detection of contextual outlier, which is used to search contextual outliers using local outlier factors [47]. For high-dimensional dataset, a local projections method based on LOF and RobPCA is proposed as a novel approach, in which the quality of local description interpretation might influence the observations in local projection [44].

Classification methods are almost used to predict class for a given data. Supervised classifications train a classifier model via training data with known label, then unknown data is assigned into one or more classes by classifier. Several classification algorithms are applied to predict musical preference of a person with four features [27]. In the prediction, binary-classification is used to created a dataset, which is examined by different classification algorithms including cart, knn, random forest. The result shows that emotion factor should be considered into dataset features in algorithms. The dataset with label in supervised classification and unsupervised classification is also researched in fields. For searching various stars demanding less human efforts, an unsupervised classification of various stars is proposed as an untraditional method, which relies only on the similarity among light curves [28]. The results of experiments show high performance of different types of various stars. In addition, the unsupervised classification based on k-means algorithm is applied to divide stellar data into discrete classes, in which the proposed method is able to separate the bulge and halo populations, distinguish dwarfs, sub-giants, RC and RGB stars [29]. The neural network is also used to classify in artificial intelligence field. ANN based on the chromosome classifiers is the main topic in the survey, which provides comprehensive review of past and recent research in the area of automatic chromosome classification systems [30]. And the classifiers based on other algorithms are also expected to employ in chromosome classification as additional techniques.

As a supervised pattern classifier, SVM(Support Vector Machine) was supposed to maximize the margin between decision boundary and training pattern by Boser *et al.* [31]. Since then, the algorithms based on thought of support vector are presented consequently. The one of most important part of SVM is kernel function which aims to map input data into a high-dimensional feature space. The data points are linearly separable in high-dimensional space by choosing appropriate kernel mapper as much as possible. So Amari *et al.* modified kernel functions to improve performance of SVM based on structure of the Riemannian geometry in 1999 [32]. In [33], results of SVM with four different kernel functions(linear, polynomial, radial, sigmoidal) are compared to get best performance model in landslide susceptibility maps. Multiple kernel functions are combined to develop a new and effective kernel, showing power and usefulness of this approach [34]. And the kernel classification algorithm based on binary SVMs is introduced to solve QP optimization problem [35]. The improvement or innovation of SVM are most based on modification of kernel functions [36], [37]. Comparisons about SVM are between not only inner kernels, but also different classifications, including random forest, k-nearest neighbor [38], in which the SVM produces highest overall accuracy than other algorithms. SVM is more and more widely applied in various fields, such

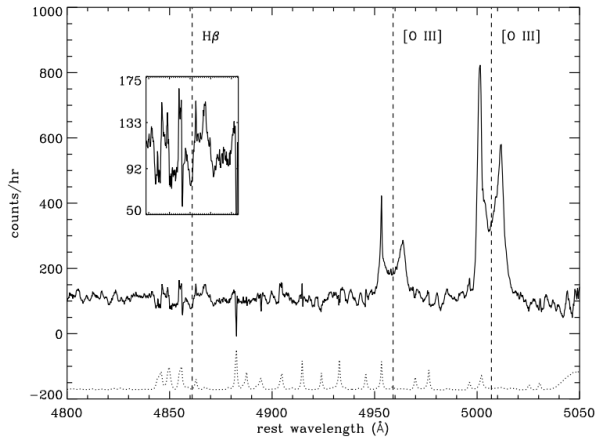


FIGURE 1. An instance of double-peaked profiles.

as identification of power transformer fault [39] and digital images [40], detection of intelligent chatter [41], allocation of wireless networks [42] etc. It is obvious that SVM algorithm with other techniques shows better performance and application in different fields.

In the latest works, the proposed methods show better higher performance in different fields. However, there are not approach used in specific background of astronomy. Developing a new method for searching double-peaked profiles data is meaningful due to the characteristics and distinct of dataset.

III. PRELIMINARIES

A. DATA DESCRIPTION

Some standard datasets for classification are existed, however, they are submitted without double-peaked profiles spectra. Spectra of LAMOST is different with other datasets due to their particularity. Thus the dataset used in this paper is constructed manually, in which positive samples and negative samples are included. Spectra in this paper are obtained from LAMOST, which is a large field of view and large aperture telescope. There are 4000 filers on focal plane of LAMOST, can obtain 4000 spectra at the same time. LAMOST becomes a powerful tool to observe sky with high spectra acquisition rate. Observation of celestial objects over total available northern sky is the main tasks of LAMOST. In LAMOST survey, two components are conducted - LEGUE (LAMOST Experiment for Galactic Understanding and Exploration) and LEGAS (LAMOST Extra Galactic Survey). LEGUE focuses on Galactic anti-center, the disk at longitudes away from the Galactic anti-center, and the halo. As another major part of LAMOST survey, LEGAS includes galaxy survey and QSO survey. Number of galaxies and quasars is lesser than stellar due to the higher red-shift of extra galactic objects. After LAMOST DR5 survey is finished, about 15,000 galaxies and 52,000 QSO are obtained, are the source data of our research.

Profiles of double peaks are composed of less than two peaks generally. An example is shown in Figure 1, where the abscissa is wavelength values in the rest-frame wavelength, and the ordinate presents flux which is consistent with wavelength. It can be seen that the characteristics of double-peaked profile are obvious in Figure 1, and weak characteristics on other lines are needed to search in particularity. Different forms of double-peaked profiles are existed in spectra due to different physical machines. So the detection of spectra with double-peaked profiles is difficult relatively.

The spectra with double-peaked profiles display various distinct characteristics, which can be summarized as below:

- Basic information. The basic information of extragalactic spectra in LAMOST dataset include that the resolution is R $\approx$ 1800, the wavelength coverage is 3800Å$\sim$9000Å. Therefore, traditional artificial identification as become impossible.
- Massive and high dimensional. After a five-year survey, the first stage of observation task has ended. More than 100 million spectra are collected. And for each spectrum, the data dimension is up to 4000.
- The characteristics are sparse. The double-peaked profiles are always shown on the emission lines of the spectra. In general, the dimensions of the emission lines covered are very small relative to the total dimension. Sometimes, the spectra only show a part of the emission lines.
- Composite spectra with signals and noises. The astronomical spectra, especially for the galaxy spectra observed by the ground-based telescopes, have not only the components from the target celestial body, but also other ingredients (noises) such as from cosmic rays, earth atmosphere, data precessing errors, etc. They will increase the recognition difficulty of the main characteristics.
- Double-peaked profiles vary a lot. Firstly, influenced by the Doppler effect, the emission lines would be shifted to other wavelengths. That is, the stations of characteristics appear in spectra are uncertain. Secondly, the shapes of double-peaked profiles are varied. Some spectra show stronger left(blue) peaks, some show stronger right (red) ones, and some show strange shapes relative to Gaussian profiles. In addition, the double-peaked profiles are not shown in all expected locations in the spectra. These uncertainties are the key problems in construction of templates and recognition of double-peaked profiles.

B. SUPPORT VECTOR MACHINE

Support vector Machine(SVM) is a supervised machine algorithm of two-classification, in order to find optimal hyperplane with maximal margin between separating hyperplane and other data. SVM can train a prediction model using given training data firstly. Then category class of unknown data is calculated with higher accuracy than other machine learning algorithms. If kernel functions were applied in support vector machine, non-linear data would be divided into two classes linearly. In this paper, spectral from LAMOST are used in algorithm with line kernel function.

Now, there is a given train set includes train data $X = \{\boldsymbol{x_i}\}$ and train label $Y = \{y_i | y_i = \pm 1\}$, i=0,1,...,n. All data are

constrained as follows:

$$y_i(\boldsymbol{wx_i} + b) \geq 1 \qquad i = 1, 2, ..., n \tag{1}$$

Parameters $\boldsymbol{w}$ and y_i are weight vector and class label of data $\boldsymbol{x_i}$ respectively in (1). Support vectors satisfy (2), including positive and negative vectors at above and lower part.

$$\begin{cases} \boldsymbol{wx_i} + b = 1 & y_i = 1 \\ \boldsymbol{wx_i} + b = 1 & y_i = -1 \end{cases} \tag{2}$$

The predict function of unknown class data is followed (3). Data x is belongs to positive sample if f(x) is closed 1. By contraries, it is geared another class.

$$f(x) = \boldsymbol{wx_i} + b \tag{3}$$

SVM aims to find maximum margin between positive and negative regarded as support vectors $2/ \parallel \boldsymbol{w} \parallel$. Thus, the task of margin maximization of SVM can be defined as:

$$\begin{aligned} &\min \ \frac{1}{2}\|\boldsymbol{w}\|^2 \\ &s.t.\ y_i(\boldsymbol{wx_i} + b) \geq 1 \end{aligned} \tag{4}$$

It is a dual problem in (4), which can be transformed by Lagrange equation as:

$$\begin{aligned} &\max_{\alpha} \ \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j (\boldsymbol{x_i} \cdot \boldsymbol{x_j}) \\ &s.t. \begin{cases} \sum_{i=1}^{n} \alpha_i y_i = 0 \\ \alpha_i \geq 0 \qquad i = 1, 2, ..., n \end{cases} \end{aligned} \tag{5}$$

where α_i denotes Lagrange multiplier, which can be confirmed by optimized algorithm.

