INTERNATIONAL JOURNAL OF CIRCUITS, SYSTEMS AND SIGNAL PROCESSING
DOI: 10.46300/9106.2021.15.131 Volume 15, 2021

Algorithm for key classification feature selection of big data based on Henie theorem

Wei Wang*,
Department of Information Engineering, Henan Industry and Trade Vocational College,
Zhengzhou, 451191
China

Received: March 15, 2021. Revised: August 3, 2021. Accepted: August 28, 2021. Published: August 31, 2021.

Abstract—**With the extensive application of the database system, the available data of enterprises or individuals are expanding, and the existing technology is difficult to meet the data analysis requirements of the big data age. Therefore, the selection of key classification features of big data needs to be carried out. However, when the key classification features of big data are selected by the current algorithm, the distance between the samples can not be given accurately, and there is a large error in the classification. To solve this problem, a key classification feature selection algorithm based on Henie theorem is proposed. In this algorithm, the second programming algorithm is firstly used to make the weighted distance between the intra-class and the inter-class as the quadratic term and linear term parameter in the target function, and balance the relationship between the data features and the different categories. The optimized vector is used as the weight vector to measure the contribution of the feature to the classification. According to the feature importance degree, the redundancy feature is gradually deleted, and the problem of selecting the key classification features of big data into the resolution principle is fused into the Henie theorem. The function limit and sequence limit of the key classification features of big data are obtained. Based on this, the key classification features of big data are selected. Experimental simulation shows that the proposed algorithm has higher classification accuracy and can effectively meet the needs of data analysis in the era of big data.**

Keywords—Big data, feature selection, Henie theorem, key classification.

I. Introduction

WITH the rapid development of computer measurement technology, big data has also entered a new era [1] [2]. The concept of big data originated in 1980s. The big data special publication created by Nature in 2008 has made big data the focus of research and application in recent years, and the potential application prospects of big data have been paid much attention by the governments of many developed countries. For example, in 2012, the US government announced the $200 million “Big data research and development plan”. In September 2015, China's State Council issued the “Action plan for promoting the development of big data”. In the meantime, at least 50 countries have issued corresponding documents to support the development of big data. Although the application value and Prospect of big data have been universally recognized, as a new thing, the technology of big data is still imperfect, especially for the selection of key classification features of data, it still cannot accurately give the distance between the data features in the class sample, and make the implementation of selecting the key classification features of big data to fall into a bottleneck. In this case, the selection of key classification features of big data has become an important factor restricting the development of big data fields, which has attracted a lot of experts and scholars to pay attention to [3].

At present, there are many researches on the selection of key classification features of big data, and some results have also been achieved. Based on the generalized theory of information, each feature in the feature set of big data is independently assessed. From the high-dimensional feature space, the effective features of data classification are selected, to complete the selection of key classification features of big data [4]. The efficiency of the algorithm is high, but when the key classification features of big data are selected by the current algorithm, the distance between the samples can not be accurately given, and there is a large error in the classification. The support vector machine is used as a classifier, and the key classification features of big data are set up to select the optimal classification surface, so that the key classification features of the data are selected [5]. The algorithm has high stability in classification feature selection, but the selection process is

rather cumbersome and time-consuming. The rough set feature selection algorithm for neighborhood is firstly applied to select the key classification features of big data [6]. The redundant features are removed step by step according to the importance degree of the feature [7]. The algorithm has high accuracy in feature selection, but the algorithm has large limitations.

In view of the above problems, an algorithm for key classification feature selection based on Henie theorem is proposed. Experimental simulation shows that the proposed algorithm has higher classification accuracy and can effectively meet the needs of data analysis in the era of big data.

II. Algorithm For Key Classification Feature Selection Of Big Data Based On Henie Theorem

A. Calculation of weighted distance between data samples in the same category

Feature selection is a pre-processing process before sample classification, and is very important for classification of big data samples. In order to improve the precision of the selecting key classification features of big data, in the process of selecting the key classification features of big data, the second planning theory is used to define the target function and constraint conditions of the key classification characteristics of big data based on the second planning, to obtain the weight of the data features, and to balance the similarity degree between the features [8]. The detailed steps are as follows:

Assuming that H represents the symmetric matrix and A represents the constraint matrix. For data sets $D_{M\times N}$, 4 M represents the number of characteristics of data classification, and N represents the number of samples, then the objective functions and constraints based on the second planning are defined by (1) and (2).

