ISSN 1330-3651 (Print), ISSN 1848-6339 (Online)
https://doi.org/10.17559/TV-20191111095139
Original scientific paper

# Unsupervised Text Topic-Related Gene Extraction for Large Unbalanced Datasets

Li JING-MING, Sun JING-TAO, Huang WEN-HAN, Zhang QIU-YU, Tian ZHEN-ZHOU, Lu NING

**Abstract:** There is a common notion that traditional unsupervised feature extraction algorithms follow the assumption that the distribution of the different clusters in a dataset is balanced. However, feature selection is guided by the calculation of similarities among features when topic keywords are extracted from a large number of unmarked, unbalanced text datasets. As a result, the selected features cannot truly reflect the information of the original data set, which thus affects the subsequent performance of classifiers. To solve this problem, a new method of extracting unsupervised text topic-related genes is proposed in this paper. Firstly, a sample cluster group is obtained by factor analysis and a density peak algorithm, based on which the dataset is marked. Then, considering the influence of the unbalanced distribution of sample clusters on feature selection, the CHI statistical matrix feature selection method, which combines average local density and information entropy together, is used to strengthen the features of low-density small-sample clusters. Finally, a related gene extraction method based on the exploration of high-order relevance in multidimensional statistical data is described, which uses independent component analysis to enhance the generalisability of the selected features. In this way, unsupervised text topic-related genes can be extracted from large unbalanced datasets. The results of experiments suggest that the proposed method of extracting unsupervised text topic-related genes is better than existing methods in extracting text subject terms from low-density small-sample clusters, and has higher prematurity and feature dimension-reduction ability.

**Keywords:** CHI statistical selection method; density peaks; factor analysis; information entropy; independent component analysis; text feature gene extraction

## 1 INTRODUCTION

As society gradually enters the age of "big data" [1], increasing amounts of information are available from webpages, microblogs, forums, and multimedia files, etc. [2]. Meanwhile, the time available to read and process information is decreasing, so efficient and accurate information analysis is becoming an effective means of understanding large datasets and discovering value. Such analysis is applicable to public opinion monitoring and early warning on the internet, such as filtration of harmful information from networks, emotion analysis, personalized recommendations for products[3], etc. Moreover, during data processing, it is generally necessary to process a lot of data that is redundant or has uncorrelated features with causing the efficiency of learning algorithms significantly. This can be a fatal link in machine learning and data mining, as feature extraction has direct impacts on model building, analysis efficiency and accuracy.

At present, feature extraction may be classified as *supervised* and *unsupervised* [4]. In text content analysis processes, regardless of the class, a vector space model [5] is required to express the text in a vector space consisting of a certain quantity of feature words. This causes two inevitable issues in practical applications: (1) the distribution of sample categories (clusters) in the dataset is not balanced. Various measures have been used for feature subset quality evaluation, including independent correlation analysis [6], similarity analysis [7], distance-based Euclidean distance, Mahalanobis distance [8], and the most widely-used technique: information entropy-based mutual information and information gain [9]. All of these techniques identify similarities between the sample categories (clusters) in the dataset. Assuming that most of the identified features come from the "big class", which contains most of the categories (clusters), and none or very few features come from the "small class", the most distinguishing features can be selected. Subsets cannot accurately reflect the information of the entire sample space, which reduces the ability of subsequent learning methods to solve practical problems. (2) The subject to be processed becomes more complicated and the data dimensions increase rapidly due to the very large size of some datasets. Analyses of ultrahigh-dimension datasets have high memory and computational requirements [10]. In spaces with high-dimensional features, various feature points have strong dependency, which causes high redundancy and even noise. Hence, the ability to generalise the features of traditional methods deteriorates sharply, and "empty space" is caused in highly-dimensional data space, making it difficult to solve multi-element density estimation problems. It is increasingly important to extract the substantive characteristics of things from complicated information; i.e., to determine out mutual independence and potentially hidden information, remove high-order redundancy, extract the genetic data of complete and independent subjects, and improve feature generalisability.

In order to overcome the defect that the traditional feature extraction method cannot truly reflect the information of the original data set under the imbalanced data set, this paper proposes an unsupervised text topic-related gene extraction method (UTTGE). The contributions of this paper are listed as follows:

- The UTTGE method combines factor analysis method with density peak algorithm. The factor analysis method is used to find the optimal low-dimensional base describing the original high-dimensional vector space, which makes it possible for the density peak algorithm [11] to quickly find the sample clusters of large-scale data sets.
- The UTTGE method introduces average local density and information entropy [12] into the definition of feature item weighting, so as to construct the feature item's discrimination matrix for sample categories (clusters), which can eliminate the defects existing in the feature selection for uneven sample sets by the traditional method.
- The UTTGE method uses Independent Component Analysis (ICA) [13] for the topic mining tasks of unsupervised texts. By analyzing the high-order dependence between multi-dimensional statistics, it finds hidden information components which are mutually independent, and accurately selects the optimal feature subset comprehensively and truly

reflecting texts' topic information in imbalance large-scale data sets. In this way, the classification and recognition of texts are improved.

