Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

WEDA: A Weak Emission-line Detection Algorithm based on the Weighted Ranking

YONGXIANG ZHOU[1], HAIFENG YANG[1,2*], JIANGHUI CAI[1,2*], XUJUN ZHAO[1], YALING XUN[1], AND CAIXIA QU[1]

[1]School of Computer Science and Technology, Taiyuan University of Science and Technology, Taiyuan 030024, China
[2]Shanxi Key Laboratory of Advanced Control and Equipment intelligence, Taiyuan 030024, China

*Corresponding authors: Haifeng Yang (e-mail: hfyang@tyust.edu.cn), Jianghui Cai (e-mail: jianghui@tyust.edu.cn)

The work is supported by the National Natural Science Foundation of China (Grant Nos. U1931209, U1731126) and Shanxi Province Key Research and Development Program (Grant Nos. 201803D121059, 201903D121116).

ABSTRACT The Hα emission line in rest wavelength frame of optical spectra is valuable characteristics for nebulae detection. Searching and recognizing the spectra with Hα emission line from massive data are necessary for the further study, while the most of methods existed currently do not adapt to such spectral data, especially for the spectra with weak Hα emission line. To address this issue, a new algorithm (named WEDA) for detection of spectra with Hα emission line is provided in this paper. Firstly, the difference factor μ between the line characteristics of the specific data is defined as its weight in recognizing of the whole lines table. Secondly, a tuning function $f(\tau, \delta)$ based on the momentum formula is defined to update the weights during the process. In this step, the spectra with Hα emission line are analysed and classified as 3 different situations. The amount of spectra with Hα emission line is different in 3 different situations, so the speed of weight of update is different in 3 different situations. The weight of update helps us detect the data containing weak Hα emission line in the 3 situations. Based on this, a new integrated algorithm especially for the detection of the spectra with Hα is provided. In the end, by using several spectral datasets from the DR5 of LAMOST survey, experiments results indicate that the WEDA shows higher accuracy basically unaffected by the dataset size and the signal to noise ratio(SNR) than the other similar algorithms.

INDEX TERMS Ranking Algorithm, Hα Emission Line, Binary Classification, Imbalance Classification.

I. INTRODUCTION

WITH the development of technology, increasing amounts of spectral data have been obtained by astronomical telescopes. The challenge we face today is how to find the spectral data we need from massive amounts of spectral data. There is also a lot of work to analyze the spectral data [1], even facilities [2] Due to the large amount of work, a lot of the work cannot be finished manually, and some of the work can be considered to be classification tasks in machine learning. Therefore, the work seen as classification tasks can be finsihed automatically via computers; for example, the detection of the Hα emission line can be seen as a binary classification task that attempts to classify spectral data as {1, -1}, where 1 represents that this spectral data contains the Hα emission line, and -1 represents that this spectral data does not contain the Hα emission line. However, the results of many classification methods do not meet the requirements of many complex data situations. Furthermore, data that meet our needs are often rare, which makes model training harder. The ranking algorithm proposed is used to solve these problems. To adapt to a complex data environment, we will weight the data to distinguish it. The ranking algorithm and weight algorithm are combined to highlight the data we need. In this paper, WEDA is used to find data with the Hα emission line on the LAMOST survey dataset.

A. MOTIVATIONS

The motivations of this paper can be summarized as follows:

1. The Hα emission line in the rest wavelength frame data are very valuable materials for the further study of nebulas in our Galaxy. Therefore, to search out and recognize them from the massive spectral data would be of great significance for the astronomers.

2. The Hα emission line always show some complex characteristics such as weak features, noise interference, various profiles, etc., which largely increased difficulty in searching and identifying. Design a specific method for this problem is necessary.

3. The ranking algorithm is alway a useful method in such areas, however, it is not suitable for solve the above problems directly.

Motivation 1 The Hα emission line can be used to detect nebulas, which is vital to work such as studying nebulas [3], star-forming galaxies [4] and late L dwarfs and T dwarfs [5]. The first step for these studies is to detect spectral data with the Hα emission line. However, data with the Hα emission line in many studies is obtained by manual identification [6], which has high precision but also high cost. When massive amounts of data need to be processed in research, such a large amount of work cannot be finished by manual identification. A new algorithm needs to be proposed to finish this work automatically. In this paper, the ranking algorithm and weight algorithm are combined to find spectral data with the Hα emission line.

Motivation 2 In recent years, an increasing number of data mining algorithms have been improved and have begun to be applied to astronomy research [7], for example, classification of star spectral data [8], finding young stellar populations [12], analyzing the variation characteristics of the sky background [9], star-galaxy classification [13] and detection of faint γ-ray sources [14]. These data mining algorithms have obtained good results in many astronomical studies. Hence, data mining algorithms are considered for use in detecting Hα emission line. Although some work, such as detecting emission lines, can be seen as classification problems, many traditional classification models, such as SVM [15] and ANN [18], cannot be directly used for this work due to complex characteristics. These algorithms need to be optimized before being applied to astronomical research, such as SVM [16] and RS [17]. Deep neural networks are powerful, but training requires a considerable amount of data instead of small-scale data sets [19]. This should be considered when the corresponding algorithm is proposed. Some clustering algorithms [10] have also been applied to spectral data [11].

Motivation 3 The ranking algorithm has been applied to some fields such as search engines [20] and recommendation systems [21]. In data mining areas such as information retrieval, there are many classic and excellent algorithms, for example, pageRank [22] and HITS [23]. Simple ranking algorithms cannot obtain all data with the Hα emission line. If a ranking function with good quality is learned [24], the task is likely to be finished. However, if all data are directly used in the ranking algorithm without data preprocessing, the ranking algorithm will have a large time complexity. Before application to big data sets, the ranking algorithm needs to be improved.

B. CONTRIBUTIONS

A novel integrated algorithm, called WEDA, is proposed in this paper based on the above motivations. The main idea of WEDA is that the ranking algorithm and changeable weights are combined to find data with the Hα emission line. First, the weights in WEDA are initialized according to differences between the specific data. Afterward, the weight update function is proposed based on a momentum function [25]. The weights will update according to changes in the data. The data are sorted by continuously selecting data during the weight update process, which forms an ordered sequence. When an ordered sequence is obtained, all data are classified only by a cutoff threshold. This algorithm can be used to find spectral data with the Hα emission line.

The contributions of this paper are as follows:

1. The difference factor μ between the line characteristics of the specific data is defined to initialize weights of primary information and secondary information.

2. The tuning function $f(\tau, \delta)$ based on momentum function is designed as the basis for weights update.

3. A new integrated algorithm named WEDA especially for the detection of the spectra with Hα is proposed based on the difference factor μ and the tuning function $f(\tau, \delta)$. Meanwhile, this algorithm is evaluated by using the spectra from the DR5 of LAMOST survey.

C. ROADMAP

The rest of the paper is organized as follows. Section II introduces work related to WEDA. Section III first gives the main idea of WEDA. Then, we give an in-depth introduction to WEDA, and a theoretical analysis of WEDA is performed. Then, Section IV discusses all experimental results and the quality of the algorithm itself to compared algorithms.

II. RELATED WORK

Traditional detection methods for nebulae depend on the observation of infrared telescopes and radio telescopes. However, the detection of lower-density nebula is not ideal due to factors such as resolution. The Hα emission line in a rest wavelength frame can be prevented from being absorbed by the star population, and it can be discovered and measured. The SNR of LAMOST spectral data is higher on the r-band, and therefore, the detection of the Hα emission line has a low dependence on the overall quality of the spectral data. The Hα emission line in a rest wavelength frame is used to detect nebulae. With the development of observational means, increasing amounts of spectral data are obtained by the LAMOST telescope. A considerable amount of meaningful information can be mined from the massive amounts of data. Spectral data with Hα emission line in LAMOST [26] can be chosen for use in WEDA. Previous work has found data with Hα emission line and used these spectral data, for example, to study peculiar A-type stars [27] and to search for classical Be stars [28]. Among spectral data with Hα emission line, some data only have a weak Hα emission line, which must be judged by experts. Many algorithms cannot

find data with weak Hα emission line. From an astronomical background, we can make use of other related information to judge the existence of the Hα emission line. Although the Hα emission line is disturbed by noise, some other emission lines chosen from the frame can verify its existence, such as Hβ, NII, and OII. If data show high confidence in these emission lines, it is likely that these spectral data contain Hα emission line. Furthermore, neighbor data can also be used to demonstrate the existence of the Hα emission line in general, but there are some outliers in which the neighbor data do not contain the Hα emission line, while the data do contain the Hα emission line. Compared with neighbor data, the confidence of other emission lines chosen from the frame is more persuasive.