It is assumed that data are linearly separated in theory. However, the most data are non-linearly in reality. The kernel function is proposed to map original data space to a new space, in which the data could be separated linearly. SVM can learn a linear classification model in new space by employing kernel function. The object function of dual problem with kernel function as:

$$\begin{aligned} &\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(\boldsymbol{x_i}, \boldsymbol{x_j}) \\ &s.t. \begin{cases} \sum_{i=1}^{n} \alpha_i y_i = 0 \\ \alpha_i \geq 0 \qquad i = 1, 2, ..., n \end{cases} \end{aligned} \tag{6}$$

where $K(\boldsymbol{x_i}, \boldsymbol{y_i})$ is a kernel function, is most obtained from expert experience. The frequently-used kernels include linear kernel, Gaussian kernel, sigmoid kernel as (7), (8), (9) respectively.

$$K(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x} \cdot \boldsymbol{y} \tag{7}$$

$$K(\boldsymbol{x}, \boldsymbol{y}) = \exp(-\frac{\|\boldsymbol{x} - \boldsymbol{y}\|^2}{2\sigma^2}) \tag{8}$$

$$K(\boldsymbol{x}, \boldsymbol{y}) = \tanh(\alpha \boldsymbol{x}^t \boldsymbol{y} + c) \tag{9}$$

It is difficult to find an optimal hyperplane to divide data into different classes absolutely, even though kernel function is applied. The parameters C and ξ are introduced to access the problem. C is the penalty coefficient presenting misclassified punishment. The greater C value, the greater penalty of misclassified. Slack variable ξ is added into every data to tolerate exception point. Introduction of the new parameters considers the existing of exceptions, which has influence on classification results. Thus object functions (4) and (5) are changed into (10) and (11) respectively.

$$\begin{aligned} &\min \frac{1}{2}\|\boldsymbol{w}\|^2 + C \sum_{i=1}^{n} \xi_i \\ &s.t. \begin{cases} y_i(\boldsymbol{wx_i} + b) \geq 1 - \xi_i \\ \xi_i \geq 0 \end{cases} \end{aligned} \tag{10}$$

where C and ξ_i are used to determine the strength of punishment for exception points.

$$\begin{aligned} &\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(\boldsymbol{x_i}, \boldsymbol{x_j}) \\ &s.t. \begin{cases} \sum_{i=1}^{n} \alpha_i y_i = 0 \\ 0 \leq \alpha_i \leq C \qquad i = 1, 2, ..., n \end{cases} \end{aligned} \tag{11}$$

The application of kernel function and penalty coefficient is helpful for improving accuracy of SVM classification model. As a two-class classifier, SVM is suitable for detecting double-peaked profiles spectra, which can be classified into one class with special characteristics. The characteristics of double-peaked profiles can be added into training process of SVM. From this, the training classification model with double-peaked profiles is more targeted.

IV. SEARCH METHOD

DoPS method is proposed as new approach to search double-peaked profiles in this paper. Characteristics extraction of double-peaked profile spectra based on relevant subspace is worked to reduce dimension of data. The subspace which is relevant with double-peaked profiles is named characteristics subspace after expert certification. In Characteristics subspace, a classification of characteristics is implemented by frequent patterns and rough set theory. Then, the classification method DoPS is used to test the performance and availability according to recall rate, reduced rate, accuracy.

A. CHARACTERISTICS EXTRACTION BASED ON RELEVANT SUBSPACE

In high-dimensional data sets, it is impossible that all dimensions are attribute for data mining. It is meaning that some of all dimensions are helpful for our searching, while others should be abandoned actually. Selecting dimensions with useful information is necessary and feasible. Relevant subspace can efficiently find the local distribution of various data sets, which is defined by local sparseness of attribute

dimensions [47]. The most frequent attributes with maximum density are included into relevant subspace. Determine if an attribute is regarded as a member of relevant subspace dependent on density of object's local data set which is defined by nearest neighbor.

Determination of relevant subspace is based on computation of local dataset(LDS). It is assumed that DS is a d-dimensional dataset with length n, $FS = \{F_1, F_2, ..., F_d\}$ is a characteristics space, and x_{ij} denotes the ith data on jth dimension. The basic concept of relevant subspace is the uniformity of local dataset on dimensions. LDS(x,F) is a local dataset including nearest neighbors of object x on attribute F. If F is a relevant subspace, distribution of LDS(x,F) is non-uniform. In contrast, if F is not a relevant subspace, LDS(x,F) is a uniform distribution, which can efficiently embody the valuable information of double-peaked profiles. However, the uniform distribution cannot provide meaningful information of double-peaked profiles. So measurement of local sparsity on dimensions is a key step of obtaining relevant subspace.

The sparsity or density of local dataset LDS of object reflects an attribute distributions. The distribution of attributes is more uniform, the density of local dataset is more similar. Relevant subspace removes dimension with lower density of LDS and receives dimension with higher density. The density of object on each dimension is defined by the local density factor λ_{ij}:

$$\lambda_{ij} = \frac{\sum_{y \in p(x_{ij})} (y - x_{ij})^2}{k + 1} \tag{12}$$

where λ_{ij} denotes the local density factor of ith data on jth dimension. $p(x_{ij})$ is a sequence of $LDS(x_{ij}, F_j)$, which is composed of k nearest neighbors of object x_{ij} on the attribute dimension F_j. We can know λ_{ij} is the mean square value of Euclidean distance between object x_{ij} and local dataset on dimension F_j. The local dataset LDS will be more sparse if the λ_{ij} is larger. On the contrary, the smaller the λ_{ij}, the more dense of LDS. λ_{ij} of each object on attribute dimensions is computed to generate a density factor matrix $[Z_\lambda]_{n \times d} = [\lambda_{ij}]$. The local matrix on dimensions $[Z_\lambda]_{k \times d}(obj)$ is obtained from the matrix, where k is length of LDS(obj).

For further describing the difference of LDS's local density on dimensions, the local density difference factor of LDS(obj) on jth attribute dimension δ_{ij} is defined:

$$\delta_{ij} = \sqrt{\left| \frac{\lambda_{ij} - C\lambda_{ij}}{C\lambda_{ij}} \right|} \tag{13}$$

where $C\lambda_{ij}$ is the mean square of $[Z_\lambda]_{k \times d}(x_i)$ on attribute A_j. And δ_{ij} reflects the difference of local density. δ_{ij} is larger, degree of difference of local density factor larger, λ is larger, then local dataset LDS(obj) on dimension is not uniform. Conversely, the smaller the δ, the smaller degree of difference of local density factor, the smaller the λ, then local dataset LDS(obj) on dimension is uniform. In a short, the local dataset of each object on attribute dimensions is uniform or not uniform is according to δ. Thus the uniformity of local dataset on dimensions can directly be judged via δ, which is an important step in characteristics detection using relevant subspace.

To measure whether the attributes dimension and characteristics of double-peaked profiles are related, a threshold of local density difference factor α is given. Then a subspace $V = \{V_1, V_2, ..., V_j, ..., V_d\}$ is built, where d is the length of dimensions. And $V_j = \{V_{1j}, V_{2j}, ..., V_{ij}, ..., V_{nj}\}$, where n is the size of dataset. The definition of V_{ij} is:

$$V_{ij} = \begin{cases} 1 & \delta_{ij} \leq \alpha \\ 0 & else \end{cases} \tag{14}$$

where V_j value is 1 or 0. If $V_{ij} \leq \alpha$, $V_j = 1$, the jth dimension is related with double-peaked profiles, being a member of relevant subspace. On the contrary, the jth dimension will be removed into non-relevant subspace if $V_j = 0$. The relevant subspace is obtained by selection of relevant dimensions in V.

The description algorithm of characteristics extraction based on relevant subspace is shown in Algorithm1.

Algorithm 1 Characteristic Extraction Based on Relevant Subspace

Input: Training dataset DS; number of neighbors K; local density difference factor threshold α.
Output: Characteristics subspace

1: Data Selection. Double-peaked profiles spectra which red-shift<0.3 are selected as training data;
2: Data preprocessing. Training data are in rest wavelength, and flux are normalized.
3: Wavelength bands selection. Flux data between 4000Å∼5700Å and 5900Å∼5700Å is selected as the valid training data.
4: **for** each dimension in characteristics space **do**
5: **for** each object in dataset **do**
6: Local data set of the object in the dimension is obtained using k nearest neighbor algorithm;
7: Compute the local density factor of the object λ_{ij} according to (12);
8: Calculate the local density difference factor δ_{ij} according to (13);
9: Relevant attributes are gained according to (14), and characteristics subspace is obtained;
10: **end for**
11: **end for**
12: Return characteristics subspace

In the characteristics extraction algorithm, KNN is firstly used to get a local dataset. Then the attributes distribution of every object on each dimension is measured by the local density. Finally the relevant dimensions are obtained from generated subspace, the V_j value is determined by voting method. In the algorithm, KNN is used for computing the local dataset of objects on dimensions. The time complexity

of KNN algorithm is $O(n \cdot logn)$. The local density factor λ and local density difference factor δ are descriptions of local dataset on dimensions to get relevant subspace. The time complexity of the function of computing λ is $O(n \cdot d \cdot k)$, and so do δ function's time complexity. Therefore, the time complexity of characteristics extraction is $O(n \cdot logn + n \cdot d \cdot k + n \cdot d \cdot k)$.