The remainder of this paper is organised as follows. Section 2 reviews relevant work by Chinese and other scholars. Section 3 proposes a compatible, new, unsupervised, text topic-related gene extraction method based on an unsupervised clustering method for unbalanced datasets and text topic-related genes. Section 4 provides the experimental results and compares the new method's performance with other similar methods. Finally, we conclude our paper in Section 5.

## 2 RELATED WORK

At present, Chinese and other scholars have performed some research on the analysis of unbalanced datasets. Such issues are mainly solved in two ways: 1) optimisation of the existing feature descending dimension method and 2) improvement of sample class distribution rebalance and sorting algorithms. The core idea of class distribution rebalance is data resampling. The more common resampling techniques include oversampling and undersampling. Chawla et al. [14] provided the synthetic minority over-sampling technique (SMOTE) and improved the generalisability of the oversampling method by artificial synthesis of small classes. However, this method requires a high sample training time and increases the possibility of sample redundancy. Chen et al. [15] provided a step-by-step optimisation-based anti-random undersampling algorithm. This algorithm can remove noise and repetition information from training samples and make the classifier more suitable for small samples. In addition, improvement of the classification algorithm is not based on changing the class distribution of the original unbalanced dataset but on identifying small class samples by making the classifier more sensitive to them. Fang et al. [16] provided a method for detecting internet spam using the SMOTE oversampling method together with the random forest classification algorithm. Li et al. [17] provided an improved kernel density estimation-based data classification algorithm, and the space information of the method is still defined as the distance between the detection point and the class-centre, which inevitably reduces this method's robustness. Except for these methods, many studies have provided improvements to the classification algorithm, e.g. boosting [18], FCM-KFDA [19], AdaBoost-SVM [20], etc. The feature subsets selected in these methods are more optimised, but these methods generally have low efficiency for large, highly-dimensional datasets.

For treatment of the unbalanced issue, there is little research on the first aspect of feature dimension reduction. However, it is an effective method of solving unbalanced issues and provides powerful support for solving a series of issues arising from highly-dimensional data. In the traditional feature selection method [21] was classified into unilateral and bilateral methods. The positive class features (sample of feature words belonging to a certain class), and combined positive and negative features (sample of feature words not belonging to a certain class) are selected using unilateral and bilateral methods, and frame combination is established according to feature selection effectiveness to obtain an optimised feature subset. However, this method still relies on traditional feature selection methods and is unsatisfactory for the selection of features in unbalanced datasets. Khoshgoftaar et al. [22] provided an iterative feature selection model to select optimal feature subsets, in which the data features are ranked by the clustering results obtained by an iterative process. However, in the model, selection of iterative functions and the number of iterations have large impacts on problem solving, and the performance of the model is limited to a certain extent. Through an unsupervised feature dimension-reduction model, this paper intends to minimise the information loss that occurs during dimension reduction and present a data subset that is closer to the original data. Current mainstream methods include PCA (principal Component Analysis), mutual information-based methods, MDS, ISOMAP, and manifold-based methods (LLE, LE, LPP, NPE, etc.). Lin et al. [23] provided a direct, unsupervised, orthometric locality-preserving algorithm. This algorithm resolves the matrix using a Laplacian matrix and may directly extract a projection matrix from the original space of the high-dimensional sample to solve the issue of small samples in the unsupervised identification analysis algorithm. Xu et al. [24] provided a mutual, information-based, unsupervised, feature selection method (UFS-MI). In this method, standard UmRMR is selected after comprehensive consideration of relevancy and redundancy features to evaluate the feature importance. Zhu et al. [25] provided an unsupervised feature selection model for regularised self-representation (RSR), in which each feature may be represented as a linear combination of relevant features in a low-dimension space, and the l2 norm is regularised to select the representative feature and ensure its robustness. Li et al. [26] provided a strongly-robust unsupervised feature selection algorithm (RUFS), which uses the $l_{2,1}$ norm minimization method to deal with redundancy and noise in tag learning and feature selection. This method provides an unsupervised, unbalanced, dataset feature selection method. In unsupervised environments, according to changes in the cluster size and using the same features of different clusters, this method assigns weights according to a feature importance function to adjust the unbalanced nature of the data distribution. Alibeigi et al. [27] provided an unsupervised feature selection method for unbalanced datasets. In unsupervised environments, the probability density of different feature spaces is used to analyse the distribution of each data feature. The data distribution relationship is used for feature selection. However, this method does not take into account the characteristics of the data distribution, which have a great impact on classification performance.

## 3 THE PROPOSED METHOD

### 3.1 Basic Framework

In unsupervised environments, to extract the feature information from unbalanced datasets, it is necessary to determine how to: 1) build models for unlabelled high-dimensional data; 2) effectively measure feature similarity; 3) reduce feature dimensions and effectively reduce redundancy and 4) ensure rapid acquisition of the optimal feature subset.