The detection of the Hα emission line can be transformed into a binary classification problem in imbalanced data. There are many applications that can be transformed into this kind of problem: in the diagnosis of disease, the detection of breast cancer [29] [30] and the choice of cardiac care [31] can be seen as binary classification problems.Among financial problems, credit-card fraud detection [32] and bankruptcy forecasting [33] can also be solved in this way. In information security, the same is true of intrusion detection [34] and spam detection [35]. The most serious problem of this type of binary classification is that the imbalance between the numbers of the two classifications will lead to a difference in the sensitivities of predictions [36]. The sensitivities of predictions will be more inclined to the majority class target [37], and the minority target class is often ignored. However, we are often interested in the minority target class, which also prompts us to not just look at the overall precision of the algorithm. For imbalanced classification, there are four aspects of interest: the training set size, class prior, cost matrix and placement of the decision boundary [38]. The training set size idea is to alter the size of training sets, which increases the number of minority target classes or decreases the number of majority target classes [39]. Under-sampling or oversampling makes the number of the two classifications the same, which transforms the imbalance into balance. This method is simple to implement and can solve the imbalance problem [40] [41]. Representative algorithms that use this strategy are SMOTE [42] and easyEnsemble [41]. The cost matrix strategy [43] increases the weight of misclassified data so that the model can fit these misclassified data; examples include adaboost [44]. The theory of class priors is based on the assumption that the distribution of the positive class is known [45]. The placement of the decision boundary moves the classified threshold instead of increasing and decreasing the amount of data [46]. The above four methods are usually used to solve imbalanced classification. This data skewness problem also has adverse impacts on parallel operations [47]. For parallel operations on imbalanced data, algorithms also need to be designed according to the specific situation.

Ranking algorithms are often used in information retrieval and recommendation systems [48] to mine the data we need from massive amounts of data. The machine learning field has considered bipartite ranking algorithms [49]. These algorithms can assign a score to each data point, and all data are sorted by this score. The scores of positive instances are higher than the scores of negative instances [50], which helps us to classify all the data. To achieve this goal, the bipartite ranking algorithm needs to determine a function in which the rank of positive instances is higher than the rank of negative instances [51]. Currently, there are many studies on how this bipartite ranking algorithm can be applied to data mining, such as CBR [52] and the Bayesian multiple kernel bipartite ranking model [53]. Therefore, the bipartite ranking algorithm can be used to find data with the Hα emission line among massive amounts of data, and the key aspect of this algorithm is how to design the score function.

To date, there are many studies about binary classification on imbalanced data and bipartite ranking algorithms. However, There is no useful approach to detect data with weak Hα emission line from spectral data sets. In this paper, a newly designed bipartite ranking algorithm is used to detect data with weak Hα emission line.

III. SEARCH METHOD

In this section, the method of WEDA will be introduced in detail. An overview of the whole algorithm is shown in Section A, and in Section B, we give more in-depth information on the contents of Section A. The theoretical analysis is displayed in Section C.

A. THE MAIN IDEA

The initial spectral data are denoted by A, and they need to be preprocessed to extract meaningful information. In this paper, the entire dataset is divided into three parts based on the SNR, the ranges of which are 0-10, 10-50 and above 50, because different SNRs will lead to different amounts of data with Hα emission line and different qualities of the performance of WEDA. The three parts of the data are processed separately and are denoted by $\{A_1, A_2, A_3\}$. Then, we preprocess each part of the data in A. Preprocessing can help to exclude some of the data that clearly lack the Hα emission line. The data that cannot be classified by data preprocessing are denoted by $D_i \in A_i$, and for these data it is necessary to utilize other useful information around the Hα emission line. With this information, we can obtain $D_{in} = \{(a_n, b_n^1, b_n^2, \ldots, b_n^j, y_n)\}_{n=1}^N$, where a_n is the sum of the confidence of the Hα emission line and the Hβ emission line; there are two b_n^j in this paper: one is the sum of the confidence of two NII emission lines and two SII emission lines, and one is the sum of the confidence of six emission lines in data formed by superposition of neighbor spectral data. y_n is the data classification, that is, whether the data contain the Hα emission line, where $y_n \in \{-1, 1\}$. N is the number of data points that are not determined by data preprocessing. Before WEDA is processed, we combine all the secondary information by predefined weights $\{\varphi_1, \varphi_2, \ldots, \varphi_j\}$. The

formula is as follows.

$$b_n = b_n^1 \times \varphi_1 + b_n^2 \times \varphi_2 + \cdots + b_n^j \times \varphi_j \quad (1)$$

WEDA can utilize the information $\{(a_n, b_n, y_n)\}_{n=1}^N$ to obtain the confidence $\{C_n\}_{n=1}^N$ of all data by $\{\omega_1, \omega_2\}$. The purpose of considering the confidence C_n of the data is to assess the current possibility of the Hα emission line. The currently highest confidence C_n of data χ_n^* is put into an ordered sequence of probabilities χ and removed from D_i. The most likely sample χ_n^* is selected from the undetermined data D_i by the formula below.

$$\begin{aligned} \chi_n^* &= argmax\{C_i\}_{i=1}^N \\ &= argmax\{a_n \times \omega_1 + b_n \times \omega_2\}_{n=1}^N \end{aligned} \quad (2)$$

Finally, the confidence measures the probability of each data point having the Hα emission line. An ordered sequence of probabilities $\chi = \{\chi_1^*, \chi_2^*, \ldots, \chi_n^*\}$ is obtained, which is in descending order.

B. THE WEIGTH UPDATE ALGORITHM(WEDA)

In section A, we finish the overview of the whole algorithm. The entire algorithm process will be described in detail in this section. The algorithm contains two steps: extracting information and calculating confidence. Data preprocessing is used to extract information and exclude data without an apparent Hα emission line from the entire spectral data set. Then, WEDA is used to calculate the confidence based on information extracted from data preprocessing.

1) Extracting information

From an astronomical background, other emission lines b_n^1 chosen from the frame and neighbor data b_n^2 can demonstrate the possibility of the existence of the Hα emission line. At this stage, we not only extract information of the Hα emission line and Hβ emission line a_n but also determine the information of other emission lines b_n^1 chosen from the frame and six emission lines in the data formed by the superposition of the neighbor spectral data b_n^2. In this paper, we choose four emission lines from the frame (NII:6548,6584, SII:6717,6731) to obtain b_n^1 and calculate the confidence of six emission lines (Hβ:4862, Hα:6564, NII:6548,6584, SII:6717,6731) in data formed by superposition of neighbor spectral data to obtain b_n^2. Among them, three kinds of information need to be extracted, which are the quality of the Hα emission line and Hβ emission line a_n, the quality of other emission lines b_n^1 chosen from the frame and the superposition of neighbor spectral data b_n^2. The four other emission lines are chosen from the frame, which are two NIIs and two SIIs. The description of the process of extracting information is shown in Algorithm 1.

Algorithm 1 Extracting information

Input: dataset A_i; distance threshold θ_1
Output: unclassified data D_i
  for Hα emission line and Hβ emission line **do**
    extract information of each emission line
    if the value of Hα emission line is 0 **then**
      $a_n = 0$
      $y_n = -1$
    end if
  end for
  a_n = sum(Hα emission line and Hβ emission line)
  for four related emission lines chosen from frame **do**
    extract information of each emission line
  end for
  b_n^1 = sum(four related emission lines)
  for each data point in A_i **do**
    if y_n == 0 **then**
      D_i = D_i∪data
    end if
  end for
  for d_i in D_i **do**
    for d_j in D_i **do**
      if Euclidean distance(d_i, d_j) < θ_1 **then**
        $ds_i = ds_i \cup d_j$
      end if
    end for
    superposition of spectral data in ds_i
    b_n^2 = mean(confidence of six emission lines)
  end for
  D_i

The evaluation of emission lines To evaluate an emission line, we give an appropriate detection wavelength range for the emission line that can avoid covering other emission lines. After fixing the specified wavelength range, we can extract all peak values to check for the existence of this emission line. If there are no peak values around the emission line location, we conclude that this emission line does not exist; furthermore, the confidence of this emission line is recorded as 0. The wavelength of the ideal peak value should be close to the wavelength of the specific emission line, and the ideal peak value should be symmetrical. The peak value closest to this emission line is selected, and the wavelength distance between the peak value and the emission line is calculated. The smaller the wavelength distance is, the higher the confidence of this emission line. The inverse of the wavelength distance is obtained. To assess the symmetry of the peak value, the shape of this peak value also needs to be evaluated. The left and right sides of the ideal emission line should be similar, so the height and width of the left side and right sides are used for comparison. The widths of the two sides are recorded as w_l, w_r according to the change in slope, and the heights of the two sides are also recorded as h_l, h_r. The difference d can be obtained by the width and height of the two sides, and it is defined by the following formula:

$$d = \frac{min(w_l, w_r)}{max(w_l, w_r)} + \frac{min(h_l, h_r)}{max(h_l, h_r)}$$

The smaller the difference between the two sides is, the higher the confidence, and therefore, the inverse of the difference is also obtained. The inverse of the difference and the inverse of the wavelength distance together form the value of this emission line.

The primary information a_n includes two emission lines, Hα and Hβ. The precondition of spectral data with Hα is that there is a peak value at the Hα emission line location. Therefore, if the confidence of the Hα emission line is 0, the classification of data can be marked as -1. Only if the confidence of the Hα emission line is not 0 can the confidence of other emission lines be calculated. For primary information, the confidence of the Hβ emission line also needs to be calculated. The primary information a_n is the sum of the confidences of these two emission lines. When the primary information a_n is obtained, the four emission lines in the frame need to be calculated to obtain the value b_n^1. The four emission lines are two NII emission lines and two SII emission lines. Their confidence calculation method is the same as the above method. The sum of the four confidences of the emission lines is the value b_n^1.