The members in relevant subspace are analyzed by reading various papers and expert certification. The characteristics subspace is obtained after the analysis of relevant subspace.

B. CHARACTERISTICS CLASSIFICATION BASED ON FREQUENT PATTERN AND ROUGH SET THEORY

1) ASSOCIATE RULES MINING

Frequent pattern mining aims to find different association rules, which can help users to make a decision in a base area. For example, shopping black analysis is to find the relations between various goods, then customer's shopping habits are analyzed to produce a better marking strategy.

Support and confidence are two measurements in association rules, which denote the usefulness and determinacy of associations respectively. Suppose an association $A \Rightarrow B$ with support s and confidence c, the association can be presented by a probability formula:

$$support(A \Rightarrow B) = P(A \cup B) \tag{15}$$

$$confidence(A \Rightarrow B) = P(B\,|A\,) \tag{16}$$

where the support of association $A \Rightarrow B$ is the probability in dataset. Confidence can obtained via the support of A and $A \cup B$ as:

$$confidence(A \Rightarrow B) = \frac{support(A \cup B)}{support(A)} \tag{17}$$

The Apriori algorithm is a typical method for mining frequent items. It adopts an iterative method for layer by layer search. The priori property which is applied in Apriori improves performance of searching frequent items. The association with larger confidence is regarded as a strong association rule. There are two steps of finding association rules:

- Finding all frequent items, whose occurrence number is at least larger than set min support;
- Producing the strong association rules, which satisfy the demand of min support and min confidence.

The algorithm 2 shows detailed process of finding association rules.

The reliability of association rules can be measured by the confidence. The larger confidence, the more credible of association rule. As all we known, the characteristics of double-peaked profiles are associated with some wavelength bands of line. So the association rules between lines can be used for grouping basis of characteristics subspace according to the confidence. The strong association rules satisfied min support and min confidence can directly be obtained when the frequent items of database are found. On the one hand, the candidate items are produced by the connecting frequent items, which will cause a memory problem in big size of database. on the other hand, every time the supports of candidate items are counted, the database is scanned in Apriori algorithm. So the method of finding association rules is only available for smaller size of database. In our case, not all strong association rules are received as good works, the n associations rules with largest confidence are selected as the remembers of strong association rules. The selection method can effectively obtain association rules between characteristics of double-peaked profiles and wavelength bands.

Algorithm 2 Finding Association Characteristics in Subspace

Input: Database in characteristic subspace, min support threshold ms

Output: Association rules

procedure frequent items processing
    Every item in database is regard as the candidate 1-item;
    Scanning the database to count the supports of candidate 1-item, then the frequent 1-items are obtained whose supports are bigger than ms;
    for k from 2 to length of dataset **do**
        Candidate k-items are produced by connecting the frequent (k-1)-items;
        Removing the non-frequent items from candidates using the priori property;
        Frequent k-items are obtained by comparing supports with ms;
    end for
end procedure
procedure finding associations rules
    for Association rule in frequent items **do**
        Computing the coefficient of association rules according to (17);
        Association rules with coefficient are recorded;
    end for
    Selecting the first 5 association rules with bigger coefficient as the strong association rules;
    Return association rules.
end procedure

2) ROUGH SET THEORY

The rough set theory is proposed based on incomplete data and knowledge which cannot be conveyed, learned and concluded precisely. The main idea of rough set is to approximately describe the uncertain or undefined knowledge using known knowledge base. This theory don't need any priori information's outside dataset to be processed. So the rough set is more objective on description of knowledge than other methods.

The rough set theory is based on the establishment of equivalence class. Suppose X is a subset of a given non-empty finite field U named $X \subseteq U$, and R is an equivalence

relation. X is R-definable if X can be described precisely by R attribute set. That means X can be represented using the union of R basic set. Contrarily, X is not R-definable. The R-definable set named R exact set can be defined precisely in the knowledge base. And the R non-definable set is named as R rough set. For searching the case of non-resolvable equivalence class of every set $X \subseteq U$, the rough set is defined approximately by two exact sets, including the upper approximate set and lower approximate set.

$$R\_(X) = \cup\{Y_i \in U \,|ind(R) : Y_i \subseteq X\} \tag{18}$$

$$R^{-}(X) = \cup\{Y_i \in U \,|ind(R) : Y_i \cap X \neq \emptyset\} \tag{19}$$

The two functions in (18) and (19) are lower approximate set and upper approximate set respectively. The $R\_(X)$ is composed of sets can be included in X. $R^{-}(X)$ is composed of sets can or may be included in X. Namely, $R\_(X)$ is the union of all Y_i included X and $R^{-}(X)$ is the union of Y_i whose intersection is 0 with X. $bn_R(X)$ is a dataset which cannot be sure to fall into X and -X. $pos_R(X) = R\_(X)$ is named the R positive field of X, $neg_R(X) = U - R\_(X)$ is also named the R negative field of X, and $bn_R(X)$ is named the boundary field.

$pos_R(X)$ or lower approximate is a set can totally included in set X. Similarly, $neg_R(X)$ is a set can included in set X non-certainly, which is the complementary set of X. The boundary set is regarded as a non-certain field which may be included in X. So the upper approximate is the union of positive field and boundary field:

$$R^{-}(X) = pos_R(X) \cup bn_R(X) \tag{20}$$

Rough set can be used for classifying, selecting attribute set and correlation analysis. The initial purpose of rough set is based on analysis and inference of data, finding hidden information or knowledge. It has become a valid method to process imprecise, in consistent, incomplete data, which profit from the mature mathematics foundation and non-necessity of priori knowledge. We can apply the association rules in rough set to get lower and upper approximate. The results of rough set will be basis of working in grouping. The characteristics subspace is associated with double-peaked profiles, however, the strengths of relation are different for each member in characteristics subspace. So the rough set is an effective way to help users divide wavelength bands which in characteristics subspace.

C. A SEARCH METHOD DOPS BASED ON SVM

1) MAIN IDEA

Support vector machine is a two-class classify, which can gain a classification model by training dataset. The formula of predicting class of a unknown data in support vector machine is $f(x) = \boldsymbol{wx} + b$, where $\boldsymbol{w}$ is the weight vector of data. The expending classification function is:

$$f(x) = w_1x_1 + w_2x_2 + \ldots + w_ix_i + \ldots + w_mx_m + b \tag{21}$$

where $\boldsymbol{w_i}$ is computed by optimizing method based on initial data. However, these coefficients are used to make prediction for the weak characteristics of double-peaked profiles. So we should set extra coefficient to enhance the strength of optimized coefficient.

Every member in characteristics subspace has different weight for double-peaked profiles. The weight of subspace is obtained by statistics as the probability of members in characteristics subspace. Suppose the probability set $P = \{P_1, P_2, \ldots P_i, \ldots P_L\}$, where L is the length of characteristics subspace, P_i is the probability of occurring double-peaked profiles on ith member in subspace. In a given dataset with double-peaked profiles, statistic method is used to computed the probability. The value of P_i is defined:

$$P_i = \frac{num(F_i)}{S(data)} \tag{22}$$

where $num(F_i)$ is the number of data with double-peaked profiles on ith characteristics subspace in dataset. And S(data) is size of dataset. The probability P_i actually is the frequency of occurrence for the ith member in characteristics subspace. We can know the probability on every characteristics subspace by set P.

The degree of influence on characteristics of double-peaked profiles is different, so the P value is applied in training classification model. Every characteristics subspace with a corresponding probability P_i is used in training process as follows:

$$f(x) = \sum_{i=1}^{L}\sum_{j=1}^{m} P_i w_{ij} x_{ij} \tag{23}$$

where L is length of characteristics subspace, P_i is the probability of ith member in characteristics subspace, m is the length of data on one characteristic. We determine the class label of a given object by approximate if f(x). If f(x) value is closed 1, the object is included in positive samples. On the contrary, the object is belonged to negative samples if f(x) value is closed to -1.

Support vector machine trains a classifier for double-peaked profiles using probability of subspace. The parameter space of SVM includes C(penalty coefficient), I(max number of iterating) and the type of kernel function. The optimization process of iterating has to stop when the iteration number is reached I, although optimized value has not obtained. So the initial value of I can be greater to get optimum solution, which is a global optimize in optimization procedure. The several results of SVM programming may be various because the random α is selected to begin optimization. The best result is our better selection for showing support vector machine.

2) THE ALGORITHM DESCRIPTIONS

To build a template for identifying double-peaked profiles data, a method named DoPS is proposed based on support vector machine. The description of DoPS algorithm based on SVM is shown in Algorithm 3.