$$x'_{zcj}=\frac{N\otimes D_{M\times N}}{\{M\mp N\}}M''\otimes\frac{H\mp A}{Q\in R}\times\frac{Q(i,j)}{F} \tag{1}$$

$$e'_{sdg}=\frac{x'_{yup}\mp e'_{opk}}{C'_{e'_{zvj}}}\pm x'_{zcj} \tag{2}$$

where, $Q\in R$ represents the symmetric positive definite matrix, $Q(i,j)$ represents the degree of similarity between i th features and j th features in the data set $D_{M\times N}$. F represents the M'' dimensions vector, and F expresses the correlation between the feature of the data set $D_{M\times N}$ and the sample category. x'_{yup} represents the data sample category information, e'_{opk} represents the minimum value of the row domain of the target function x'_{zcj}, and the correlation threshold between features of the sample category information is represented by $C'_{e'_{zvj}}$.

Assuming that v'_{iop} represents the attribute space of the weight vector of the feature, the weight vector x'_{opk} of the data feature is obtained by using (3):

$$x'_{opk}=v'_{iop}\frac{\phi'_{juo}\mp E'_{asxjj}}{b'_{gjkl}\oplus\Phi(d)}\oplus\frac{y_d\oplus x'_{zcj}}{D_{M\times N}} \tag{3}$$

where, E'_{asxjj} represents the Gauss kernel function, ϕ'_{juo} represents the element value in the weight vector x'_{opk}, b'_{gjkl} represents a data sample in the data set $D_{M\times N}$, $\Phi(d)$ represents the distance between the same samples, and y_d represents the category label of the sample b'_{gjkl}.

Each element value in the weight vector x'_{opk} is expressed as the characteristic weight of each data. Its weight value reflects the similarity between the feature and other features and its correlation with the sample category. The higher the weight value is, the more important the sample classification is [9]. The distance η'_{ghj} between the similar samples and the distance S'_{erj} between different samples are respectively as the quadratic term and linear term parameter of the target function by the second planning algorithm, and the second planning problem of the key data classification features is defined by using (4).

$$W'_{wsxj}=\frac{S'_{erj}\oplus x'_{opk}}{\eta'_{ghj}}\pm\frac{\{e'_{afh}\mp d'_{dhk}\}}{f'_{fjk}}e'_{sdg} \tag{4}$$

where, e'_{afh} represents the distribution dispersion in the heterogeneous samples, d'_{dhk} represents the inter-class dispersion, and the f'_{fjk} represents the intra-class tightness.

Assuming that $Dist_k(i,j)$ represents the distance between i th features and j th features in k th types of samples, the weighted distance between the similar samples can be calculated by using (5).

$$H(i,j)=\frac{\mu_k(i)\oplus Dist_k(i,j)}{W'_{wsxj}}\otimes\frac{\sigma_k(i)}{|\mu_k(i)-\mu_k(j)|}\times(card(D)) \tag{5}$$

where $\mu_k(i)$ represents the mean of i th feature in the k th type of samples, and $\sigma_k(i)$ represents the standard deviation in the i th features of the k th intra-class sample, $|\mu_k(i)-\mu_k(j)|$ represents the intra class space of the two features, and $(card(D))$ represents the total number of data sets.

Assuming that O'_{upo} represents the total number of data set samples, ∂'_{rtp} represents the mean difference of the normalized feature, and $card(D_k)$ represents the number of k th types of samples in the data set $D_{M\times N}$, and the weighted distance between the heterogeneous samples is calculated using (6).

$$e'_{rtu}=\frac{O'_{upo}\mp card(D_k)}{D_{M\times N}\mp\partial'_{rtp}}\mp\{\partial''_S\mp E'_{UIP}\}\times H(i,j) \tag{6}$$

where, ∂'_{rtp} represents the proportion of k th types of samples in the data set, E'_{UIP} represents the distance between the i th feature in k th types of samples and other categories of samples.

and ∂_S'' represents the distance error between the k th sample and the other categories.

To sum up, in the process of selecting the key classification features of big data, the second programming theory is used to define the objective function and constraint conditions of the key classification features of big data based on the second planning, to obtain the weight vector of the feature, give the similarity degree between the features, and the relation between the homogeneous and the heterogeneous [10]. It lays a foundation for optimizing and selecting key classification features of big data.