Superposition of neighbor spectral data First, a distance threshold θ_1 needs to be given to determine the neighbor spectral data of all spectral data. After extracting the latitude and longitude from the spectral data, the distance between objects is calculated by the Euclidean distance. If we want to obtain the neighboring data of a spectral data point, this spectral data point needs to be compared with all spectral data to obtain all distances. If this distance is smaller than θ_1, the spectral data corresponding to this distance will be considered neighboring data of this data point and are added to the current data point's distance set ds_n. When all data are assigned, each data point has a distance set ds_n. We obtain b_n^2 by the distance set ds_n. The idea is that if most neighboring data have a high probability of having the Hα emission line, the probability that the data point has the Hα emission line will also be high. To obtain the probability that all neighboring data have the Hα emission line, all neighboring data will be superposed. We only need the superposition of Six emission lines' specified wavelength ranges instead of all wavelengths. Before the superposition of the specified wavelength range, the flux in the specified wavelength range needs to be normalized to eliminate differences between all neighboring spectral data. After completing the superposition of spectral data, the data formed by superposition of neighbor spectral data is used to obtain b_n^2. The confidence of six emission lines will be evaluated by the above evaluation method of emission lines. The mean of the confidence of the six emission lines is b_n^2. Finally, b_n^2 can be obtained for each data point to support the existence of the Hα emission line.

All information has been extracted by data preprocessing. Furthermore, some data have been determined by data preprocessing. For unclassified data $D_i = \{(a_n, b_n^1, b_n^2, y_n)\}_{n=1}^N$, the last two attributes b_n^1 and b_n^2 are obtained to show whether the unclassified data contain the Hα emission line.

2) Data ranking

For data D_i that cannot be classified by data preprocessing, the WEDA is proposed to improve the weight algorithm. Weights in the weight algorithm are fixed and unchanged, which makes it unsuitable for complex environments. Thus, the WEDA is proposed to improve the weight algorithm. We can have weights update constantly to adapt to complex situations. We first need to combine b_n^1 and b_n^2 to obtain b_n from predefined weights $\{\varphi_1, \varphi_2\}$; b_n is shown below.

$$b_n = b_n^1 \times \varphi_1 + b_n^2 \times \varphi_2 \tag{4}$$

The whole process of data ranking is shown in Algorithm 2.

Initial weights In real data, there are many data situations. To adapt to different data situations, the initial weights needed to be determined by the data. Based on primary information a_n and secondary information b_n, all unclassified data D_i are sorted. The top K data can be chosen to calculate initial weights $\{\omega_1, \omega_2\}$ by the difference between primary information a_n and secondary information b_n. The difference between primary information a_n and secondary information b_n is the weight difference. There are many ways to calculate the difference, and the difference should be related to specified values rather than the total difference. In view of the above, we need to determine the maximum difference in the top K data. The weight ω_1 is related to the primary information plus the maximum difference, while the weight ω_2 is related to the secondary information minus the maximum difference. All weights $\{\omega_1, \omega_2\}$ are limited to values between 0 and 1.

We can choose the top K data from these data to represent the data that can be classified without secondary information b_n; the selection method is as follows:

1. First, the primary information of all data is normalized as [0,2]. All data with primary information a_n larger than 1 can be separated by the following formula.

$$\kappa = (a_n \% 1) \times 10 \tag{5}$$

$$\lambda_\kappa = \lambda_\kappa + 1 \tag{6}$$

where κ is an integer that represents which level the data should be assigned to, and the range of κ is [0, 9]. The values of all levels λ_κ are initialized to 0. The value of the level is λ_κ plus 1 when κ is equal to the level. After that, all data have been processed, and then we use it to choose K.

2. Second, if the value in the previous level λ_κ is two times larger than the value of the next level $\lambda_{\kappa-1}$ and the value of the next level $\lambda_{\kappa-1}$ is larger than the minimum threshold of 4, the next level $\lambda_{\kappa-1}$ is called the cutoff level. The reason that the minimum threshold is set to 4 is that K should be limited to prevent too few numbers being used, making the factor of two meaningless. All data in the levels $\{\lambda_\kappa\}_{k=\kappa}^9$ before the cutoff level are chosen to be the top K data, and half of the data in the cutoff level $\lambda_{\kappa-1}$ are also chosen as the top K data.

Before calculating the difference of the top K data, the primary information a_n and secondary information b_n must have uniform standards. In this paper, a sorted ranking is

Algorithm 2 Data ranking

Input: unclassified data D_i; predefined weights $\{\varphi_1, \varphi_2\}$; learning rate τ; dissipation coefficient δ
Output: confidence sequence C
  for b_n^1 and b_n^2 in D_i **do**
    $b_n = b_n^1 \times \varphi_1 + b_n^2 \times \varphi_2$
  end for
  a_n is normalized as [0,2]
  for a_n in D_i **do**
    $\kappa = (a_n \% 1) \times 10$
    $\lambda_\kappa = \lambda_\kappa + 1$
  end for
  for i in range(0,9,-1) **do**
    if $\lambda_\kappa / \lambda_{\kappa-1} > 2$ and $\lambda_{\kappa-1} > 4$ **then**
      $topK = topK \cup$ half data that $((a_n \% 1) \times 10 = i)$
      break
    end if
    $topK = topK \cup$ all data that $((a_n \% 1) \times 10 = i)$
  end for
  sorted($\{a_n\}$)
  sorted($\{b_n\}$)
  for k in topK **do**
    $V_k^{a_n} = \frac{\ell - \gamma_k^{a_n}}{\ell}$
    $V_k^{b_n} = \frac{\ell - \gamma_k^{b_n}}{\ell}$
    d = max(d, $V_k^{a_n} - V_k^{b_n}$)
  end for
  $\omega_1 = min(0.5 + d, 1)$
  $\omega_2 = max(0.5 - d, 0)$
  while len(C) $<$ number of unclassified data points **do**
    for a_n, b_n in D_i **do**
      $c_n = a_n \times \omega_1 + b_n \times \omega_2$
    end for
    $C = max(\{c_n\}) \cup C$
    $C = D_i \setminus max(\{c_n\})$
    $gap = \tau * (\delta \times W + |a_{x_n^*} - a_{x_{n-1}^*}|)$
    $\omega_1 = \omega_1 - gap$
    $\omega_2 = \omega_2 + gap$
    if ω_1, ω_2 meet the requirement **then**
      $\omega_1 = max(0.6, \omega_1 - 0.2)$
      $\omega_2 = 1 - \omega_1$
    end if
  end while
  return C

applied to create a uniform standard. The higher the ranking is, the larger the value. The formula is defined as follows.

$$V_n = \frac{\ell - \gamma_n}{\ell} \tag{7}$$

where ℓ represents the total number of data and γ_n is the ranking of the data. Both primary information and secondary information can have their own V_n, and the two V_n are subtracted to obtain the difference d_n. d is the maximum of all differences d_n.

$$\omega_1 = min(0.5 + d, 1) \tag{8}$$
$$\omega_2 = max(0.5 - d, 0) \tag{9}$$

Weight update Based on the initial weights $\{\omega_1, \omega_2\}$ obtained previously, our purpose is to have the weight ω_1 related to the primary information slowly decrease when the weight ω_2 related to the secondary information slowly increases. We divide the process of changing weights into three stages. Because we assume that the data that have the Hα emission line must have a high primary information value a_n or secondary information value b_n, the two weights $\{\omega_1, \omega_2\}$ should be quite different rather than close. In the first stage, the data that only depend on primary information a_n are found, and then they remain in the second stage for a short time. In this period, the two weights $\{\omega_1, \omega_2\}$ are very close, so we can detect data that have both a value of primary information a_n and a value of secondary information b_n slightly higher than those of other data. In the third stage, the weight ω_2 related to secondary information is greater than the weight ω_1 related to primary information, and data are chosen that only depend on secondary information b_n. When the number of chosen data points is greater than the threshold of 0.7 or the secondary information of the data b_n is very small, the algorithm will go to the next iteration. In the second iteration of processing, the weight ω_1 is obtained by the weight ω_1 in the first iteration minus the iteration threshold 0.2. This is because the previous iterative process detects the most data that depends on primary information a_n, so the weight ω_1 does not need to be set as large as before. However, the weight ω_1 in all iterations is greater than or equal to 0.6, which makes the primary information dominant at the first stage of processing. The weight ω_1 is calculated by the following formula.

$$\omega_1 = max(0.6, \omega_1 - 0.2) \tag{10}$$

Finally, the above three-stage process is repeated until all data are processed.

The data become increasingly dependent on secondary information, so the speed of descent should increase. The spacing distance can be calculated by a formula defined as follows.

$$gap = \tau * (\delta \times W + |a_{x_n^*} - a_{x_{n-1}^*}|) \tag{11}$$

where τ is the learning rate, δ is the dissipation coefficient, $a_{x_n^*}$ represents the value of primary information of the chosen data, and $a_{x_{n-1}^*}$ represents the value of primary information of the next data. W is the sum of all previous weights. The weights $\{\omega_1, \omega_2\}$ are updated by the following formula.

$$\omega_1 = \omega_1 - gap \tag{12}$$
$$\omega_2 = \omega_2 + gap \tag{13}$$

$\{\omega_1, \omega_2\}$ can be used to calculate all the data's confidence $\{C_n\}_{n=1}^N$ at the current stage. In practical applications, we can take the first iteration into account. In the first iteration, in

this paper, we set the minimum gap in the first two stages and the maximum gap in the second stage to prevent the weights from decreasing too quickly or too slowly. The maximum gap and minimum gap are also set according to the number of data in this stage. All data can be sorted through confidence $\{C_n\}_{n=1}^{N}$, and we can choose the data corresponding to the highest confidence C_n; the data with the highest confidence C_n are considered to have the Hα emission line. When the number of chosen data points is greater than the threshold 0.7 or the secondary information of data b_n is very small, the weight-changing process will go to the next iteration. In all subsequent iterations, complete iterative processing does not include three stages, but we need to set gap restrictions. After multiple iterations, all data have been processed and sorted.