Algorithm 3 DoPS Algorithm Based on SVM

Input: Training dataset DS; testing dataset; penalty coefficient C; max iterations I; selection of kernel function.
Output: Training model.

```
for e doach characteristic in characteristic subspace
    Matrix a training dataset;
    Generate a kernel value matrix using selected kernel function;
    Generate Lagrange multiplier matrix composed of initial α;
    while iterations number < I and α no changed do:
        Optimize α value;
        Update α over all training samples ;
        Update α over samples which are not on boundary;
        Compute value b using α according to constraint condition;
        Update α by probability using (23);
        Return value b and α.
    end while
end for
```

3) ALGORITHM ANALYSIS

In Algorithm 3, the training procedure is the main body of algorithm, in which the optimization process is the most important step. The initial Lagrange multiplier α is set 0 in α matrix. The optimization according to the constraint condition is begun from a random α value. The time complexity of the algorithm for linear-data is $O(n \cdot d)$, where n and d are the size and number of attribute dimension of training dataset respectively. The non-linear samples are used in DoPS, which has a time complexity $O(n^2 \cdot d)$. Penalty C is a most influential parameters in SVM. The larger C value, the smaller ξ_i, the smaller distance between hyperplane, then the case of over-fitting in training procedure may be occurred. On the contrary, the smaller C value, the larger ξ_i, the greater distance between hyperplanes, so under-fitting is going to happen. In addition, the over-fitting also can arise when the selection of kernel function is more powerful.

Characteristics subspace is obtained by the characteristics extraction based on relevant subspace. On the other hand, the dimension of high-dimensional dataset is greatly reduced. The DoPS algorithm on lower-dimensional dataset runs with less time than original dataset. So our algorithm is available with higher efficiency.

Different classification results may occur due to the random beginning of optimization. A number of results can be received as the determination candidates. There are two solutions relative to different results:

- Different α values are obtained in optimizing processing, which could be repeated some times. The mean of a number of α is selected as the final α.
- Different classification results are obtained by different Lagrange α. We can select the most frequent results in a number of results as the final classification result.

V. EXPERIMENTS RESULTS ANALYSIS

We evaluate the performance of Dops method at three processors in which the one is Intel(R) Core(TM) i7-6500U-2.50GHz processor with 12.0GB memory, and Windows 7 operating system is installed.

A. DATA SELECTION AND PREPROCESSING

In this paper, the dataset in our method is obtained from two papers about searching double-peaked profiles. 20 double-peaked profiles spectra are searched in LAMOST DR1 [16], among them, 10 have published by others and the rest 10 is first discovered. 325 spectra with double-peaked profiles are presented in LAOMST DR4 [17]. There are 345 spectra with double-peaked profiles in LAMOST, among them 4 spectra without red bands due to bigger red-shift. The known positive samples will be used in training data in follow-up works.

The datasets used in characteristics extraction and DoPS are known certified double-peaked profiles data and selected randomly from LAMOST respectively. The data need be preprocessed before application in method, the main steps of preprocessing are as follows:

- Ensuring the wavelength is on rest-frame, in which the red-shift is used to be implied;
- Data normalization. The dataset is in a common scale by normalizing processing;
- Cutting wavelength. The flux of wavelength between 4000Å∼8000Å are cut as the useful dimensions. In addition, the blue and red bands are connected on 5700Å∼5900Å, where the information can also be cut.

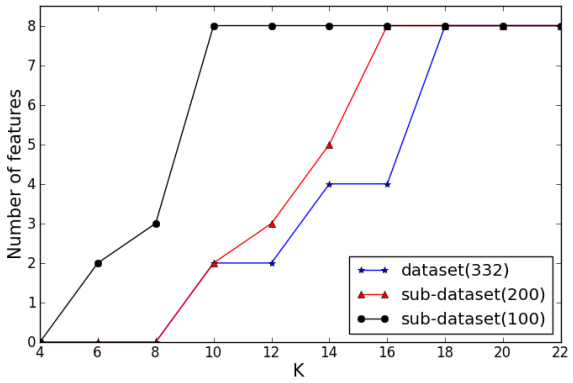
B. CHARACTERISTICS EXTRACTION

1) PARAMETERS TESTING α AND K

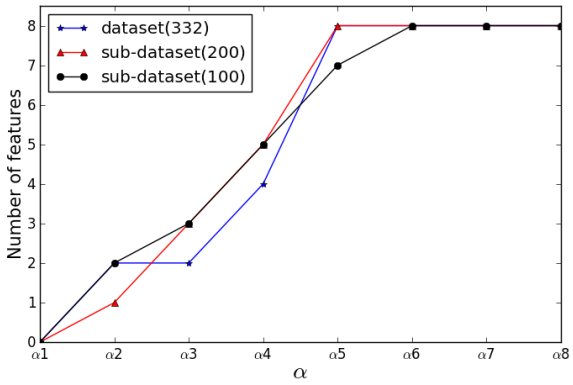
In the characteristics extraction method based on relevant subspace, KNN algorithm is implied to restrain the involved spectra data. Three datasets are selected to ensure the number of members in relevant subspace.

The number of extracted characteristics is obtained by comparing different results with the parameters K and α. To make sure the finally number, different K and α are used in experiments. Figure 2 shows the number change of characteristics with parameters. The Figure 2(a) and (b) are trend of number of relevant subspace with K and α respectively. Three datasets with double-peaked profiles are used in the characteristics extraction. The first dataset including 332 spectra are certified by researchers. And the second and third are randomly selected form the first sample including 200 and 100 spectra respectively. The datasets with different size are used to decide that the final length of relevant subspace.

In Figure 2, the left sub-graph is a decision trend of number about characteristics when K is from 4 to 22. The number is increased from 0 with bigger K, and it converges to 8 for the three datasets. K determines number



(a) Number of characteristics in relevant subspace with K



(b) Number of characteristics in relevant subspace with α

FIGURE 2. Decision for number of extracted characteristics.

of extracted characteristics as one of parameters of characteristics extraction. The smaller k value, the denser distribution of extracted attributes, but the smaller continuous points in centralized position and discontinuous points.

The right sub-figure in Figure 2 shows a number trend with different α, which are various due to size of dataset. For the three datasets, α is fixed to 0.9996, 0.9992, 0.9978 respectively. In the right sub-figure, the number of characteristics tends to 8 with α, which is included in a list from α 1 to α 8. In the α list, the values are getting bigger belong the abscissa. Value α has influence on number of relevant attributes. The larger α, the more relevant attributes, the more discontinuous attributes outside centralized position. On the contrary, the opposite result will be happen.

It is concluded from Figure 2 that the relevant subspace is obtained including 8 members by observing the number trend of characteristics with various experiments. To determine the members in relevant subspace, a characteristics distribution of a spectrum is shown in Figure 3 when K=18, $\alpha = 0.9996$.

There are 3 parts in Figure3. The red points included in a rectangle frame in the top part are extracted attributes, including continuous and discontinuous points. The extracted attributes are focused on two wavelength ranges of 4800Å∼5100Å and 6540Å∼6750Å in Figure 3. The middle and bottom parts are enlarged drawings of the left

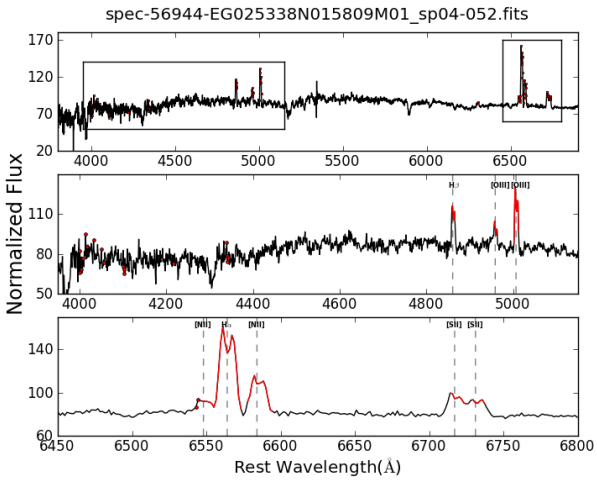


FIGURE 3. Characteristics distribution of a spectrum.

and right rectangular frames in the top respectively. In the last two parts, the red points are non-continue extracted attributes, and the red lines are regarded as extracted characteristics which are composed of continue extracted points.

2) THE CHARACTERISTICS IDENTIFICATION

After surveying the relative literature, these characteristics subspace is identified by astronomers. The results indicate that there are 8 emission lines in the subspace, that is Hα, Hβ, [OIII]$\lambda\lambda$4959, 5007, [NII]$\lambda\lambda$6548, 6584, [SII]$\lambda\lambda$6717, 6731.