B. The calculation of the weight vector of the features to the contribution of the classification

Assuming that μ_{jk}' represents the positive definiteness of the second programming for the H matrix, ∂_{oiu}' represents the diagonal elements of the H matrix, and the utilization (7) is used to normalize the quadratic term and the linear term.

$$E_{sxcv}' = \frac{\partial_{oiu}' \oplus \mu_{jk}'}{H \mp E_{et}'} \mp \frac{\{c_{tyu}'\}}{\mu_{upo}'} \otimes e_{rtu}' \tag{7}$$

where E_{et}' represents the weight coefficient of E_{sxcv}' on each sample, c_{tyu}' represents the set of samples in the original feature space, and μ_{upo}' represents the feature importance of the high-dimensional dataset.

The value of the best solution vector element reflects the closeness between the corresponding features in the intra-class and other features, and the distance relationship between the features among the heterogeneous samples. Assuming that σ_{wep}' and f_{wep}' represent the degree of tightness within the intra-class and the degree of dispersion in the inter class, and s_{wer}' represent the data blocks within the historical window. Then in (8), E_{sxcv}' is as a weighting vector to measure the contribution of the features to the classification.

$$E_{asj}' = \frac{\sigma_{wep}' \pm f_{wep}'}{s_{wer}' \mp E_{sxcv}'} \pm l_{yup}' \tag{8}$$

where l_{yup}' represents a threshold for normalization of feature spacing.

The normalization of the quadratic term and linear item parameter can promote the better feature weight after the optimization to show the role of the feature in the similar and heterogeneous samples, and can obtain the classification characteristics with rich category information [11].

To sum up, we can explain that in the process of selecting the key classification features of big data, the quadratic term and linear item are normalized, the optimized solution vector is used as the weight vector to measure the classification contribution, which lays the foundation for the optimization and selection of the key classification features of big data.

C. Deletion of key classification's redundant features of big data

In order to improve the ability of data classification in the selection of the key classification features of big data, the neighborhood feature selection algorithm is used to divide the key classification decision features of the data by the equivalence relation, and the neighborhood decision table is set up to obtain the neighborhood dependence of the decision features to the condition features, and the process of the neighborhood feature selection is given. According to the importance of features, the redundant features are gradually removed [12]. The detailed steps are as follows:

It is assumed that E_{eu}' represents the non-empty finite set, δ represents the range of the feature a, and ∂_{ui}' represents the continuous data. Then, the neighborhood feature selection algorithm is used to divide the data key classification decision feature set by the equivalence relation.

$$Q_{sdp}' = \frac{\partial_{ui}' \mp E_{eu}'}{\delta \oplus (a)} \oplus \frac{\{\partial_{ui}' \otimes p_{dgj}'\}}{d_{wr}'} \tag{9}$$

where, p_{dgj}' represents the vector space of the data classification decision feature, and d_{wr}' represents the maximum weight of the data classification decision feature.

By σ_{rtp}'' representing the approximate set of the classified feature neighborhood of the data set B_{erj}', and v_{sgh}' representing the approximate set of the classified feature neighborhood of the data set B_{erj}' .The neighborhood decision table u_{ey}' can be obtained by using (10).

$$u_{ey}' = \frac{\sigma_{rtp}'' \pm B_{erj}'}{v_{sgh}'} \oplus \sigma_{sgj}' \mp s_{sgh}' \tag{10}$$

where, σ_{sgj}' represents the maximum constraint scope of the conditional feature, and s_{sgh}' represents the number of class identities of the data set.

It is assumed that χ_{yup}' represents the spatial vector of arbitrary data classification decision making characteristics, ∂_{jwer}' is the attribute of the condition feature, then based on (11), the neighborhood dependency of decision features on conditional features is to obtain.

$$S_{edgk}' = \frac{\partial_{jwer}' \oplus \chi_{yup}'}{D_{ert}' \oplus v_{fhk}'} \mp \frac{\{f_{fjl}' \times b_{fhk}'\}}{u_{ey}' \times Q_{sdp}'} \tag{11}$$

where, D_{ert}' represents the neighborhood reduction function, v_{fhk}' represents the selected key feature set, and b_{fhk}' represents the minimum feature of importance degree.

To sum up, we can explain that in the process of selecting the key classification features of big data, the second programming theory is used to define the objective function and constraint conditions based on the second planning, to obtain the weight vector of the feature, to give the similarity between the features, and to balance the relationship between the similar and the heterogeneous [13]. It lays the foundation for selecting the key

classification features of big data.