C. THEORETICAL ANALYSIS

In this algorithm, data preprocessing is indispensable to the entire algorithm because it greatly reduces the amount of data that the algorithm operates on, which can significantly reduce time complexity and spatial complexity. After removing most of the easily determined data, only a small portion of undetermined data will be used in WEDA. The time complexity of WEDA is $O(N^2)$, where N is the number of data points that fail to be determined by data preprocessing. The learning rate τ and dissipation coefficient δ have a great impact on the algorithm. The two parameters can affect which data the algorithm prefers. The larger the value of the two parameters is, the faster the weight ω_1 decreases; the algorithm can select most of the data with more secondary information, which indicates that the algorithm is biased towards data that only depend on secondary information. The smaller the value of the two parameters is, the more slowly the weight ω_2 decreases; the data with more primary information will be chosen by this algorithm, which means that the algorithm prefers data that only depend on primary information. In this algorithm, the speed at which the weights decrease affects the data selected by the algorithm; therefore, we need to limit the size of the weight decrease.

The three subsets have their own ordered sequences. The higher the rank in the sequence is, the higher the probability of the appearance of the Hα emission line. We only need a certain threshold as the boundary value of the two types of data, and the boundary value is called the cutoff threshold.

IV. EXPERIMENTAL RESULTS

In this section, we rigorously evaluate our method from two perspectives on an Intel(R) Core(TM) i7-6700HQ with 8.0 GB memory with a Windows 10 operating system. In addition, to verify whether our method works efficiently on spectral data, we choose five different sizes of data as classification data and six different classification algorithms to compare. The update of the weights can be analyzed based on figure 6.

We implemented our method and all the classification algorithms with Python. In this paper, spectral data from LAMOST DR5 are used in all algorithms. For the description of the characteristics of emission lines during data preprocessing, the wavelength range used for the description of the characteristics needs to be determined. Table 1 shows the wavelength ranges of the six emission lines. The six wavelength ranges are used to extract the emission lines' information. The distance threshold θ_1 is initialized to 1, and the predefined weights $\{\varphi_1, \varphi_2\}$ are set to $\{0.8, 0.2\}$ for all tasks. The learning rate τ and the dissipation coefficient δ are set to 0.1 when the number of data points is fewer than 20000. In practical applications, we need to constrain the gap to decrease and increase the control weights. In the first stage of weight update, there is a minimum gap to prevent the primary weight from decreasing too slowly. The minimum weight related to the primary information must be greater than 0.8, and the number of data points chosen with a level greater than or equal to 3 should be 0.9 of the total number of data points with a level greater than or equal to 3 when the number of all data points is smaller than 20000. The minimum gap is based on the ratio of the weight range related to primary information and the number of data points for which κ is greater than or equal to 3. The minimum gap and maximum gap also need to be set in the second stage. The weight related to primary information is limited to a range of 0.5 to 0.8. The difference between the minimum gap and maximum gap is the difference in the number of data points in which κ is greater than or equal to 1 and smaller than 3.0. The number of data points in the minimum gap is set to 0.5, and the number of data points in the maximum gap is set to 0.3. In subsequent iterations, we can set the minimum gap and the maximum gap for complete iterative processing. The number of data points in the minimum gap and maximum gap are set to 0.2 and 0.25, respectively.

For each classification method, we repeat the same experiments multiple times with different amounts of spectral data and obtain the execution time, recall rate and precision. For each experiment, our method and the compared methods share the same data for a fair comparison. The process of updating the weights in 19411 data points is also shown and analyzed. The data sets used in this paper are described in detail in section A. The other implementation details and experimental results for each task and algorithm are shown in sections B and C.

A. DATA DESCRIPTION

In this paper, the dataset for all algorithms is obtained from LAMOST DR5 V3. LAMOST, also called the Guo Shou Jing Telescope, can take 4000 spectral images in a single exposure. LAMOST DR5 spectral data were obtained during a six-year sky survey from October 2011 to June 2017, and they include 4154 astrometric fields and 9026365 total spectral data. The number of high-quality spectra that have an SNR greater than 10 reached 7775981. For different amounts of data that all the experiments require, we only need to set different position constraints and choose star and galaxy data. The amount of data is 617, 3611, 8202, 12782, and 19411.

The standard data with Hα emission line are shown in

TABLE 1: The characteristic description of emission lines

	Hα	Hb	NII	NII	SII	SII
wave_len	(6553,6576)	(4851,4874)	(6538,6561)	(6574,6597)	(6707,6730)	(6721,6744)

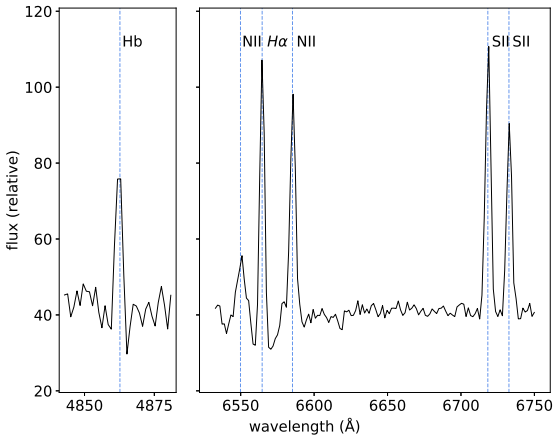


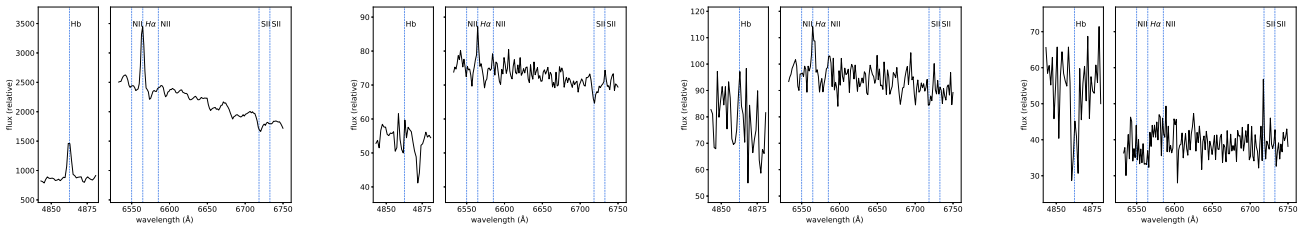
FIGURE 1: Standard spectra with the Hα emission line

figure 1. The horizontal coordinate represents wavelength and the vertical coordinate represents flux. The emission lines at specified wavelengths are obvious and have good characteristics in figure 1. We find that these data with Hα emission line also contain two NII and two SII emission lines and an Hβ emission line. This is because these six emission lines are related. These correlations are divided into two categories: one is the correlation between the Hα emission line and the Hβ emission line, and the other is the correlation between the Hα emission line and four other emission lines. Based on the above idea, data with the Hα emission line can be divided into two categories. However, these data are easy to detect because their characteristics are very obvious and there are six emission lines. Many data in reality can be disturbed by noise, which weakens the characteristics so that it is difficult to distinguish whether the data contains the Hα emission line. These data with weak characteristics require other emission lines to judge whether they contain the Hα emission line.

In general, spectral data with the Hα emission line also contain other related emission lines, which means that if there are only obvious characteristics in the Hα emission line location but no characteristics in other emission lines, it cannot be demonstrated that the spectral data contain the Hα emission line. The spectral data with the Hα emission line are bound to contain other emission lines, such as the Hβ emission line. Based on the above two correlations, we can choose spectral data with the Hα emission line.

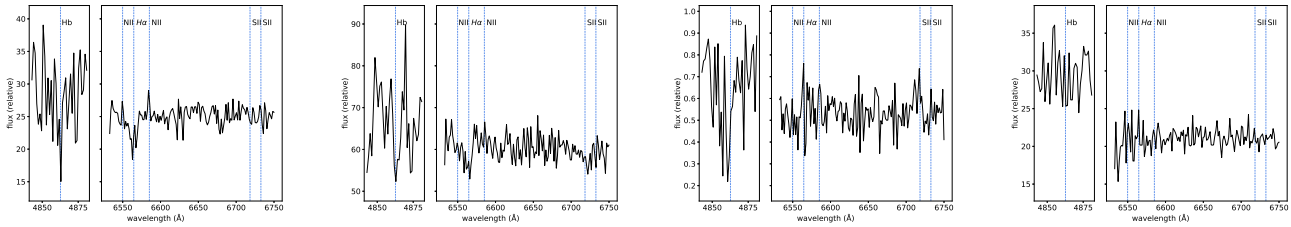
The first correlation is between the Hα emission line and the Hβ emission line. When two SII and two NII emission lines are disturbed by noise, this kind of data still contain an Hβ emission line, and the Hα and Hβ emission lines are strongly related. Figure 2 shows some examples with this connection. As seen from the figure, these data contain at least the Hα emission line and Hβ emission line, but each data point has its own unique situation. In general, data with the Hα emission line have other emission lines in the frame, such as NII, SII and Hb. In these figures, each data must contain an Hβ emission line. Moreover, different spectral data have different situations; for example, in figure 2-a, the spectral data have an obvious Hβ emission line and two weak NII emission lines, and the spectral data in figure 2-b also have an NII emission line and an SII emission line, but not the Hα emission line. Figure 2-c and figure 2-d also show that spectral data with the Hα emission line also have an Hβ emission line at least. Based on these spectral data, the quality of Hα and Hβ emission lines are used as primary information. Therefore, this type of data can be judged directly by the Hα emission line and Hβ emission line instead of requiring additional information such as frame and neighbor data classification. This type of data only needs the weight related to primary information because some data of this type do not have other emission lines chosen from the frame or are surrounded by data without the Hα emission line. If the weight related to the secondary information exists, or is greater than the weight related to the primary information, this type of data is difficult to detect. Moreover, there are some special cases, for example, it may be observable by the human eye that spectral data contain the Hα emission line and other emission lines but no Hβ emission line. This type of data can be detected by the second iteration or even the third iteration in the algorithm.