In this paper, consideration is first given to the solution to the problems of valid dimension, and dimension when density peak clustering is performed for unmarked high-dimensional data. Dimension reduction is performed for high-dimensional vectors in the factor analysis method. The clustering algorithm is indicated for density peak by the neighbourhood similarity of the sample point to achieve the clustering and automatic marking of the unmarked text set. Then, a weight is introduced to improve the calculation of the existing $\chi^2$ statistical magnitude, and a CHI statistical matrix is constructed for the feature and sample classes (clusters), and a low-dimensional embedded space is built on the basis of maintaining the amount of original feature information. Finally, the topic gene is extracted in the dependent component analysis method. Fig. 1 shows the framework of the unsupervised text topic-related gene extraction method (UTTGE).

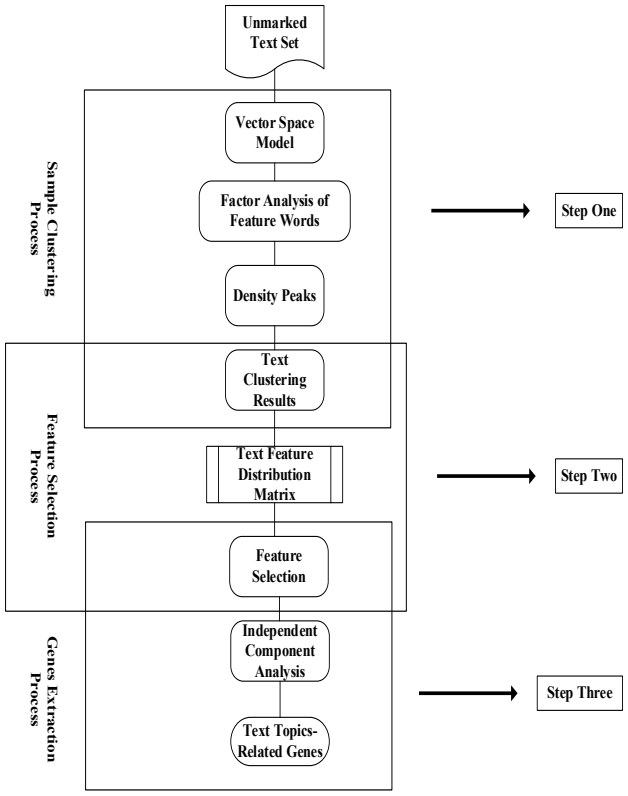



**Figure 1** Framework of unsupervised text topic-related gene extraction method

The following provides details on the three main steps of the UTTGE method.

### 3.2 Density Peak-Based Text Clustering Method

Generally, much informational content does not provide effective labelling due to poor processing and arrangement. However, if one needs to perform exploratory classification and marking for such information in an unsupervised sample, clustering—an unguided learning method—should be used. When the sample size is very high in an actual clustering process, the computational load is likely to exceed the capacity of the computer. Therefore, before clustering, it is necessary to perform dimension reduction for a certain class of variables in the sample. In this paper, we first analyze the characteristic variables of the sample using factor analysis, and then use the fast search and discovery density peak algorithm to cluster the samples according to the obtained factors.

#### 3.2.1 Factor Analysis of Sample Features

Suppose the sample set $X$ includes $n$ samples, $x_1$, $x_2$, …, $x_n$. Each sample $x_i$ consists of $m$ feature indexes, and is recorded as $X = (x_{ij})_{n\times m} = (X_1, X_2, \ldots, X_m)$.

(1) Before the factor analysis, the degree of correlation of $X_1$, $X_2$, …, $X_m$ is judged by the Kaiser Meyer Olkin method (KMO)[28] to determine whether factor analysis is necessary. The value of KMO ranges within (0,1). The closer the KMO value is to 0, the weaker the correlation of $X_1$, $X_2$, …, $X_m$; the closer it is to 1, the stronger the correlation. Generally, it is considered that when the value of KMO is > 0.5, the factor analysis is of actual significance.

(2) The covariance matrix $\boldsymbol{\Sigma} = (h_{ij})_{m\times m}$ is calculated for $X_1$, $X_2$,…, $X_m$. The characteristic root $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ may be obtained for the covariance matrix by the characteristic equation $|\boldsymbol{\Sigma} - \lambda I| = 0$ of $\boldsymbol{\Sigma}$, and corresponding unit feature vector is $T_1$, $T_2$, …, $T_p$.

(3) According to the solution principle for the actual problem, the first $u$ characteristic roots and feature vectors are taken. The sum of their characteristic roots is made to be > 85% of the sum of all characteristic roots to determine the quantity of public factors.

(4) The factor loading matrix

$$\boldsymbol{A} = (T_1\sqrt{\lambda_1}, T_2\sqrt{\lambda_2}, \cdots, T_m\sqrt{\lambda_m}) = \begin{pmatrix} a_{11} & \cdots & a_{1u} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mu} \end{pmatrix}$$

is calculated by the characteristic root and feature vector of $\boldsymbol{\Sigma}$. If the load of each factor has no significant difference in the different feature indexes, the factor loading matrix must be rotated. The factor loading matrix is generally rotated by varimax rotation to obtain $\boldsymbol{A'} = \begin{pmatrix} b_{11} & \cdots & b_{1u} \\ \vdots & \ddots & \vdots \\ b_{m1} & \cdots & b_{mu} \end{pmatrix}$.