3) CHARACTERISTICS REPRESENTATION

In the characteristics subspace, to further describe characteristics on 8 spectra lines, the wavelength coverage, red/blue shift interval(RBS) and line strength ratio(LSR) of double-peaked profiles are introduced theoretically. Suppose a space $A =< A_1, A_2, \ldots, A_i, \ldots, A_8 >$, where $A_i =< line, RBS, LSR >$ is a triple including the name of ith line, RBS and LSR of line. Definition of RBS and LSR as follows:

$$RBS = [min(RBS_j), max(RBS_j)] \tag{24}$$

$$LSR = [min(LSR_j), max(LSR_j)] \tag{25}$$

In (24) and (25), j $= 1, 2, \ldots,$ length(training dataset), the RBS_j and LSR_j are red/blue shift interval and line strength ratio of the jth spectra respectively. The values of RBS and LSR are between maximum and minimum value. The coordinate position of double peaks are $(wave_1, flux_1), (wave_2, flux_2)$, then RBS_j and LSR_j are defined:

$$RBS_j = (\frac{wave_2}{rest\_wave} - 1) - (\frac{wave_1}{rest\_wave}) \tag{26}$$

where RBS_j is the red/blue shift interval of the jth spectra, $rest\_wave$ is the line core in rest-frame wavelength. This function can measure the distance of double peaks, if a given distance of double peaks is outside the range, the

TABLE 1. Characteristics description of double peaks on blue band.

symbol	Hβ	[OIII]λ4959	[OIII]λ5007
$wav\_len(\mathring{A})$	[4859,4866]	[4956,4963]	[5002,5012]
RBS	[0.0002,0.0020]	[0.0004,0.0022]	[0.0004,0.0023]
LSR	[0.3024,3.1063]	[0.4292,2.9010]	[0.3213,3.6389]

TABLE 2. Characteristics description of double peaks on red band.

symbol	[NII]λ6548	Hα	[NII]λ6584	[SII]λ6717	[SII]λ6731
$wav\_len(\mathring{A})$	[6542,6559]	[6555,6573]	[6577,6594]	[6717,6726]	[6721,6737]
RBS	[0.0004,0.0025]	[0.0004,0.0022]	[0.0004,0.0018]	[0.0002,0.0019]	[0.0004,0.0021]
LSR	[0.3077,5.9030]	[0.2249,2.0969]	[0.3940,2.2307]	[0.3618,2.0050]	[0.3943,2.8830]

characteristic is unbelievable.

$$LSR_j = \frac{flux_2}{flux_1} \tag{27}$$

where LSR_j is the line strength ratio of the jth spectra. The LSR is also a measurement of double peaks. It is suspect when a given LSR value is outside the range of defined LSR.

The characteristics of double peaks in training set are recorded by observations of double-peaked profiles manually. The tables about characteristics of red and blue lines is generated, including wavelength range, RBS and LSR, computed according to (26) and (27). The directed descriptions of double-peaked characteristics are seen in table 1 and 2, which show characteristics descriptions of double peaks on the blue and red bands respectively. A given detected double-peaked profile can be proved further if the characteristic covers range in the tables. Otherwise, the line is regarded as non double-peaked profile.

C. CHARACTERISTICS GROUPING

The characteristics based on relevant subspace includes 8 lines, which are (Hα, Hβ, [OIII]$\lambda\lambda$4959, 5007, [NII]$\lambda\lambda$6548, 6584, [SII]$\lambda\lambda$6717, 6731). The lines in characteristics subspace should be divided into several subsets according frequent patterns and rough set theory. 1000 spectra with double-peaked profiles are selected as a dataset to compute the probability of every line. A new dataset $D = x_1, x_2, \ldots, x_i, \ldots, x_{1000}$ is generated, where $x_i = x_{ij}$, j = 1,2...,8, and x_{ij} is 1 or 0. If x_{ij} value is 1, the double peaks is occurred on the jth characteristic line. Otherwise, the characteristic without double peaks is recorded.

The dataset shaped 1000*8 is applied in mining association rules, finding the association between double-peaked profile lines. The Apriori algorithm is used to produce the strong association rules with higher confidence. The dataset with double peaks on line(value is 1) is to find strong associations by Apriori. Suppose the found association is $A \Longrightarrow B$, A and B are element or set, then $A \cup B$ will be a subset of characteristics subspace.

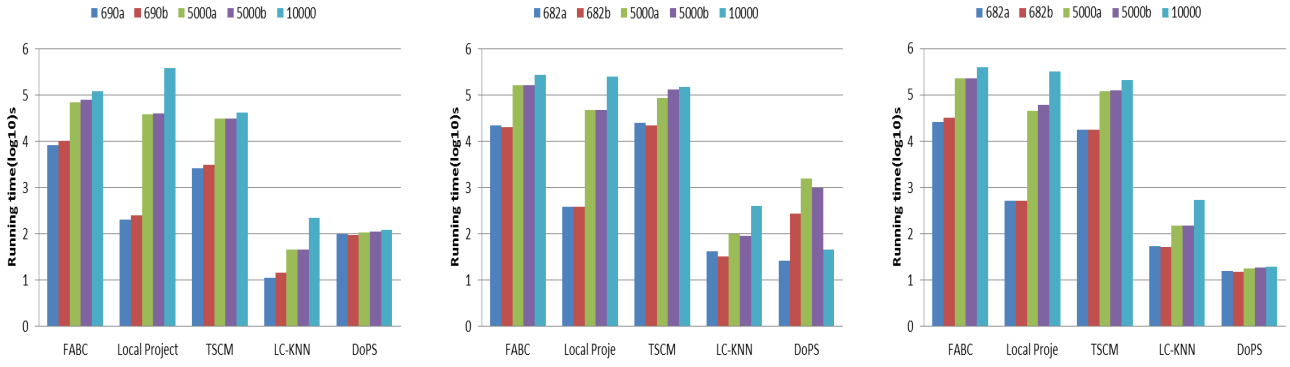
The number of association rules are produced by Apriori algorithm. The intersection and union of subsets of top 5 strong association rules can be as the lower approximate and upper approximate of rough set respectively. It is meaning that the first 5 subsets are $\{G_1, G_2, G_3, G_4, G_5\}$, the upper approximate is $U = \{G_1 \cup G_2 \cup G_3 \cup G_4 \cup G_5\}$, and the lower approximate is $L = \{G_1 \cap G_2 \cap G_3 \cap G_4 \cap G_5\}$. After adjusting the parameters of Apriori, the upper approximate is a set including all lines of characteristics subspace. The lower approximate is $\{[NII]\lambda\lambda6548, 6584, H\alpha, [SII]\lambda\lambda6717, 6731\}$ when confidence is 0.2. So the characteristics subspace can be divided into 3 subsets:

a. $\{H\beta, [OIII]\lambda\lambda4959, 5007\}$
b. $\{[NII]\lambda\lambda6548, 6584, H\alpha, [SII]\lambda\lambda6717, 6731\}$
c. $\{H\beta, [OIII]\lambda\lambda4959, 5007, [NII]\lambda\lambda6548, 6584, H\alpha, [SII]\lambda\lambda6717, 6731\}$

The subset a and b are characteristics lines of black and red bands respectively, which compose the subset c. Subset a is more frequent than subset b because the red bands may be non-existent with a bigger red-shift. The grouping of characteristics subspace based on association rules and rough set is also justified by astrophysical background for red-shift.

D. DOPS METHOD

In the method of DoPS, the lines in characteristics subspace are divided into 3 subsets to analyze double-peaked profiles spectra. The training process and testing process are included in SVM, which will be confirmed spectra data. The 345 spectra with double-peaked profiles are found in previous works by researchers, where the 8 lines are all included in 341 spectra due to smaller red-shift. The less positive samples are used for training classifier based on the known discovered data, which reflect rare scenes. For the subset a, 690 spectra are selected as the training set including 345 positive and 345 negative samples. 682 samples(341 positive and 341 negative samples) are used to be training dataset in subset b and c. In the testing process, 10000 samples from LAMOST DR5 are selected randomly as the testing dataset in all subsets. Then four algorithms including FABC [43],



(a) Subset a (b) Subset b (c) Subset c

FIGURE 4. Running time of algorithms in three subsets.

Local Projection [44], TSCM [45] and LC-KNN [46] are implemented to testify the available of DoPS method.

In the measurement space, accuracy, recall rate, reduced rate and running time are all compared between five algorithms. Higher accuracy, reduced rate and lower running time are our main object with higher recall rate, which should be ensured 1 due to the rarity of physical machine of double-peaked profiles spectra. The spectra of LAMOST are low-quality data, which is presented by the inverse variance. To test the identification influence of the signal and noise, experiment results are discussed with different inverse variance computed by inverse variance of every point in data. In addition, 5 testing datasets is also compared with different size of dataset in DoPS method.

1) EFFICIENCY VERIFICATION

As a measurement of efficiency, the running time is recorded to test the efficiency of our proposed method. The running time of five methods in 3 subsets is shown in Figure 4, where the 5 testing datasets are used to compare the running time. For the subset a, the first 4 dataset sets are randomly selected including 345 positive data. The sample in the last dataset sized 10000 is randomly selected, including rare positive ones. There are 341 positive points in the first dataset for subset b and c. In Figure 4, for the subset a, the running time of DoPS is smaller than the first three methods, while is bigger than LC-KNN generally due to the lower-dimensional dataset in subset a. It is seen that the running time of DoPS in subset c is far smaller than others with higher-dimensional dataset. In addition, the running times of the first 4 methods are greater with bigger size of dataset. The running time is stable no matter how the size of dataset changes in DoPS method in the subset a and c. However the running time is unstable in subset b when the size of dataset is unequal. The more running time is costed in the dataset sized 5000 than 10000, which don't convergence quickly in the training processing.