The second correlation is between the Hα emission line and four other emission lines; this is shown in figure 3. In figure 3, the spectral data have an SII emission line or NII emission line but not an Hβ emission line. In contrast to the above spectral data, these spectral data do not require an Hβ emission line. Figure 3-a and figure 3-b have two NII emission lines, and an SII emission line and three emission lines can demonstrate the existence of the Hα emission line. The two spectral data in figure 3-c and figure 3-d have two SII emission lines and two NII emission lines. The more emission lines there are, the higher the confidence. Based on these data situations, the quality of NII and SII emission lines are used as secondary information. For this type of data, the quality of the Hα emission line and Hβ emission line used as primary information is useless and even interferes with detection. The detection of this type of data only depends on four emission lines in frame and neighbor data classification. Therefore, only the weight related to the secondary information needs to be used to detect data with



(a) Hα, Hβ (b) Hα, Hβ, SII (c) Hα, Hβ, NII (d) Hα, Hβ, NII, SII

FIGURE 2: Spectra that depend on primary information



(a) Hα, NII (b) Hα, NII (c) Hα, SII (d) Hα, SII

FIGURE 3: Spectra that depend on secondary information

the Hα emission line. From the four images in figure 3, all data are ultimately found to have three emission lines in the frame. Most data with weak Hα emission line have SII emission lines or NII emission lines, or even both emission lines. The neighboring data are also used to demonstrate the existence of the Hα emission line. If most neighboring data contain the Hα emission line, the data point is more likely to contain the Hα emission line. For frame and neighboring data classification, the frame is more useful than neighboring data classification because these data may be outliers and data with the Hα emission line generally has other emission lines. For the frame, there are ultimately two emission lines, and the frame can be used to demonstrate the existence of the Hα emission line. If the frame only has one emission line, the frame information becomes useless. The secondary information can be dominant when data do not contain an Hβ emission line. The weights in the algorithm need to be adjusted to adapt to this situation, so that the algorithm can detect data containing other emission lines in the frame.

B. ANALYSIS OF THE QUALITY OF WEDA

The objective of the experiments presented in this subsection is to analyze the quality of WEDA when data sets with different numbers are used as test data sets. The five data sets are used to test the quality of the algorithm, which can efficiently test the quality of the algorithm for various amounts of data.

Two perspectives are used to analyze the quality of our algorithm. The first perspective is different SNRs. Before preprocessing, the entire dataset is divided into three parts based on the SNR: 0-10, 10-50 and above 50. Experiments with different SNRs can help us to determine which area's SNR data are difficult to classify and reduce the overall quality of this algorithm, which is a drawback. of the algorithm. The second perspective is the amount of data. Each part is divided into three parts based on its own situation and cutoff threshold. We choose a cutoff threshold such that the recall rate is 1 to compare the quality of our algorithm under different amounts of data.

1) Different SNRs and different cutoff thresholds

In this subsection, the influence of different SNRs on quality is compared. As mentioned above, we use five data sets with different amounts of data and obtain the recall rate and precision for each. The results are shown in figure 4. The descriptions of five data sets are shown in Table 2. In figure 4, each color represents a data set. In each line, there are three images: the images in the first line show the recall rates on the five data sets, and the images in the second line show the precision on the five data sets. The horizontal axis represents the cutoff threshold, which is the ratio of the number of data points with the Hα emission line in the data set D_i and the number of the entire data set D_i; the data set is all data points such that it cannot be determined whether they have the Hα emission line by data preprocessing.

The first cloumn of figure 4 demonstrates the quality of the algorithm on five data sets when the SNR is between 0 and 10. The recall rates in the five data sets are shown in figure a. From figure a, we can find that as the amount of data increases, the cutoff threshold consistently become

TABLE 2: The descriptions of five data sets

	number of data set	0 < SNR < 10	10 < SNR < 50	SNR > 50
data set A	617	54	202	361
data set B	3611	476	1953	1182
data set C	8202	772	3889	3541
data set D	12872	1803	4801	6268
data set E	19411	1903	7191	10317

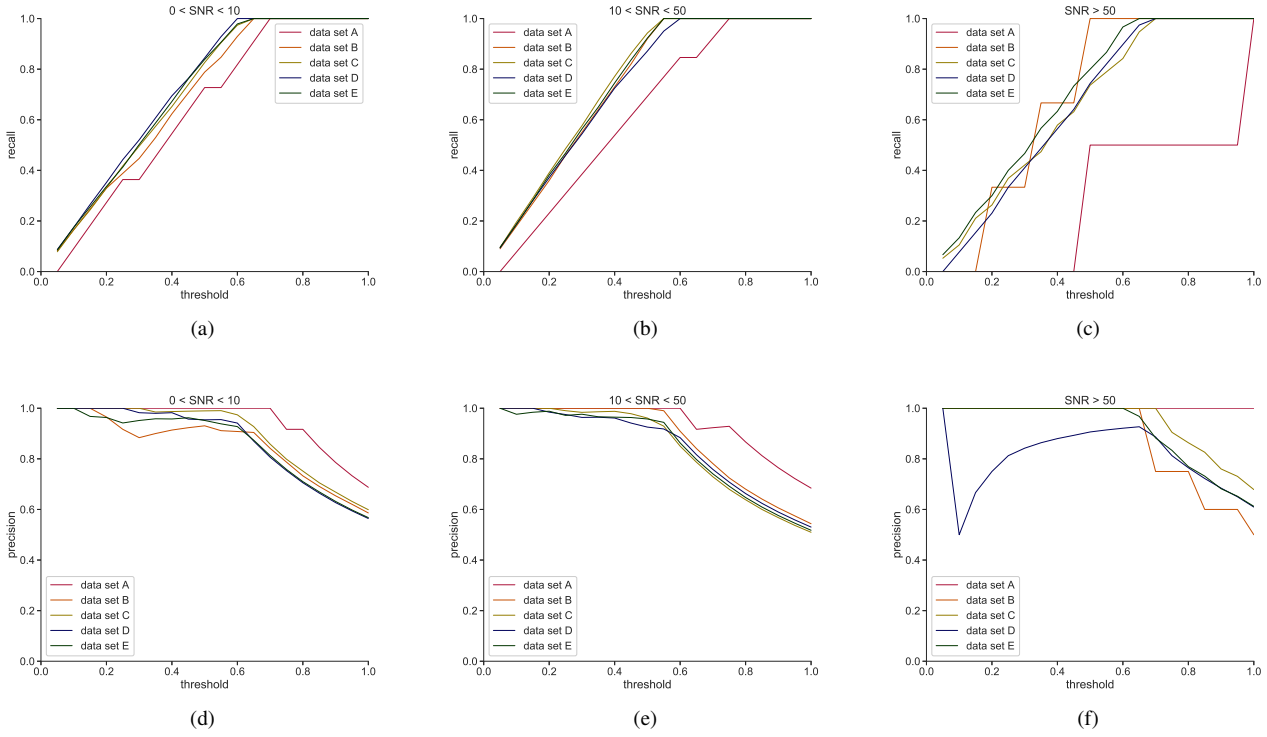


FIGURE 4: This figure shows the influence of different SNRs on quality is compared. Five data sets that are 617, 3611, 8202, 12872, 19411 are used and the recall rates and precisions are obtained. The first column represents data that SNR is between 0 and 10. The first line in the first column that is figure a is recall rate of data that SNR is between 0 and 10. The second line in the first column that is figure d is precision of data that SNR is between 0 and 10. The second column represents data that SNR is between 10 and 50. The first line in the second column that is figure b is recall rate of data that SNR is between 10 and 50. The second line in the second column that is figure e is precision of data that SNR is between 10 and 50. The third column represents data that SNR is above 50. The first line in the third column that is figure c is recall rate of data that SNR is above 50. The second line in the third column that is figure f is precision of data that SNR is above 50.

stable and eventually converge to a stable value of 0.65. The data preprocessing has excluded some data that are believed to not contain the Hα emission line, and these data are not used in the following processing of the data ranking. The data preprocessing cannot identify data with the Hα emission line. When the cutoff threshold is 0, the recall rates are 0 because no data are believed to contain the Hα emission line in the stage of data preprocessing. The whole process of data ranking only processes data that cannot be determined by data preprocessing. The cutoff threshold is approximately 0.7 when the amount of data is 617. During this time, the amount of data used in the algorithm is small, which makes the data situation simple and the confidence of data with the Hα emission line higher in general, so that a larger cutoff threshold can find all data with the Hα emission line. As the amount of data increases, the data situation becomes gradually comprehensive, and there are some special cases among these data. In the figure, when the amount of data reaches 19411, the cutoff threshold begins to remain stable. Compared to the cutoff thresholds in small data sets, the cutoff threshold was close to 0.65 at this time and did not change much. The reason for this phenomenon is that the

number of data situations has been increasing, however, the number of data points with the Hα emission line has also increased.