The selection steps of key classification features of big data based on Henie theorem

Henie theorem is a bridge between the limit of data function and the limit of sequence. Henie theorem is given by German mathematician Henie. Using the Henie theorem, the key classification features of the data can be selected to choose the number of series of functions. Therefore, it can also be called the principle of resolution. In the selection process of key classification features of big data, the Henie theorem is introduced to the selection process of key classification characteristics of big data. The function limit of key classification features of data is transformed into the problem of resolution principle, and the function limit and number limit of the key classification features of big data are obtained. The detailed steps are as follows:

Assuming that $\lim f(x)$ is a proof function that represents an arbitrary column and $\{x_n\}$ represents the limit problem of a function. Using (12), the function limit problem of selecting the key classification features of data is transformed into the principle of resolution.

$$E'_{pkjh}\frac{\{x_n\}\oplus\lim f(x)}{o'_{opi}}\mp k''_{plo} \tag{12}$$

where o'_{opi} represents the nature of the sequence limit of data's key classification features, and k''_{plo} represents the necessary condition for the Henie theorem.

It is assumed that the μ'_{oiu} represents the two side clips of the series, ξ'_{iuk} represents the limit of the function of the data classification features, and σ'_{opk} represents the number of an arbitrary monotonous increase. Then, the function limit ξ'_{iuk} and the sequence limit ∂'_{dfu} of the key classification features of the big data are calculated by using (13) and (14).

$$\xi'_{iuk}=\frac{\mu'_{oiu}\oplus\left\{d'_{jkl}\oplus b'_{hjk}\right\}}{\sigma'_{opk}}\oplus E'_{pkjh} \tag{13}$$

$$\partial'_{dfu}=\xi'_{iuk}\frac{\mu'_{oiu}\oplus\left\{d'_{jkl}\oplus b'_{hjk}\right\}}{\sigma'_{opk}}\oplus E'_{pkjh} \tag{14}$$

where d'_{jkl} represents the difference limit of the functions, and b'_{hjk} represents the set of arbitrary monotonic increasing numbers.

The Henie theorem can reveal the intrinsic relationship between the discrete variation and the continuous change of the data classification features, assuming that η'_{eru} represents the Cauchy convergence criterion, the key classification features of the big data are selected by (15).

$$x'_{opl}=\frac{\partial'_{dfu}\pm\xi'_{iuk}}{E'_{pkjh}}\otimes\eta'_{eru} \tag{15}$$

III. Simulation Results

In order to prove the validity of the proposed algorithm based on Henie theorem for key classification feature selection of big data, an experiment is needed. In the MATLAB simulation environment, we set up the experimental simulation platform for the key classification features of big data. The classification features of 6 data sets are selected, of which 4 data sets are derived from the data set AcuteLeuKe-mia, Multiple myeloma, Colon and DLBCL in the University of California at Irvine, and the other two datasets are derived from the dataset Ionosphere and Promter. The basic information of the dataset is shown in Table I.

Table I. Structural information of experimental data sets

Dataset	Number of features	Number of samples	Number of training set
DLBCL	7 139	265	211
Colon	7 139	107	64
AcuteLeu Ke- mia	2 010	73	43
Multiple myeloma	36	62	63
Ionosphere	58	77	38
promter	1 500	89	37

A. Setting of evaluation index

In the experiment, in order to better verify the feasibility of selecting the key classification features of big data based on the Henie theorem algorithm, the experiment is divided into two stages. In the first stage of the experiment, we use the false alarm rate and the leakage rate as the evaluation index to define the performance of the algorithm for the key classification feature selection of big data. In the second stage of the experiment, in order to show the comprehensiveness and impartiality of the experiment, the algorithm in the study is used as the contrast algorithm to analyze and compare, and the quality of selecting the key classification features of the big data is verified by the accuracy of the data classification [5].

B. Test of false alarm rate and leakage rate by the proposed algorithm

Based on the traditional algorithm and the proposed algorithm based on Henie's theorem, we want to compare the key classification feature selection experiments of big data [5]. The proposed algorithm is used to test the false positive rate and false negative rate of key data classification for big data. Using the second programming theory to classify the target functions and constraints of features, the weight of data features is obtained. According to the big data correlation threshold, the weight vector of features is calculated, and the quadratic term and linear term are normalized. The results are shown in Figs. 1 and 2.