(5) $b_{ip} = \mathrm{Max}\{b_{i1}, b_{i2}, \ldots, b_{iu}\}$, $i = 1, 2, \ldots, m$, $p \in \{1, 2, \ldots, u\}$ is operated by the line vector of the rotated factor loading matrix $\boldsymbol{A'}$, and the maximum load $b_{ip}$ of the feature index $X_i$ of the matrix $\boldsymbol{A'}$ in $u$ factors to obtain matrix $\boldsymbol{A}^* = (b'_{ij})_{m\times n}$, $i = 1, 2, \ldots, m$; $j = 1, 2, \ldots, u$, where, $b'_{ij} = \begin{cases} b_{ip}, & j = p \\ 0, & \text{other} \end{cases}$.

(6) The sample set $X$ is simplified into the finite sample set $X^{\Delta}$ comprising $n$ samples, where each sample $x_i$ consists of $u$ feature index factors, and the feature index matrix $X^* = (x^*_{ij})_{n\times u} = (x^*_1, x^*_2, \cdots, x^*_n)$ is constructed for $n$ samples from this, where $x^*_{ij}$ is the $j^{\text{th}}$ feature index factor of the $i^{\text{th}}$ sample. Its formula is as follows:

$$x^*_{ij} = \sum_{q=1}^{m} x_{iq} b_{qj};\ i = 1, 2, \cdots, n;\ j = 1, 2, \cdots, u, \tag{1}$$

#### 3.2.2 Density Peak Search Discovery-Based Text Clustering Algorithm

Rodriguez et al. [29] provided a rapid search and discovery density peak-based clustering algorithm that can automatically discover cluster centres and achieve efficient clustering of arbitrarily-shaped data. The algorithm assumes that the local density of data point $x_i$ is $\rho_i$, that the distance from $x_i$ to the local density is larger than this, and that the closest data point $x_j$ in the cluster is $\delta_i$. The clustering decision graph is built by calculating the $\rho_i$ and $\delta_i$ of the arbitrary data point $x_j$. The relative high data points of $\rho_i$ and $\delta_i$ are marked as the central point of the cluster, and the remaining data points are distributed in the cluster of data points closest to it with higher density. In the algorithm, $\rho_i$ and $\delta_i$ are defined as follows:

$$\rho_i = \sum_{j \neq i} \aleph(d_{ij} - d_c) \tag{2}$$

$$\delta_i = \min_{j:\rho_j > \rho_i} (d_{ij}) \tag{3}$$

where, $\aleph(\bullet) = \begin{Bmatrix} 1, \aleph < 0 \\ 0, \aleph > 0 \end{Bmatrix}$, $d_{ij}$ is the distance between different data points, and $d_c$ is the cut-off distance (hyper-parameter).

To better remove interference from noisy samples and provide true and reasonable clustering results, we use the feature index matrix $X^* = (x_{ij}^*)_{n \times u} = (x_1^*, x_2^*, \cdots, x_n^*)$ obtained in Section 3.2.1 as the algorithm input. We then calculate the sample similarity by adjusting the cosine similarity to redefine the variable $d_{ij}$ in the clustering algorithm, and select the cut-off distance $d_c$ in the selection method provided in literature [29]. The value of $d_c$ is obtained by defining the data point $x_i$ as the circle centre, the radius as $d_c$, and the accumulative number of $\rho_i$ as $|X| \times 2\%$. The similarity between vectors $x_i^*$ and $x_j^*$ is:

$$Sim(i,j) = \frac{\sum_{t=1}^{u} (x_{it}^* - \overline{x}_t^*)(x_{jt}^* - \overline{x}_t^*)}{\sqrt{\sum_{t=1}^{u} (x_{it}^* - \overline{x}_t^*)^2} \sqrt{\sum_{t=1}^{u} (x_{jt}^* - \overline{x}_t^*)^2}} \quad (i,j = 1, 2, ..., n) \tag{4}$$

where, $\overline{x_t^*} = \frac{1}{n} \sum_{i=1}^{n} x_{it}^*$, and $u$ is the attribute number of the subject.

Thus, we divide the data set $X$ into $C$ clusters.

The algorithm is described below:

Input: feature index matrix $X^*$ of $n$ samples

Output: $C$ sample clusters

Step 1: calculate the distance $Sim(i,j)$ between any two data points $x_i^*$ and $x_j^*$ in Eq. (4)

Step 2: calculate the local density $\rho_i^*$ of any data point $x_i^*$ in $X^*$ and the distance $\delta_i^*$ between this point and the point with higher local density.