The SNR of this data set is 0 to 10, and the SNR of all the data is relatively close, so when the amount of data is similar, their cutoff thresholds are also close. Because the SNR is relatively low, the confidence will be disturbed by noise. Some data that depend on primary information can only be determined during the second iteration, or even the third iteration, which makes the cutoff threshold larger.

Figure d shows all precision values in the five data sets in which the SNR is 0 to 10. The green line represents that the amount of data is 19411 and is the first to begin to slump. Because the SNR is low, the data that depend on primary information are disturbed by noise. The weight related to the primary information remains dominant at the stage that identifies data that depend on primary information, therefore, some data that do not contain the Hα emission line also have high confidence. The orange line representing 3611 has the largest drop when the cutoff threshold is 0.15. Due to the small amount of data with the Hα emission line in small data sets, even if the amount of unclassified data is small, the precision will slump. When the cutoff threshold is 0.15, the recall rate does not reach 1, which shows that some spectral data without the Hα emission line have a high ranking. The pink line starts to drop when the recall rate has reached 1. Small data sets have simple data situations, so misclassified data are relatively less common. Other lines also show cases in which the line only drops slightly because the amount of data is large.

The number of data points for which the SNR is between 0 and 10 is the smallest of the three parts, and the amount of data for the Hα emission line is also smaller. In a small dataset, no lines have a steady trend, because the data situation is not comprehensive and the value generated by the data preprocessing may be inaccurate in classifying some data due to the lower SNR.

The second cloumn of figure 4 shows the recall rate and precision of datasets for which the SNR is between 10 and 50 using different sizes of data sets. All recall rates in the five data sets are shown in figure b. When the amount of data is small, the cutoff threshold varies greatly. For example, the pink line reaches 1 when the cutoff threshold is 0.75. From this point of view, the cutoff threshold has no reference value at this time, and the true cutoff threshold cannot be determined. One of the reasons for this result is that small data sets have unique situations that are a small part of the overall situation. From figure b, the pink line, green line and blue line reach 1 at different cutoff thresholds, and the three cutoff thresholds have great differences. As the number of data points increases, the cutoff threshold will slowly stabilize. The green line representing 19411 reaches 1 when the cutoff threshold is 0.55, and the precision falls to 0.94. The pink line reaches 1 when the cutoff threshold is 0.75. The blue line reaches 1 when the cutoff threshold is 0.6. This is mainly because data that have an SNR between 10 and 50 are complicated, and the range of the SNR is too large, resulting in a large difference in the SNR of each data point. It is difficult for the algorithm to detect all data perfectly, which can be seen below in figure e. When the recall rate reaches 1, most precision values slowly declined to 0.9.

The quality of the data for which the SNR is between 10 and 50 is very steady, and the precision is better than that of the data for which the SNR is between 0 and 10. As mentioned before, As the SNR increases, the characteristics of emission lines in spectral data become increasingly obvious. When the number of data situations begin to increase, the result does not change very much. However, the data with the Hα emission line become increasingly complicated as the amount of data increases. Some data are regarded as outliers for which the values of both primary information and secondary information are small, which makes the confidence low. The ranking of these data is decreased. Therefore, when the recall rate reaches 1, the precision also drops slightly. Eventually, the precision can become low.

It can be seen that the precision values are not high when the recall rate is 1. This is because the confidence of some special data is low and it is ranked lower in the sequence. There is often difficulty in extracting valid feature information of such data through data preprocessing.

The results of data for which the SNR is greater than or equal to 50 are shown in the third column of figure 4. The line is not similar to the above two results and has greatly changed. From figure c, the recall trend of the pink line representing 617 is not smooth. It is easy to detect the data with the Hα emission line in the datasets with a higher SNR. In this range of SNR, data with the Hα emission line have high confidence, and many data can be determined by data preprocessing. There is only a small amount of data left, including some data with the Hα emission line. The value of the primary information of data with the Hα emission line is high because the characteristics of this spectral data are very obvious, so it is easy to detect. For the green line representing 19411, when the cutoff threshold is 0.6, the green line reaches 1. From figure f, all lines except the blue line remain at 1 when the recall does not reach 1, which shows that all spectra are correctly classified. The blue line representing 12872 drops to 0.5 when the cutoff threshold is 0.1. This is because there are misclassified spectral data initially. Then, the blue line starts to rise to 0.91 until the cutoff threshold reaches 0.7, which shows that the rest of the spectral data are correctly classified. On the whole, spectral data with SNR above 50 are still easy to classify. When the SNR is greater than or equal to 50, the characteristics of all data with the Hα emission line are very obvious.

We analyze the quality of the three parts of the data from the above six figures. Data for which the SNR is 50 and above are easy to classify based on high confidence in data with the Hα emission line.

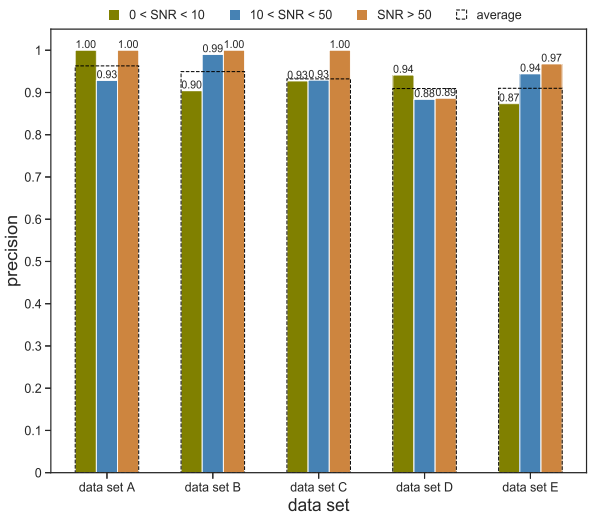


FIGURE 5: Precision of WEDA on different data sets

2) Different amounts of data

After analyzing the recall rate and precision, the impact of the amount of data on the algorithm is analyzed in this subsection. The algorithm can divide all data into three parts based on SNR, and each part of the data is processed separately. According to the above results, we choose a cutoff threshold based on the recall rate. The requirement of the experiment is to identify all data with the Hα emission line, so the recall rate should reach 1. Under the premise that the recall rate reaches 1, the cutoff threshold is chosen to calculate the overall precision. Five different sizes of datasets are used in the experiment. The result is shown in figure 5.

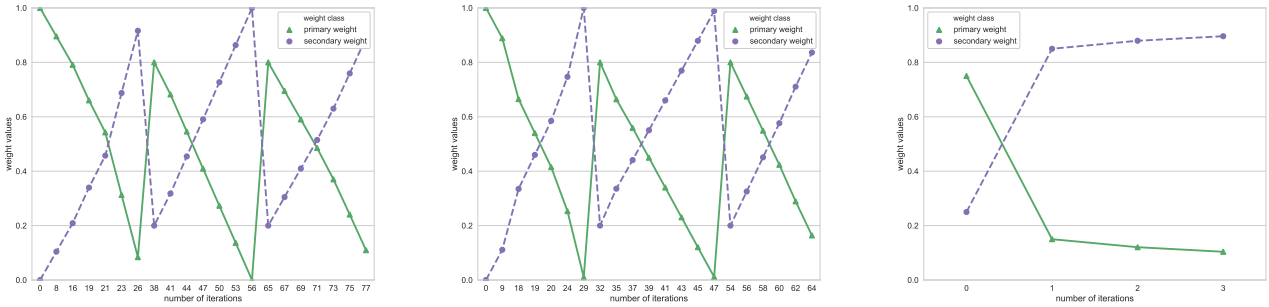
In figure 5, the green bar represents the precision of data which SNR is between 0 and 10, and the blue bar represents the precision of data which SNR is between 10 and 50, and the precision of data which SNR is above 50 is represented by the orange bar. The relationship between the SNR and the precision can be found by the comparison of three bar. In general, SNR is larger and precision is higher, which also is demonstrated by figure 5. However, data set A and data set D have an abnormal situation. The precision of data which SNR is between 0 and 10 is the highest. In data set A, the amount of data is small. Even if the number of misclassified data is small, the precision of data is lower. In data set D, both data sets that SNR is between 10 and 50 and SNR is above 50 have special situations. The data that SNR is between 0 and 10 in data set D only has simple situation. From five bars representing average, are consisted of black dotted line, the precision falls from 617 to 12872 due to the more complicated data situation and larger dataset with Hα emission lines. Therefore, the quality of the algorithm is still not stable. At the same time, the amount of data has a great influence on the algorithm. A small dataset only represents a part of the situation. As the amount of data increases, the precision changes. The overall trend of the pink line is downward, while the precision begins to stabilize at 0.91 in the dataset sized 19411. The larger the dataset, the more data situations it contains; this stabilizes the algorithm, so we can see that the precision is stable when the data size is increased from 12872 to 19411. For large data sets, there are too many data situations to accurately extract information from spectral data. The height of green bar starts at 0.96, which shows that the algorithm still has misclassified data in small data sets. There should be less misclassified data with an SNR between 10 and 50. The height of green bar eventually increases slightly, but the amplitude is small, which shows that the data situations remain stable.

3) Weight analysis

From the above analysis, we find that the SNR affects this algorithm because data sets in different SNRs have their own data situations, which affects the weights in the algorithm. In this subsection, we will analyze how the weights update in the process of this algorithm. This data set contains 19411 pieces of data and is put into WEDA to obtain the figure 6 showing the updates of the weights.