The trend of running time of DoPS in subset b is not similar with other two subsets. The quality of the testing dataset is an important influence factor for running time in subset b. The testing dataset in subset b is composed of randomly selected samples, which may be an aspect of the trend of running time in subset b.

Overall, the Figure 4 shows better performance of DoPS than other methods, which tests the efficiency of our proposed method from perspective of time.

2) INFLUENCE OF DATA QUALITY

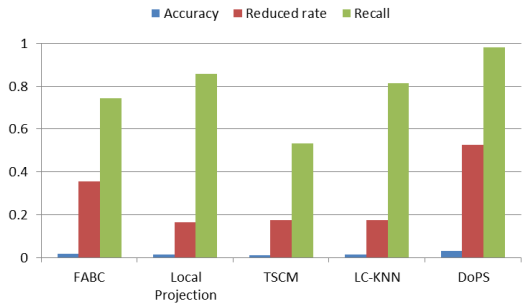
The quality of dataset can be measures by the signal and noise(SNR) which obtained via inverse variance. The higher SNR of spectra, the better results of classification. The dataset sized 10000 randomly selected in three subsets is used to test the influence of SNRs. For subset a, the testing dataset SNR>0.06 and 0.1 is considered into testing processing. Figure5 shows the accuracy, reduced rate and recall rate of five algorithms with SNR>0.6 and 0.1.

There are 23 dimensions in subset a, which is composed of 6-dimensional Hβ set and 17-dimensional [OIII] set. It is can be seen that the recall rate of DoPS is smaller with bigger SNR. The more data will be removed from positive data when SNR is bigger and bigger, which leads to the fewer and fewer recall rate. The reduced rate of DoPS with SNR>0.1 is higher than 0.06, which is also concluded that higher SNR can improve the method performance in subset a.

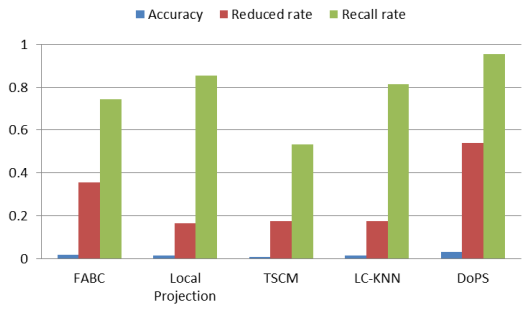
53 attributes are included in subset b, which is selected according to SNR>0.5 and 1. So the subset c including 76 attributes, composed of subset a and b. The results of subset b and c are shown in Figure6 and 7 respectively.

In Figure6, the accuracy and reduced rate of DoPS are higher than others. And the recall rate with SNR>0.5 is bigger than SNR>1 due to the more positive samples loss with bigger SNR. So the threshold of SNR should be considered to ensure the temperate between accuracy and reduced rate.

The highest dimensional subset c is also to show the influence of SNR on classification result. 0.5 and 1 are also selected to test the accuracy, reduced rate and recall rate like subset a and b, which is shown in Figure 7. For subset c which includes 76 attributes, the Figure 7 shows the trends

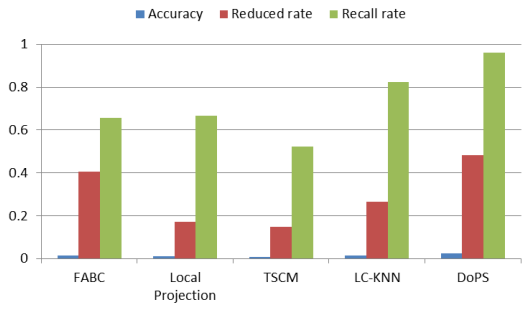


(a) SNR>0.06

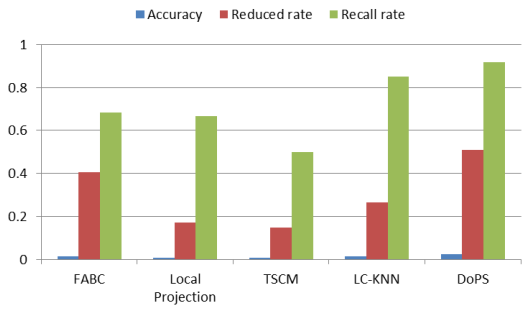


(b) SNR>0.1

FIGURE 5. Evaluations of 5 algorithms with different SNRs in subset a.



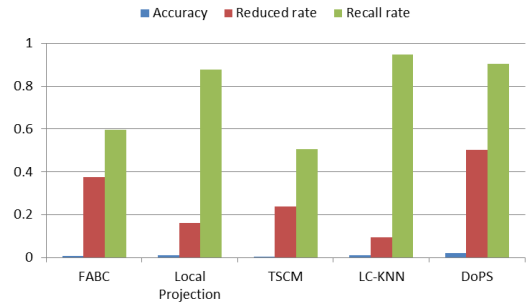
(a) SNR>0.5



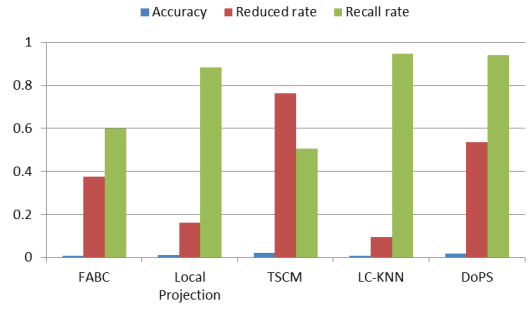
(b) SNR>1

FIGURE 6. Evaluations of 5 algorithms with different SNRs in subset b.

similar to the first two subsets. The higher reduced rate is appeared with bigger SNR, while the lower recall rate is existed.



(a) SNR>0.5



(b) SNR>1

FIGURE 7. Evaluations of 5 algorithms with different SNRs in subset c.

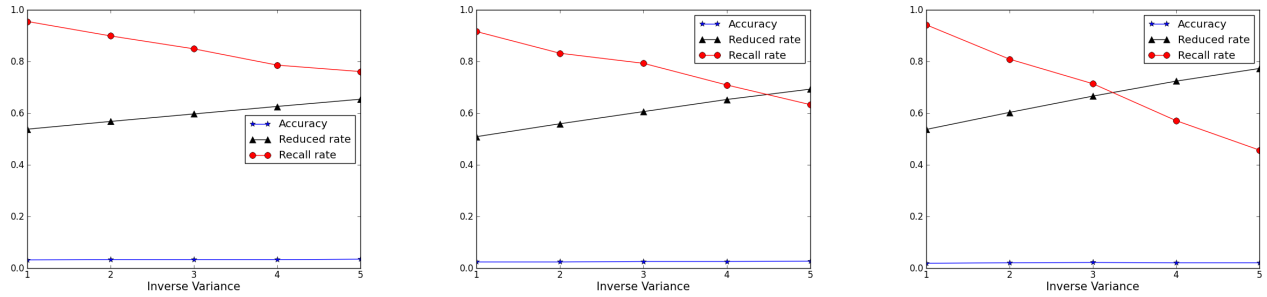
To test the accuracy, reduced rate and recall rate with the change of SNR, the trends of evaluations with different SNRs are shown in Figure 8.

It is can be seen that the bigger SNR, smaller recall rate, while greater reduced rate are always happened. The more and more positive samples are not considered into positive class when SNR is bigger and bigger. So it is a trade-off in evaluations and selection of SNR.

3) INFLUENCE OF DATASET SIZE

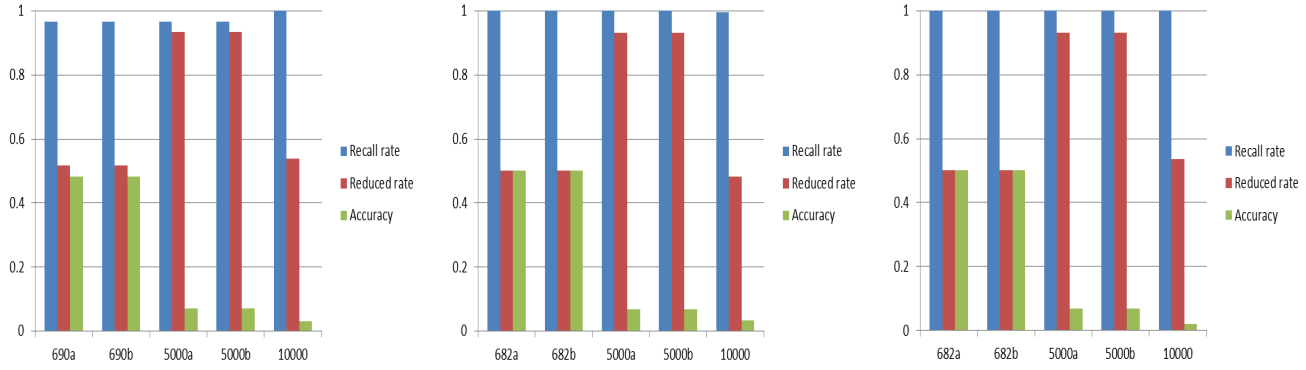
Under the high-dimensional data background, the three subsets represent different dimensions. To test the effectiveness of size of dataset, 5 datasets sized different type sizes are used to be testing datasets. Accuracy, reduced rate and recall rate are all as the evaluation measurements of the algorithms. We hope the recall rate is 1 due to the little proportion of double-peaked profiles spectra. So it is our goal that higher accuracy and reduced rate with recall rate 1. The result of accuracy, reduced rate and recall rate can be seen in Figure 9, in which the accuracy, reduced rate and recall rate are included in different size of dataset.