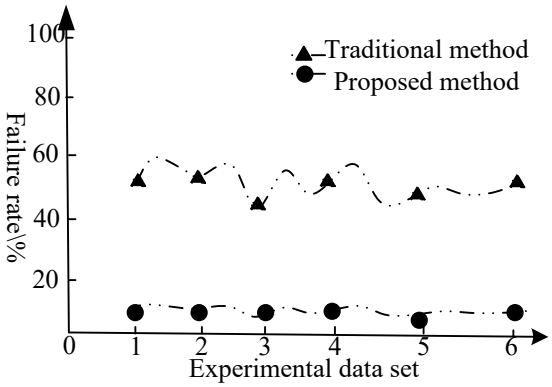

Fig. 1 the leakage rate of different algorithms

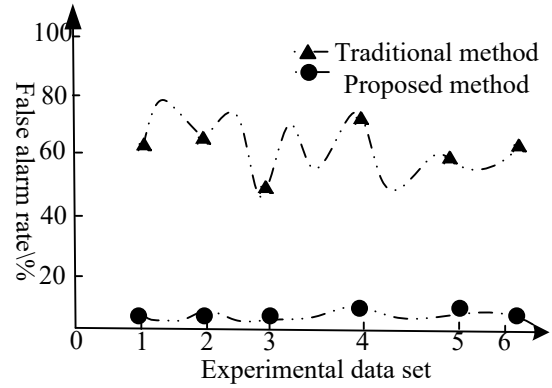

Fig. 2 false alarm rate of different algorithms

It can be seen from the experimental simulation results in Fig. 1 and Fig. 2 that the leakage rate of the traditional method is higher than that of the design method in this paper. This method can balance the relationship between the same class and heterogeneous classes by normalizing the quadratic term and linear term. On this basis, the feature set conducive to classification is selected according to the feature weight to ensure the selection quality of key classification features.

C. Comparison of data classification accuracy between different algorithms

This paper uses the proposed algorithm and the algorithm in the study to select the key classification features of big data, and compares the classification accuracy of the key classification features of big data by different algorithms [5]. The comparison results are shown in Fig. 3.

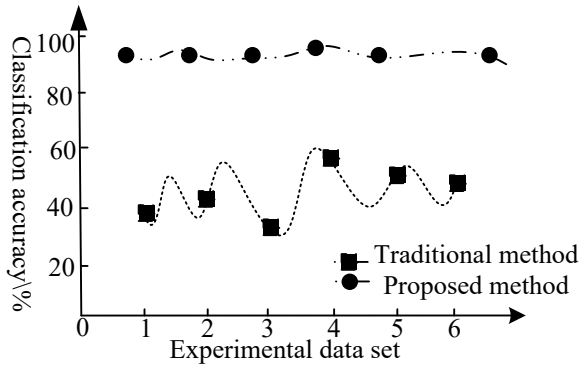

Fig. 3 comparison of classification accuracy of different algorithms

According to the simulation experiment results in Fig. 3, it can be seen that the data classification accuracy of the algorithm in the study fluctuates between 30% and 60%, and the accuracy is low [5]; The accuracy of the proposed method is always above 90%, and the highest is 95%.

The simulation results show that the key classification feature selection algorithm based on Heine theorem proposed in this paper can effectively measure the weight vector of classification contribution to features. The classification accuracy of key classification features of big data is higher than that of the previous methods, and the practical application results are better, and can effectively meet the needs of data analysis in the era of big data. This is because the design method uses the value of the optimal solution vector element to obtain the chromatic divergence of the feature compactness, and uses the weighted vector to measure the contribution of the feature to the classification. The neighborhood decision table is established, the dependence of decision features and conditional features is obtained, the redundant features are deleted, the functional limit and sequence limit of key classification features of big data are calculated by using Heine theorem, and the selection of key classification features is completed to meet the analysis of data.

IV. Conclusion

In view of the key problems that need to be solved in the selection of key classification features of big data, an algorithm for key classification feature selection based on Henie theorem is proposed. Firstly, the distance of the data features between the intra-class and the dissimilar samples is used as the quadratic term and linear item parameters of the target function by the second planning algorithm, and on this basis, the classification features in the intra-class and inter-class are searched. And the normalization of the quadratic term and linear item is used to balance the relationship between the same sample and heterogeneous samples. The optimal solution vector is used as the weight vector to measure the classification contribution, and the Henie theorem is introduced to the selection of key classification characteristics of big data. The internal relation between the discrete change of the data classification characteristic variables and the continuous change is revealed, and the accuracy of the selection of the key classification features is improved. Experimental simulation shows that the proposed algorithm has higher classification accuracy and can effectively meet the needs of data analysis in the era of big data.