The three images in figure 6 represent the update of weights in three SNRs. The different horizontal coordinates in the three images represent different ranking numbers, and figure 6-a has the highest ranking because it is difficult for data with an SNR between 0 and 10 to be identified by data preprocessing, and figure 6-c has the lowest ranking due to the high SNR, which makes data easier to identify by data preprocessing. The different SNR values lead to different ranking numbers due to different amounts of data complexity. In contrast to figures 6-a and 6-b, the initial primary weight in figure 6-c is only 0.75. From the introduction of this algorithm, initial weights are determined by differences in specific data. The initial primary weight in figure 6-c is only 0.75, which shows that the difference of data that is not determined by data preprocessing is small and the data situation for high SNRs is relatively simple.

The processes of the algorithm in figures 6-a and 6-b have three iterations. Both processes have almost the same overall trend and are different in parts. The update of weights is determined by the difference of primary information. If primary information of data with a closed ranking has a large difference, the weight related to the primary information tends to decrease faster. The data in figures 6-a and 6-b can be seen to have low SNR and complex data situations, which require multiple iterations to find all data with the Hα emission line. In figure 6-a, the weight related to the primary information at the end of the first iteration does not fall to 1, which indicates that most data with an SNR between 0 and 10 that are undetermined by data preprocessing depend on secondary information and that the difference in secondary information in these data is small. The figure 6-c shows that the detection of data with the Hα emission line only requires the first iteration. The weight related to primary information falls from 0.75 to 0.15 after the first ranking. One of the



(a) 0 < SNR < 10 (b) 10 < SNR <50 (c) SNR > 50

FIGURE 6: The update of weights

reasons for this phenomenon is that data with a high SNR have very obvious characteristics, which indicates that data with the Hα emission line have a large amount of primary information and it is easy to distinguish these data.

C. COMPARISON WITH OTHER ALGORITHMS

In this section, we carry out comparisons of WEDA and other related methods: adaptive semi-unsupervised weighted oversampling (ASUWO) [54], the postprocessing technique for a support vector machine (BSVM) [55], a minimally spanned support vector machine (MSSVM) [56], a semi-supervised heterogeneous ensemble classifier (multi-train) [57], SSO-SMOTE-SSO [58], and vote-boosting ensembles (VBensembles) [59]. The main comparisons have three aspects: precision, recall rate, and execution time.

For WEDA, data are classified when the recall rate is 1. We analyzed the quality of this algorithm from the perspective of data size and SNR, selecting the case where the recall rate is 1.

1) Precision and recall rate

The experiments are carried out on the five different data sets, and all algorithms' recall rates and precisions are obtained. The recall rate and precision of all algorithms are displayed in figure 7, in which each color represents an algorithm. In the figure, the dotted line represents the recall rate, and the solid line represents the precision.

From the results presented in the figure, it is clear that WEDA shows the best overall quality. Under the condition that the recall rate is 1, the precision of WEDA outperforms all other algorithms. In all compared algorithms, the precision of BSVM, the yellow dotted line, reaches 1, but the recall rate is very low, which indicates that it only finds a small amount of data with the Hα emission line; the result for the MSSVM is the same. These compared algorithms only detect particularly obvious data for which the Hα emission line's confidence is high. For data that need to use secondary information, such algorithms can ignore these data and be unable to judge them.

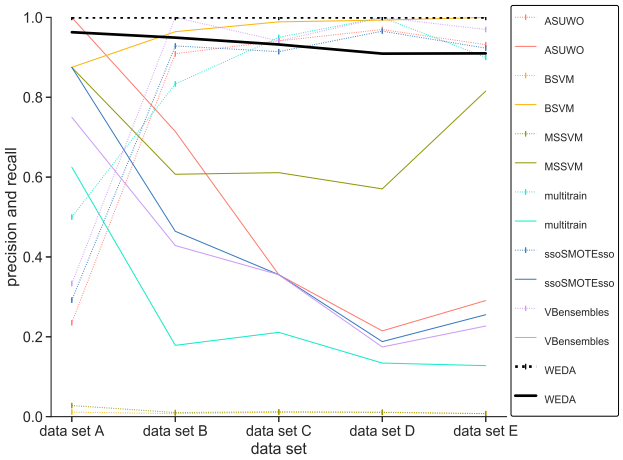


FIGURE 7: Precision of all algorithms

For other algorithms, their recall rate can reach a satisfactory level, but they have poor precision. These algorithms use features of minority data and oversample a small amount of the data. Although data recall rates increase, data precision decreases. For spectral data including stars and galaxies with the Hα emission line, these models are unsuitable.

As more and more data are input, the dataset becomes increasingly complicated. The recall rate is unable to reach 1 and the precision also remains low, which makes WEDA inevitably worse, but its recall rate and precision remain at a good level.

2) Execution time

The comparison of time complexity is shown in figure 8. The figure shows the time complexity of seven algorithms in three data sets, sized 3611, 8202 and 19411, separately. From the figure, we find that MSSVM, VBensembles, BSVM, and multitrain have a very high time cost in a small dataset and their running time does not change much with an increase in the amount of data. This is mainly related to the number

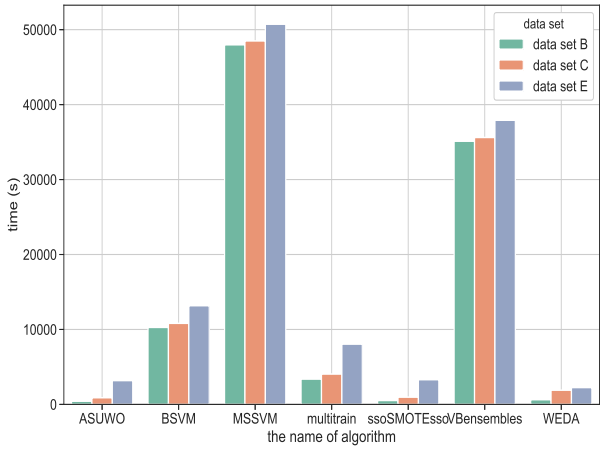


FIGURE 8: Execution time of all algorithms

of iterations of the algorithms themselves rather than the amount of data. The remaining three algorithms' time complexity can be affected by the amount of data. Their running time on all data sets is lower than that of the above four algorithms. The figure clearly shows that our algorithm has the lowest running time. The stage of data preprocessing in our algorithm determines many data points and reduces the amount of data. Although WEDA processes data in multiple iterations, the amount of data that can put into WEDA is few and the time required for the process decreases. Our algorithm's running time increases when the amount of data increases, but the magnitude of our algorithm's increase is minimal.

V. DISCUSSION

In this paper, we proposed a novel ranking algorithm called WEDA. WEDA uses changeable weights instead of fixed weights to adapt to complicated data and determine data that contain the Hα emission line in different situations. Both weight update and initialization are based on data rather than artificial settings to adapt to different data sets. In addition, we use this algorithm to detect Hα emission line in star and galaxy data, and its quality is confirmed very well compared with other algorithms by experiments; it is more effective than other algorithms in detecting Hα emission line. Regarding maintaining a high recall rate, the precision of this algorithm can also reach a satisfactory level. Even if the amount of data increases, the overall quality of this algorithm will not become very poor. The experimental results indicate that WEDA can be used to detect Hα emission line in star and galaxy data. In the future, we plan to apply it to large-scale data to detect Hα emission line and make the algorithm applicable to other types of data.

ACKNOWLEDGMENT

The Guo Shou Jing Telescope (the Large Sky Area MultiObject 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.

The work is supported by the National Natural Science Foundation of China (Grant Nos. U1931209, U1731126) and Shanxi Province Key Research and Development Program (Grant Nos. 201803D121059, 201903D121116).

REFERENCES

[1] QU Cai-Xia, YANG Hai-feng, CAI Jiang-hui, XUN Ya-ling. P-Cygni Profile Analysis of the Spectrum: LAMOST J152238.11+333136.1[J]. Spectroscopy and Spectral Analysis, 2020,40(4): 1304-1308. Doi:10.3964/j.issn.1000-0593(2020)04-1304-05

[2] Guo X, Zheng F, Li C, et al. A portable sensor for in-situ measurement of ammonia based on near-infrared laser absorption spectroscopy[J]. Optics and Lasers in Engineering, 2019, 115: 243-248.

[3] Pérez-González P G, Zamorano J, Gallego J, et al. Spatial analysis of the Hα emission in the local star-forming UCM galaxies[J]. The Astrophysical Journal, 2003, 591(2): 827.

[4] Mármol-Queraltó E, McLure R J, Cullen F, et al. The evolution of the equivalent width of the Hα emission line and specific star formation rate in star-forming galaxies at 1< z< 5[J]. Monthly Notices of the Royal Astronomical Society, 2016, 460(4): 3587-3597.

[5] Pineda J S, Hallinan G, Kirkpatrick J D, et al. A Survey for Hα Emission from Late L Dwarfs and T Dwarfs[J]. The Astrophysical Journal, 2016, 826(1): 73.

[6] Hosoya K, Itoh Y, Oasa Y, et al. Spectroscopic Survey of Hα Emission Line Stars Associated with Bright Rimmed Clouds[J]. International Journal of Astronomy and Astrophysics, 2019, 9(2): 154-171.

[7] Borne K. Scientific data mining in astronomy[J]. arXiv preprint arXiv:0911.0505, 2009.