In Figure9, 690a and 690b in the first sub-figure denote different datasets including 345 positive and 345 negative. 5000a and 5000b are testing data sized 5000 including 345 positive. The last size of dataset is 10000 randomly selected from LAMOST. In the last two sub-figures, 682a and 682b are dataset sized 682 including equal number of positive and negative sample. For subset a, the recall rate is 1 when the size of dataset is 10000. However, the reduced rate is higher than



(a) subset a (b) subset b (c) subset c

FIGURE 8. Evaluations of DoPS algorithm with different SNRs in three subsets.



(a) subset a (b) subset b (c) subset c

FIGURE 9. Evaluations of DoPS algorithm in three subsets with size of dataset.

others in dataset sized 5000. For the subset b and c, the recall rate is always 1 with higher reduced rate when the size of testing data is 5000.

The better performance occurred in Figure9 due to the bigger proportion in dataset sized 5000 than 10000. However, when the size of dataset is reached 10000, the reduced rate is 1, which may be a better phenomenon for bigger dataset. In our method, the number of iteration is needed to set manually, which has influence on classification effects. The novel proposed on-line algorithm based on worse-violators can be considered for improving our method effects, which eliminates the need for specifying the maximum number of iterations and the use of a regularization term. This approach is suitable for large-scale learning, which is also in accordance with our data quest.

VI. CONCLUSION

In this paper, we propose a method of DoPS method based on RS & SVM for searching double-peaked profiles spectra. Our detailed works are as follows three parts. Firstly, the characteristics subspace($H\beta$, OIIIλ4959, OIIIλ5007, NIIλ6548, $H\alpha$, NIIλ6584, SIIλ6717, SIIλ6731) is obtained by characteristics selection. Then the characteristics descriptions of characteristics subspace are given, including RBS, LSR and wavelength range. Secondly, the characteristics subspace is divided into 3 subsets by frequent patterns and rough set theory. Finally, several datasets selected from LAMOST are used to employ in DoPS and other four compared algorithms. It is confirmed that our DoPS method shows better performance in efficiency, noise immunity, recall rate and reduced rate.

In the DoPS method, an unbalanced between the accuracy and recall rate is a shortcoming due to the rarity of double-peaked profiles data. However, it can be changed by adjusting parameters according to different requests. This method will be employed in searching special data in LAMOST DR7 in next works, which includes rare positive sample. So the incremental method can be applied, that is the searched positive data can be included into the next positive sample to increase the size of positive sample.

ACKNOWLEDGMENT

The Guo Shou Jing Telescope (the Large Sky Area Multi-Object Fiber Spectroscopic Telescope, LAMOST) is a National Major Scientific Project built by the Chinese Academy of Sciences. Funding for the project has been provided by the National Development and Reform Commission. LAMOST is operated and managed by National Astronomical Observatories, Chinese Academy of Sciences.

REFERENCES

[1] A. Addeh, A. Khormali, and N. A. Golilarz, "Control chart pattern recognition using RBF neural network with new training algorithm and practical features," *ISA Trans.*, vol. 79, pp. 202–216, Aug. 2018.

[2] Y. Djenouri, A. Belhadi, J. C. Lin, D. Djenouri, and A. Cano, "A survey on urban traffic anomalies detection algorithms," *IEEE Access*, vol. 1, pp. 12192–12205, 2019,

[3] P. Govindarajan *et al.*, "Classification of stroke disease using machine learning algorithms," *Neural Comput. Appl.*, pp. 1–12, 2019.

[4] M. Abdel-Basset, G. Manogaran, D. El-Shahat, and S. Mirjalili, "A hybrid whale optimization algorithm based on local search strategy for the permutation flow shop scheduling problem," *Future Gener. Comput. Syst.*, vol. 85, pp. 129–145, Aug. 2018.

[5] Z. Ma, H. Ge, Y. Liu, M. Zhao, and J. Ma, "A combination method for Android malware detection based on control flow graphs and machine learning algorithms," *IEEE Access*, vol. 7, pp. 21235–21245, 2019.

[6] C. T. Cheng *et al.*, "Application of a deep learning algorithm for detection and visualization of hip fractures on plain pelvic radiographs," *Eur. Radiol.*, pp. 1–9, 2019.

[7] A.-L. Luo, Y. H. Zhao, G. Zhao, L. C. Deng, X. W. Liu, Y. P. Jing, G. Wang, H. T. Zhang, J. R. Shi, X. Q. Cui, and Y. Q. Chu, "The first data release (DR1) of the LAMOST regular survey," *Res. Astron. Astrophys.*, vol. 15, no. 8, p. 1095, 2015.

[8] J. Liu, M. Eracleous, and J. P. Halpern, "A radial velocity test for supermassive black hole binaries as an explanation forbroad, double-peaked emission lines in active galactic nuclei," *Astrophys. J.*, vol. 817, no. 1, p. 42, 2016.

[9] C.-X. Qu *et al.*, "Feature extraction and analysis of double-peaked emission line spectra based on relevant subspace," *Spectrosc. Spectral Anal.*, vol. 39, no. 2, p. 1677, 2019.

[10] Y. Lyu and X. Liu, "A high fraction of double-peaked narrow emission lines in powerful active galactic nuclei," *Monthly Notices Roy. Astronomical Soc.*, vol. 463, no. 1, pp. 24–36, 2016.

[11] J. M. Comerford, R. Nevin, A. Stemo, F. Müller-Sánchez, R. S. Barrows, M. C. Cooper, and J. A. Newman, "The origin of double-peaked narrow lines in active galactic nuclei. IV. Association with galaxy mergers," 2018, *arXiv:1810.11543*. [Online]. Available: https://arxiv.org/abs/1810.11543

[12] R. K, M. Das, P. Kharb, and M. Honey, "A candidate dual AGN in a double-peaked emission-line galaxy with precessing radio jets," *Monthly Notices Roy. Astronomical Soc.*, vol. 465, no. 2, pp. 4772–4782, 2017.

[13] H. Yang, "A study of superimposed components recognition and analysis on LAMOST extragalactic spectra," *Publications Astronomical Soc. Pacific*, vol. 129, no. 971, 2017, Art. no. 017001.

[14] A.-L. Sun, J. E. Greene, and N. L. Zakamska, "Sizes and kinematics of extended narrow-line regions in luminous obscured AGN selected by broadband images," *Astrophys. J.*, vol. 835, no. 2, p. 222, 2017.

[15] K. Rubinur, M. Das, and P. Kharb, "Searching for dual AGN in galaxies with double-peaked emission line spectra using radio observations," *Monthly Notices Roy. Astronomical Soc.*, vol. 484, no. 4, pp. 4933–4950, 2019.

[16] Z.-X. Shi, A.-L. Luo, G. Comte, X.-Y. Chen, P. Wei, Y.-H. Zhao, F.-C. Wu, Y.-X. Zhang, S.-Y. Shen, M. Yang, H. Wu, X.-B. Wu, H.-T. Zhang, Y.-J. Lei, J.-N. Zhang, T.-G. Wang, G. Jin, and Y. Zhang, "A search for double-peaked narrow emission line galaxies and AGNs in the LAMOST DR1," *Res. Astron. Astrophys.*, vol. 14, no. 10, p. 1234, 2014.

[17] M.-X. Wang, A.-L. Luo, Y.-H. Song, S.-Y. Shen, S. Feng, L.-L. Wang, Y.-F. Wang, Y.-B. Li, B. Du, W. Hou, Y.-X. Guo, X. Kong, and J.-N. Zhang, "Double-peaked narrow emission-line galaxies in LAMOST survey," *Monthly Notices Roy. Astronomical Soc.*, vol. 482, no. 2, pp. 1889–1899, 2018.

[18] P. Škoda, A. Palicka, J. Koza, and K. Shakurova, "Identification of interesting objects in large spectral surveys using highly parallelized machine learning," *Proc. Int. Astronomical Union*, vol. 12, no. S325, pp. 180–185, 2016.

[19] K. L. Smith, G. A. Shields, E. W. Bonning, C. C. McMullen, D. J. Rosario, and S. Salviander, "A search for binary active galactic nuclei: Double-peaked [O III] AGNs in the sloan digital sky survey," *Astrophys. J.*, vol. 716, no. 1, pp. 866–877, 2010.