Step 3: use $\rho_i^*$ as the horizontal axis and $\delta_i^*$ as the vertical axis to draw a decision-making diagram.

Step 4: according to the decision graph, mark the points with higher $\rho_i^*$ and $\delta_i^*$ values as the cluster centers, and mark points where $\rho_i^*$ is relatively low but $\delta_i^*$ is relatively high as noise points.

Step 5: distribute the remaining points to obtain $C$ cluster partitions of $n$ samples.

### 3.3 Feature Selection Method for Unbalanced Datasets

#### 3.3.1 $\chi^2$- Value Based on Text Feature Distribution Matrix

Different feature words have large differences in their ability to express text topics and importance. The CHI method considers feature importance according to $\chi^2$ values and, generally, features below a certain $\chi^2$ value contain no or little sample class (cluster) distinction information. However, this understanding is established through the balanced or quasi balanced sample class (cluster) differentiation in the data set. In the case of unbalanced class (cluster) differentiation, the influence of class (cluster) differentiation and feature word frequency on classification are not considered. For unbalanced datasets, the traditional CHI method has obvious defects. To avoid the deficiencies using $\chi^2$-values, after comprehensive consideration of the specific distribution of the features in each sample class (cluster), it is necessary to solve the problem of sample class (cluster) imbalance and feature selection. In this paper, existing $\chi^2$-values are blended using information entropy and average local density to establish a new, weighted, $\chi^2$-value matrix, which can better solve the problem of feature selection in unbalanced datasets. Correction of the distribution of the features in the sample class (cluster) to a certain extent not only clearly shows the actual feature distribution, but also significantly improves the performance of the CHI statistical selection method.

To solve the difference between different sample classes (clusters), the feature $t$ and the sample class (cluster) $c_i$ are simultaneously weighted, and the weighted $\chi^2$-value may be defined as $W\chi^2(t, c_i)$ in this study. Let $W = 1$ in the traditional feature selection method. If a larger weight is distributed to the small class (cluster), the $\chi^2$-value of the small class (cluster) will be increased, and the opportunity to select these features will be increased so as to improve the classification accuracy of the small class (cluster). However, oversize weighted values are distributed to the $\chi^2$-values of the small class (cluster), so it is possible to influence the selection of the feature in the large class (cluster). Therefore, the weight setting is especially important, and the weight is defined as the information entropy of feature $t$ and the sample class (cluster) $c_i$ in this study; that is to say, $W\chi^2(t, c_i)$ is expressed as follows:

$$\begin{cases} W\chi^2(t,c_i) = \chi^2(t,c_i) \times H(t \mid c_i) \times H(c_i) * D(c_i) \\ H(t \mid c_i) = -p(t,c_i) \log p(t \mid c_i) \\ H(c_i) = -p(c_i) \log p(c_i) \\ D(c_i) = \dfrac{\overline{\rho(c_i)}}{\max\limits_{c_i \in C} \overline{\rho(c_i)}} \end{cases} \tag{5}$$

where $p(t|c_i)$ is the probability of feature $t$ occurring in sample class (cluster) $c_i$, $p(c_i)$ is the probability of occurrence of the sample class (cluster) $c_i$, $p(t, c_i)$ is the occurrence probability of feature $t$ and sample class (cluster) $c_i$, $\overline{\rho(c_i)}$ is the average local density of the sample point in sample class (cluster) $c_i$, and $C = \{c_1, c_2,\ldots, c_k\}$ is the sample class (cluster) set. $\overline{\rho(c_i)}$ is defined as $\overline{\rho(c_i)} = \frac{1}{|c_i|}\sum_{c_i.rep} \rho_{c_i.rep}$, where $c_i.rep$ is the sample point in cluster $c_i$.

In Eq. (5), according to the trade-off between information entropy and the density adhesion of the samples, $H(t|c_i)$, $H(c_i)$ and $D(c_i)$ are integrated. On one hand, higher weights are granted to small classes (clusters) so that the $\chi^2$-value can objectively show the impact of the sample class (cluster) distribution on the feature selection, which facilitates the feature selection of small classes (clusters). On the other hand, the average local density of the cluster is introduced to further intensify the impact of sample adhesiveness on feature selection and treat the density imbalance of different clusters.

$$\boldsymbol{K} = \begin{bmatrix} W\chi^2(t_1, c_1) & W\chi^2(t_1, c_2) & \cdots & W\chi^2(t_1, c_m) \\ W\chi^2(t_2, c_1) & W\chi^2(t_2, c_2) & \cdots & W\chi^2(t_2, c_m) \\ \vdots & \vdots & \ddots & \vdots \\ W\chi^2(t_n, c_1) & W\chi^2(t_n, c_2) & \cdots & W\chi^2(t_n, c_m) \end{bmatrix} \quad (6)$$

The statistical matrix $\boldsymbol{K}$ is established by the weighted $\chi^2$-values. The rows and columns in $\boldsymbol{K}$, respectively, are the weighted probability distributions of the feature in different classes (clusters) and the same class (cluster). On this basis, the feature selection can avoid defects resulting from further consideration of the feature or the sample class (cluster).