[8] Liu C, Cui W Y, Zhang B, et al. Spectral classification of stars based on LAMOST spectra[J]. Research in Astronomy and Astrophysics, 2015, 15(8): 1137.

[9] Yang, Y., Cai, J., Yang, H., Zhang, J., & Zhao, X. TAD: A trajectory clustering algorithm based on spatial-temporal density analysis. Expert Systems with Applications, 139, 112846. https://doi.org/10.1016/j.eswa.2019.112846

[10] J. Cai, H. Wei, H. Yang and X. Zhao, A Novel Clustering Algorithm based on DPC & PSO, in IEEE Access, doi: 10.1109/ACCESS.2020.2992903.

[11] Y. Li, J. Cai, H. Yang, J. Zhang and X. Zhao, A Novel Algorithm for Initial Cluster Center Selection, in IEEE Access, vol. 7, pp. 74683-74693, 2019, doi: 10.1109/ACCESS.2019.2921320.

[12] Nolan L A, Harva M O, Kabán A, et al. A data-driven Bayesian approach for finding young stellar populations in early-type galaxies from their ultraviolet–optical spectra[J]. Monthly Notices of the Royal Astronomical Society, 2006, 366(1): 321-338.

[13] Kim E J, Brunner R J. Star-galaxy classification using deep convolutional neural networks[J]. Monthly Notices of the Royal Astronomical Society, 2016: stw2672.

[14] Krause M, Pueschel E, Maier G. Improved γ/hadron separation for the detection of faint γ-ray sources using boosted decision trees[J]. Astroparticle Physics, 2017, 89: 1-9.

[15] Cortes C, Vapnik V. Support-vector networks[J]. Machine learning, 1995, 20(3): 273-297.

[16] C. Qu, H. Yang, J. Cai, J. Zhang and Y. Zhou, DoPS: A Double-Peaked Profiles Search Method Based on the RS and SVM, in IEEE Access, vol. 7, pp. 106139-106154, 2019, doi: 10.1109/ACCESS.2019.2927251.

[17] H. Yang, C. Qu, J. Cai, S. Zhang and X. Zhao, SVM-Lattice: A Recognition & Evaluation Frame for Double-peaked Profiles, in IEEE Access, doi: 10.1109/ACCESS.2020.2990801.

[18] Shanmuganathan S. Artificial neural network modelling: An introduction[M]//Artificial Neural Network Modelling. Springer, Cham, 2016: 1-14.

[19] Zhou Z H, Feng J. Deep Forest[J]. arXiv preprint arXiv:1702.08835, 2017.

[20] Ifada N, Nayak R. Do-Rank: DCG optimization for learning-to-rank in tag-based item recommendation systems[C]//Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, Cham, 2015: 510-521.

[21] Zhang S, Ge Y. Personalized Tag Recommendation Based on Transfer Matrix and Collaborative Filtering[J]. Journal of Computer and Communications, 2015, 3(09): 9.

[22] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: Bringing order to the web. Stanford Digital Libraries Working Paper, 1998

[23] Jon Kleinberg. Authoritative sources in a hyperlinked environment. In Proceedings of the ACMSIAM Symposium on Discrete Algorithms, 1998.

[24] Du C, Luo A, Yang H, et al. An Efficient Method for Rare Spectra Retrieval in Astronomical Databases[J]. Publications of the Astronomical Society of the Pacific, 2016, 128(961): 034502.

[25] Ruder S. An overview of gradient descent optimization algorithms[J]. arXiv preprint arXiv:1609.04747, 2016.

[26] Hou W, Luo A L, Hu J Y, et al. A catalog of early-type emission-line stars and Hα line profiles from LAMOST DR2[J]. Research in Astronomy and Astrophysics, 2016, 16(9): 138.

[27] Hou W. Studies of the "Peculiar" A-type Stars from LAMOST Survey[J]. Publications of the Astronomical Society of the Pacific, 2017, 129(974): 047001.

[28] Lin C C C, Hou J L, Chen L, et al. Searching for classical Be stars in LAMOST DR1[J]. Research in Astronomy and Astrophysics, 2015, 15(8): 1325.

[29] M. Karabatak, M. Cevdet, Ince An expert system for detection of breast cancer based on association rules and neural network, Expert Syst. Appl. 36 (2, Part 2) (2009) 3465–3469.

[30] M. Zieba, J.M. Tomczak, M. Lubicz, J. Swiatek, Boosted SVM for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients, Appl. Soft Comput. 14 (Part A) (2014) 99–108.

[31] R.B. Rao, S. Krishnan, R.S. Niculescu, Data mining for improved cardiac care, ACM SIGKDD Explor. Newsl. 8 (1) (2006) 3–10.

[32] J.S. Yoon, Y.S. Kwon, A practical approach to bankruptcy prediction for small businesses: substituting the unavailable financial data for credit card sales information, Expert Syst. Appl. 37 (5) (2010) 3624–3629.

[33] Zelenkov Y, Fedorova E, Chekrizov D. Two-step classification method based on genetic algorithm for bankruptcy forecasting[J]. Expert Systems with Applications, 2017, 88: 393-401.

[34] Aburomman A A, Reaz M B I. A novel SVM-kNN-PSO ensemble method for intrusion detection system[J]. Applied Soft Computing, 2016, 38: 360-372.

[35] Olatunji S O. Improved email spam detection model based on support vector machines[J]. Neural Computing and Applications, 2019, 31(3): 691-699.

[36] Ebenuwa S H, Sharif M S, Alazab M, et al. Variance ranking attributes selection techniques for binary classification problem in imbalance data[J]. IEEE Access, 2019, 7: 24649-24666.

[37] R. Longadge and S. Dongre, Class imbalance problem in data mining: Review, Int. J. Comput. Sci. Netw., vol. 2, no. 1, p. 1707, 2013.

[38] Liu X Y, Wu J, Zhou Z H. Exploratory undersampling for class-imbalance learning[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 2008, 39(2): 539-550.

[39] Drummond C, Holte R C. C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling[C]//Workshop on learning from imbalanced datasets II. Washington, DC: Citeseer, 2003, 11: 1-8.

[40] G. M. Weiss, Mining with rarity: A unifying framework, ACM SIGKDD Explorations, vol. 6, no. 1, pp. 7–19, 2004.

[41] Z.-H. Zhou and X.-Y. Liu, Training cost-sensitive neural networks with methods addressing the class imbalance problem, IEEE Transactions on Knowledge and Data Engineering, vol. 18, no. 1, pp. 63–77, 2006.

[42] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: synthetic minority over-sampling technique[J]. Journal of artificial intelligence research, 2002, 16: 321-357.

[43] Elkan C. The foundations of cost-sensitive learning[C]//International joint conference on artificial intelligence. Lawrence Erlbaum Associates Ltd, 2001, 17(1): 973-978.

[44] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences, 55(1):119–139, 1997

[45] Jain S, White M, Radivojac P. Estimating the class prior and posterior from noisy positives and unlabeled data[C]//Advances in neural information processing systems. 2016: 2693-2701.

[46] Han J, Pei J, Kamber M. Data mining: concepts and techniques[M]. Elsevier, 2011.

[47] Zhao X, Zhang J, Qin X. k NN-DP: Handling Data Skewness in kNN Joins Using MapReduce[J]. IEEE Transactions on Parallel and Distributed Systems, 2017, 29(3): 600-613.

[48] Du C, Luo A, Yang H, et al. An Efficient Method for Rare Spectra Retrieval in Astronomical Databases[J]. Publications of the Astronomical Society of the Pacific, 2016, 128(961): 034502.

[49] Agarwal S, Niyogi P. Stability and generalization of bipartite ranking algorithms[C]//International Conference on Computational Learning Theory. Springer, Berlin, Heidelberg, 2005: 32-47.

[50] Menon A K, Williamson R C. Bipartite ranking: a risk-theoretic perspective[J]. The Journal of Machine Learning Research, 2016, 17(1): 6766-6867.

[51] Faramarzi N S, Ayday E, Guvenir H A. A Privacy-Preserving Solution for the Bipartite Ranking Problem[C]//2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA). IEEE, 2016: 375-380.

[52] Khalid M, Ray I, Chitsaz H. Confidence-weighted bipartite ranking[C]//International Conference on Advanced Data Mining and Applications. Springer, Cham, 2016: 35-49.

[53] Du C, Du C, Long G, et al. Online Bayesian Multiple Kernel Bipartite Ranking[C]//UAI. 2016.

[54] Nekooeimehr I, Lai-Yuen S K. Adaptive semi-unsupervised weighted oversampling (A-SUWO) for imbalanced datasets[J]. Expert Systems with Applications, 2016, 46: 405-416.

[55] Gonzalez-Abril L, Angulo C, Nuñez H, et al. Handling binary classification problems with a priority class by using Support Vector Machines[J]. Applied Soft Computing, 2017, 61: 661-669.

[56] Panja R, Pal N R. MS-SVM: minimally spanned support vector machine[J]. Applied Soft Computing, 2018, 64: 356-365.

[57] Gu S, Jin Y. Multi-train: A semi-supervised heterogeneous ensemble classifier[J]. Neurocomputing, 2017, 249: 202-211.

[58] Susan S, Kumar A. SSOMaj-SMOTE-SSOMin: Three-step intelligent pruning of majority and minority samples for learning from imbalanced datasets[J]. Applied Soft Computing, 2019, 78: 141-149.

[59] Sabzevari M, Martínez-Muñoz G, Suárez A. Vote-boosting ensembles[J]. Pattern Recognition, 2018, 83: 119-133.