The method designed in this paper can be applied to other studies that need to carry out data feature classification. Combined with this method to process data, analyze data and effectively apply data, it can better solve the problems of science and technology, business and informatization, and improve the overall level of information data processing technology in China. Due to the limited time, this paper only studies the classification of data features. In the next research, we are going to focus on the privacy and security of data, and add a guarantee for the information and data security under the condition of meeting the big data feature analysis.

ACKNOWLEDGMENT

This research is funded by the science and technology key project of Henan province of China (No.212102310085) and the key research project of colleges and universities in Henan province of China (No. 21B520004).

References

[1] Y. Zhao, G. Wang, and Y. Yin, "Improving ELM-based microarray data classification by diversified sequence features selection," Neural Comput Appl, vol. 27, no. 1, pp. 155-166, 2016.

[2] R. Blomley, B. Jutzi, and M. Weinmann, "Classification of airborne laser scanning data using geometric multi-scale features and different neighbourhood types," Isprs Annals of Photogrammetry Remote Sensing & Spatial Informa, vol. 3, no. 3, pp. 169-176, 2016.

[3] C. M. Gevaert, C. Persello, and R. Sliuzas, "Informal settlement classification using point-cloud and image-based features from UAV data," Isprs J Photogramm, vol. 125, pp. 225-236, 2017.

[4] A. Mishra, K. Dey, and P. Bhattacharyya, "Learning cognitive features from gaze data for sentiment and sarcasm classification using convolutional neural network," Meeting of the Association for Computational Linguistics, pp. 377-387, 2017.

[5] V. G. Astafurov and A. V. Skorokhodov, "Formation of a set of informative classification features for solving cloud classification problem using MODIS satellite data," Tr Spiiran, vol. 4, no. 53, pp. 118-139. 2017.

[6] S. Lin, G. D. Guo, and F. Huang, "Quantum anonymous ranking based on the Chinese remainder theorem," Physical Review A, vol. 93, no. 1, 2016.

[7] A. Molavi, A. Jalali, and N. M. Ghasemi, "Adaptive fuzzy control of a class of nonaffine nonlinear system with input saturation based on passivity theorem," Isa T, vol. 69, pp. 202-213, 2017.

[8] T. Santhi Vandanna, S. Venkateshwarlu, and K. Viswanath, "Robust and highly secure technique for wireless body sensor network using sequence of ECG data," WSEAS Transactions on Information Science and Applications, ISSN / E-ISSN: 1790-0832/2224-3402, vol. 17, pp. 138-145, 2020.

[9] D. Oreški and G. Hajdin, "Development and comparison of predictive models based on learning management system data," WSEAS Transactions on Information Science and Applications, vol. 22, no. 16, pp. 192-201, 2019.

[10] Daoud, "Data acquisition system for photovoltaic maximum power point tracking," WSEAS Transactions on Information Science and Applications, vol. 16, pp. 129-139, 2019.

[11] A. Chaleplioglou, S. Papavlasopoulos, and M. Poulos, "Minimisation of terms to describe a knowledge domain for ontology engineering and linked data generation," WSEAS Transactions on Information Science and Applications, vol. 16, no. 7, pp. 64-68, 2019.

[12] J. Y. Tan and X. Y. Zhou, "Hidden encryption simulation of big data features based on information entropy suppression," Computer Simulation, vol. 37, no. 4, pp. 192-196, 2020.

[13] J. Li, S. S. Lin, and F. Chen, "Analysis of the substation area industry clustering methods and classification characteristics based on the big data," Power Systems and Big Data, vol. 23, no. 3, pp. 1-9, 2020.

Wei Wang, female, born in August 1982, and she is an associate professor. She had got the B.S degree in 2006, from Chongqing Communication Institute, majoring in computer science and technology. Now, she is working in Henan Industry and Trade Vocational College. And her research areas include: data mining, machine learning and deep learning. She has published 23 academic papers. Meanwhile, she has hosted 3 research projects and participated 8 research projects.

Author Contribution:

In order to reduce the error of data feature classification algorithm, Wei Wang proposed a key classification feature selection algorithm based on Heine's theorem. The weighted distance is used as the objective linear function to optimize the vector relationship. Integrate and classify the characteristics of big data by using Heine's theorem. Finally, the experimental test can prove that the algorithm proposed by the author has high practicability, the accuracy is as high as 99%, and the false positive rate is low. It has a wide application prospect.