### 3.3.2 Algorithm description

Input: weighted text $\chi^2$-value matrix $\boldsymbol{K}$.
Output: text feature subset $T$
Algorithm steps:
(1) $T = \varnothing$; // $T$ is the feature set, and initialise
(2) Orderly select each row $t_i$ in $\boldsymbol{K}$ and treat as below:

a. Look up $t_i^{\max} = \max\left\{W\chi^2(t_i, c_j)\right\}$ and $t_i^{\min} \min\left\{W\chi^2(t_i, c_j)\right\}$ in each line.

b. Convert $t_i$ into the corresponding degree of membership $\mu_{ij}$ by $\frac{W\chi^2(t_i, c_j) - \min\left\{t_i^{\min}\right\}}{\max\left\{t_i^{\max}\right\} - \min\left\{t_i^{\min}\right\}}$.

c. Build vector $c_j^* = \left\{b_{ij}\right\}$ // $b_{ij}$ of the new class (cluster) as $t_i$, and arrange the degrees of membership $\mu_{ij}$ in descending order.

d. Calculate $b_i^2 = \sum_{j=1}^{m} b_{ij}^2$, , where // $b_i^2$ is the total contribution of feature $t_i$.

e. Calculate $\varphi = \frac{\sum_{i=1}^{p} b_i^2}{\sum_{i=1}^{n} b_i^2}$; where // $\varphi$ is the accumulated variance contribution rate.

f. When $\varphi \geq 0.85$, obtain the feature subset $T$.

The time complexity of the algorithm is decided mainly by Step (2). The time complexity of the algorithm is $O(n\times m)$ (where $n$ is the feature number and $m$ is the number of the class (cluster). Additionally, according to the specific algorithm step, the space complexity of the algorithm is $O(n)$.

### 3.4 Genetic Extraction Model for Text Topics

The purpose of ICA algorithm is to calculate a separation matrix and obtain a group of mutually-independent random variables. In this paper, the negentropy-based fast fixed-point algorithm (FastICA) [30] is used to find out the mutual independent implicit topic information components by analysing the high-order statistical correlations in the multidimensional data, and extract independent genetic features while removing the high-order redundancy of the components.

### 3.4.1 Negentropy-Based Fast Fixed Point Algorithm

Definition 2: if the density function of the random variable is $p_y(x)$, its differential entropy is defined as follows:

$$H(y) = -\int p_y(x)\log p_y(x)\mathrm{d}x \quad (7)$$

Definition 3: the negentropy $J$ is defined below:

$$J(y) = H(y^*) - H(y) \quad (8)$$

where $y^*$ is a Gaussian random vector with the same correlation (covariance) matrix as $y$.

It is very difficult to directly calculate the negentropy, so it must be calculated approximately. The typical method for negentropy approximation is to use high-order accumulation and a density polynomial. Its corresponding approximation is given below:

$$J(y) \approx \frac{1}{12}E\left\{y^3\right\}^2 + \frac{1}{48}kurt(y)^2 \quad (9)$$

where $kurt(y)$ is the kurtosis of $y$. However, this estimation method is not robust. Therefore, in practice, the expected form of the non-quadratic function $G$ and its corresponding approximate form is as follows:

$$J(y) \propto [E\left\{G(y)\right\} - E\left\{G(v)\right\}]^2 \quad (10)$$

where the function may be selected from $G(y) = \frac{1}{a}\log\cosh ay$ or $G(y) = -\exp(-0.5y^2)$, and $a$ ranges within $1 \leq a \leq 2$ and is generally 1.

Fast ICA algorithm finds out one unit (length) of vector $w$ to maximise the non-Gaussianity of the corresponding projection $w^{\mathrm{T}}z$. The non-Gaussianity is measured by the negentropy approximation $J(w^{\mathrm{T}}z)$, as defined in Eq. (9).

Description of the basic algorithmic form:

① Centralize the data to obtain average 0;
② Whiten the data to obtain $z$;
③ Select estimated number $m$ of independent components, and make $i = 1$;
④ Select one initialized vector $w_i$ (randomly) with unit norm;
⑤ Upgrade $w_i \leftarrow E\left\{zg(w_i^{\mathrm{T}}z)\right\} - \left\{g(w_i^{\mathrm{T}}z)\right\}w_i$, where $g$ is the derivative of the non-quadratic function $G$;
⑥ Standardize $w_i$, $w_i \leftarrow w_i / \| w_i \|$;
⑦ In cases of no convergence, return to Step ⑤;
⑧ Make $i \leftarrow i+1$. In cases of $i \le m$, return to Step ④.

### 3.4.2 Description of Genetic Extraction Algorithm for Text Topics

Obtain the text feature subset $T = t_1, t_2, \ldots, t_p$ of dataset $X$ by the algorithm provided in Section 3.3.2.