[20] I. V. Strateva, M. A. Strauss, L. Hao, D. J. Schlegel, P. B. Hall, J. E. Gunn, L.-X. Li, Ž. Ivezic, G. T. Richards, N. L. Zakamska, W. Voges, S. F. Anderson, R. H. Lupton, D. P. Schneider, J. Brinkmann, and R. C. Nichol, "Double-peaked low-ionization emission lines in active galactic nuclei," *Astronomical J.*, vol. 126, no. 4, p. 1720, 2003.

[21] B. García-Lorenzo, "Searching double-peaked emission-line profiles in the spectra of galaxies through the symmetry of the cross-correlation function," *Monthly Notices Roy. Astronomical Soc.*, vol. 429, no. 4, pp. 2903–2909, 2013.

[22] K. M. Cherry and L. Qian, "Scaling up molecular pattern recognition with DNA-based winner-take-all neural networks," *Nature*, vol. 559, no. 7714, pp. 370–376, 2018.

[23] X. Bai, J. Zhou, and A. Robles-Kelly, "Pattern recognition for high performance imaging," *Pattern Recognit.*, vol. 82, pp. 38–39, Oct. 2018.

[24] Q. Zhou, S. Kaappa, S. Malola, H. Lu, D. Guan, Y. Li, H. Wang, Z. Xie, Z. Ma, H. Häkkinen, N. Zheng, X. Yang, and L. Zheng, "Real-space imaging with pattern recognition of a ligand-protected Ag374 nanocluster at sub-molecular resolution," *Nature Commun.*, vol. 9, no. 1, 2018, Art. no. 2948.

[25] R. Zhang, J. Tao, and H. Zhou, "Fuzzy optimal energy management for fuel cell and supercapacitor systems using neural network based driving pattern recognition," *IEEE Trans. Fuzzy Syst.*, vol. 27, no. 1, pp. 45–57, Jan. 2018.

[26] J. M. Vidal, M. A. S. Monge, and L. J. G. Villalba, "A novel pattern recognition system for detecting Android malware by analyzing suspicious boot sequences," *Knowl.-Based Syst.*, vol. 150, pp. 198–217, Jun. 2018.

[27] A. R. Barghi, A. Ferdowsi, and A. Abhari, "Musical preferences prediction by classification algorithm," in *Proc. Commun. Netw. Symp.*, 2018, Art. no. 2.

[28] L. Valenzuela and K. Pichara, "Unsupervised classification of variable stars," *Monthly Notices Roy. Astronomical Soc.*, vol. 474, no. 3, pp. 3259–3272, 2017.

[29] R. Garcia-Dias, C. A Prieto, J. S. Almeida, and I. Ordovás-Pascual, "Machine learning in APOGEE: Unsupervised spectral classification with K-means," *Astronomy Astrophys.*, vol. 612, Apr. 2018, Art. no. A98.

[30] F. Abid and L. Hamami, "A survey of neural network based automated systems for human chromosome classification," *Artif. Intell. Rev.*, vol. 49, no. 1, pp. 41–56, 2018.

[31] B. E. Boser, I. M. Guyon, V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proc. 5th Annu. Workshop Comput. Learn. Theory*, 1992, pp. 144–152.

[32] S. Amari and S. Wu, "Improving support vector machine classifiers by modifying kernel functions," *Neural Netw.*, vol. 12, no. 6, pp. 783–789, 1999.

[33] W. Chen, H. R. Pourghasemi, and S. A. Naghibi, "A comparative study of landslide susceptibility maps produced using support vector machine with different kernel functions and entropy data mining models in China," *Bull. Eng. Geol. Environ.*, vol. 77, no. 2, pp. 647–664, 2018.

[34] P. Kelly and L. Longo, "An investigation into the effects of multiple kernel combinations on solutions spaces in support vector machines," in *Proc. IFIP Int. Conf. Artif. Intell. Appl. Innov.* Cham, Switzerland: Springer, 2018, pp. 157–167.

[35] J. L. Rojo-Álvarez, M. Martínez-Ramón, J. Muñoz-Marí, and G. Camps-Valls, "Support vector machine and kernel classification algorithms," in *Digital Signal Processing With Kernel Methods* 2018, pp. 433–502.

[36] Z. Zhang and Z. Mi, "Preference risk assessment method based on grey multi index kernel support vector machine model," *Cluster Comput.*, vol. 3, pp. 1–7, Feb. 2018.

[37] Y. Li, Z. Zhu, A. Hou, Q. Zhao, L. Liu, and L. Zhang, "Pulmonary nodule recognition based on multiple kernel learning support vector machine-PSO," *Comput. Math. Methods Med.*, vol. 2018, Apr. 2018, Art. no. 1461470.

[38] P. T. Noi and M. Kappas, "Comparison of random forest, k-nearest neighbor, and support vector machine classifiers for land cover classification using Sentinel-2 imagery," *Sensors*, vol. 18, no. 1, p. 18, 2018.

[39] H. A. Illias and W. Z. Liang, "Identification of transformer fault based on dissolved gas analysis using hybrid support vector machine-modified evolutionary particle swarm optimisation," *PLoS ONE*, vol. 13, no. 1, 2018, Art. no. e0191366.

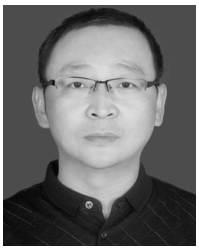
[40] A. Lawi and Y. Adhitya, "Classifying physical morphology of cocoa beans digital images using multiclass ensemble least-squares support vector machine," *J. Phys., Conf. Ser.*, vol. 979, no. 1, 2018, Art. no. 012029.

[41] Y. Chen, H. Li, X. Jing, L. Hou, and X. Bu, "Intelligent chatter detection using image features and support vector machine," *Int. J. Adv. Manuf. Technol.*, vol. 102, nos. 5–8, pp. 1433–1442, 2019.

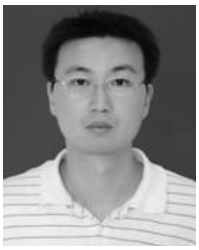
[42] X. Cao, R. Ma, L. Liu, H. Shi, Y. Cheng, and C. Sun, "A machine learning-based algorithm for joint scheduling and power control in wireless networks," *IEEE Internet Things J.*, vol. 5, no. 6, pp. 4308–4318, 2018.

[43] S. Ezghari, A. Zahi, and K. Zenkouar, "A new nearest neighbor classification method based on fuzzy set theory and aggregation operators," *Expert Syst. Appl.*, vol. 80, pp. 58–74, Sep. 2017.
[44] T. Ortner, P. Filzmoser, M. Zaharieva, S. Brodinova, and C. Breiteneder, "Local projections for high-dimensional outlier detection," 2017, *arXiv:1708.01550*. [Online]. Available: https://arxiv.org/abs/1708.01550
[45] Y. Zelenkov, E. Fedorova, and D. Chekrizo, "Two-step classification method based on genetic algorithm for bankruptcy forecasting," *Expert Syst. Appl.*, vol. 88, pp. 393–401, Dec. 2017.
[46] Z. Deng, X. Zhu, D. Cheng, M. Zong, and S. Zhang, "Efficient kNN classification algorithm for big data," *Neurocomputing*, vol. 195, pp. 143–148, Jun. 2016.
[47] J. Zhang, X. Yu, Y. Li, S. Zhang, Y. Xun, and X. Qin, "A relevant subspace based contextual outlier mining algorithm," *Knowl.-Based Syst.*, vol. 99, pp. 1–9, May 2016.

CAIXIA QU is currently pursuing the master's degree with the Taiyuan University of Science and Technology (TYUST), Taiyuan, China. Her current research interests include data mining and paralleling computing.

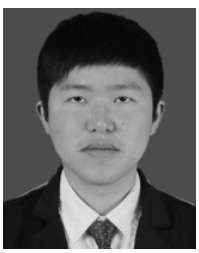


HAIFENG YANG is currently a Professor of computer application technology with the Taiyuan University of Science and Technology, Taiyuan, China. He is also the long-term Member of the Institute for Intelligent Information and Data Mining. He is a member of the China Computer Federation (CCF) and the Chinese Astronomical Society (CAS). His research concerns the data mining and machine learning methods in the specific backgrounds especially for the astronomical big data.



JIANGHUI CAI is currently the Chief Professor of computer application technology with the Taiyuan University of Science and Technology, Taiyuan, China. He is also the long-term member of the Institute for Intelligent Information and Data Mining. He is a Senior Member of the China Computer Federation (CCF). His research concerns the data mining and machine learning methods in specific backgrounds of astronomical informatics, seismology, and mechanical engineering.

JIFU ZHANG received the B.S. and M.S. degrees in computer science and technology from the Hefei University of Technology, China, in 1983 and 1989, respectively, and the Ph.D. degree in pattern recognition and intelligence systems from the Beijing Institute of Technology, in 2005. He is currently a Professor with the School of Computer Science and Technology, Taiyuan University of Science and Technology (TYUST). His research interests include data mining, parallel computing, and celestial spectrum data analysis.



YONGXIANG ZHOU is currently pursuing the master's degree with the Taiyuan University of Science and Technology (TYUST), Taiyuan, China. His current research interests include data mining and paralleling computing.