(1) Centralisation of the feature subset

Calculate the average vector of the text feature subset $T = t_1, t_2, \ldots, t_p$.

$$\overline{t} = E(T) = \frac{1}{p}\sum_{i=1}^{p} t_i$$

The centralized text feature subset is $\overline{T} = (\overline{t_1}, \overline{t_2}, ..., \overline{t_p})$, where $\overline{t_i} = t_i - \overline{t}$, $\overline{t_i} \in R^n$ and $i = 1, 2, \ldots, p$.

(2) Whitening

Calculate the covariance matrix $\boldsymbol{C}_t$ of the text feature subset $\overline{T} = (\overline{t_1}, \overline{t_2}, ..., \overline{t_p})$.

$$\boldsymbol{C}_t = \boldsymbol{E}\left\{\overline{TT'}\right\} = \boldsymbol{E}\boldsymbol{D}E'$$

where $\boldsymbol{E}$ is the feature vector matrix of $\boldsymbol{C}_t$, $\boldsymbol{E}$ is an orthogonal matrix, $\boldsymbol{D}$ is the feature value matrix of $\boldsymbol{C}_t$, and $\boldsymbol{D}$ is a diagonal matrix.

Linearly whiten $V$ into:

$$V = \boldsymbol{D}^{-\frac{1}{2}}E'$$

The data obtained after whitening is:

$$Z = V\overline{T}$$

(3) Calculate the independent components in the algorithm provided in Section 3.4.1.

## 4 EXPERIMENTAL RESULTS AND ANALYSIS

In this section, the unsupervised text topic-related gene extraction method is verified by experiment. All codes were written in MATLAB R2015a software, and the parameters of the PC used for the compilation runs were: HP Pavilion 15, Intel i7-6500U CPU, 8 GB RAM and Windows 10 64-bit operating system. To validate the proposed method, comparisons were performed between it and several publicly-available datasets: the regularized self-representation-based unsupervised feature selection algorithm (RSR) [25], the feature clustering-based feature selection method (FSFC) [31], the mutual information-based unsupervised feature selection method (UFS-MI) [24], the strong robustness unsupervised feature selection algorithm (RUFS) [26], and the model-induced term-weighted features method (tp-bnb) [32].

### 4.1 Corpus Set

To verify the differences in the performance of multiple methods in different data environments, three datasets from different sources are selected for evaluation testing in this paper.

Dataset A: Twenty newsgroup corpus, total of 20 classes, about 1000 files per class, each category is highly balanced. It contains 18,774 samples and 16,201 features.

Dataset B: Reuter-21578 corpus, testing sample set consisting of 10 classes of 7549 files with an unbalanced distribution. The greatest set included 2877 files constituting 38.11% of the files in the sample set, while the smallest set included 101 files (1.34% of the sample set).

Dataset C: Sohu news data (SogouCS) 20151022 corpus, including 12 classes of 10,902 files, where the greatest class includes 2254 files and the smallest class includes 130 files. To verify the actual treatment effects of all methods, the corpus is supplemented and optimised to a certain degree. For example, some classes of text are supplemented, some incomplete texts are removed, and six classes are selected from the processed corpus: computer games, entertainment, sports and leisure, medicine, natural science, and art: a total of 5493 texts. See Tab. 1 for the specific data structure.

**Table 1** Number of texts in each class of dataset C

| Class | Computer games | Entertainment | Sports leisure | Medicine | Natural science | Art |
|---|---|---|---|---|---|---|
| Quantity (piece) | 1546 | 2049 | 230 | 512 | 842 | 314 |

**Table 2** Composition of topics involved in dataset D

| SN | Topic | Number (Article) | SN | Topic | Number (Article) |
|---|---|---|---|---|---|
| 1 | AlphaGo | 383 | 7 | The incident of "doing politics with one's own trust" | 252 |
| 2 | Chen Sicheng's cheating | 582 | 8 | Turkish nightclub attack | 96 |
| 3 | Xichang Satellite Launch | 60 | 9 | 74th Golden Globe Award | 312 |
| 4 | National conference on science and technology awards | 158 | 10 | Lin Xinru gives birth to a daughter | 375 |
| 5 | Busy preparing for art examination | 93 | 11 | Alipay national bill | 578 |
| 6 | H7N9 avian influenza | 56 | 12 | Midea Group acquires KUKA | 51 |

Dataset D: it contains microblogs related to 12 hot topics on Sina Weibo collected from January 1 to January 10, 2017. Considering the large differences in popularity of the topics, an equal proportion sampling method is adopted, and after the sampling is completed, artificial marking is performed, which includes 2996 pieces of relevant messages and 500 pieces of noisy data for 12 topics. See Tab. 2 for the specific data structure.

In the text pre-processing stage, the Chinese corpus is processed by the ICTCLAS Chinese word segmentation tool of the Chinese Academy of Sciences. The English corpus is processed by the porter algorithm. K-nearest neighbour (KNN), naive Bayesian methods and Support Vector Machine (SVM) are used as the classification algorithm, the k-means clustering method is selected as the clustering method. The neighbour parameter used in the contrast algorithm is set to five and cosine similarity is selected as the vector similarity.

### 4.2 Evaluation Test Index

Evaluation of the algorithm's classification results was performed using the macro average recall ratio, macro average accuracy rate, macro average F1 value and other indexes. The algorithm's clustering results were evaluated according to the normalized mutual information.

1) The macro average recall ratio is given in Eq. (11).

$$Macro\_\overline{r} = \frac{1}{|C|}\sum_{i=1}^{|C|} r_i \tag{11}$$

where $r_i$ is the recall ratio of class $i$, and $|C|$ is the class number.

2) The macro average accuracy rate is given in Eq. (12).

$$Macro\_\overline{p} = \frac{1}{|C|}\sum_{i=1}^{|C|} p_i \tag{12}$$

where $p_i$ is the accuracy rate of class $i$.

3) The macro average F1 value is given in Eq. (13).

$$Macro\_\overline{F1} = \frac{1}{|C|}\sum_{i=1}^{|C|} F1_i \tag{13}$$

where $F1_i$ is the $F1$ value of class $i$.

4) The level of similarity between different partitions of the same dataset can be measured by the normalized mutual information, as shown in Eq. (14).

$$NMI(U,V) = \frac{I(U,V)}{\sqrt{H(U)H(V)}} \tag{14}$$

where $U$ and $V$ are two different partitions of the same dataset, and $I(U, V)$ is the mutual information of $U$ and $V$.

### 4.3 Experimental Testing and Analysis

To obtain experimental results with high statistical significance, this paper used the five-fold cross validation method for evaluation. Figs. 2 and 3, respectively, show the classification results for RSR, FSFC, UFS-MI, RUFS, tp-bnb and UTTGE obtained by the KNN and naïve Bayes classifiers on the 20 newsgroup corpuses. According to Figs. 2 and 3, the effect is best for UTTGE and worst for FSFC. For RSR, UFS-MI, tp-bnb and RUFS, UFS-MI was dominant. When fewer feature numbers are selected, the UTTGE method has higher classification accuracy than the other five methods and UFS-MI and tp-bnb had similar results to UTTGE. However, it can be seen that the classification performance of these methods is reduced to a certain degree at low dimensions. This is mainly because of the impact on classification performance of the many empty files that appeared in the feature dimension reduction process. Therefore, it cannot be said that selecting fewer features gives a better result. Figs. 2 and 3 show that when the feature number is increased to a critical point, the performance of the classifier declines to a certain extent mainly because of the impact of the many invalid classification features that are introduced. Hence, feature dimension reduction must be carefully chosen within a rational range.

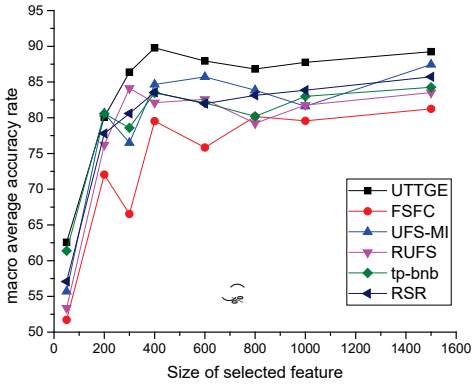



**Figure 2** Comparison of macro average accuracy rate of KNN classifier on 20 Newsgroups corpus

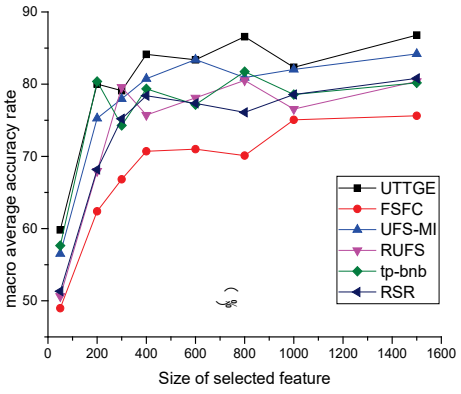



**Figure 3** Comparison of macro average accuracy rate of Naïve Bayes classifier on 20 Newsgroups corpus

Tab. 3 shows the experimental results of RSR, FSFC, UFS-MI, RUFS, tp-bnb and UTTGE obtained by the KNN classifier on the Reuter-21578 corpus. According to Tab. 3, UTTGE has a slightly lower macro average accuracy rate than UFS-MI but only when the feature number is 200. When the feature number is 200 or 500, UTTGE has a slightly lower macro average $F1$ value than tp-bnb. When the feature number is 50, the macro average accuracy rates of RSR, FSFC, UFS-MI, tp-bnb and RUFS are less than 75% while UKGE-MS achieved 76.01%. The macro

average F1 values of RSR, FSFC, UFS-MI, tp-bnb and RUFS are less than 70%, while UKGE-MS achieves 70.28%. For UTTGE, with increases in the feature number, the macro-accuracy rate tends to become stable after reaching the optimal value of 87.2%, and both exceeded 85%. The other five methods had macro average accuracy rates of less than 85%.

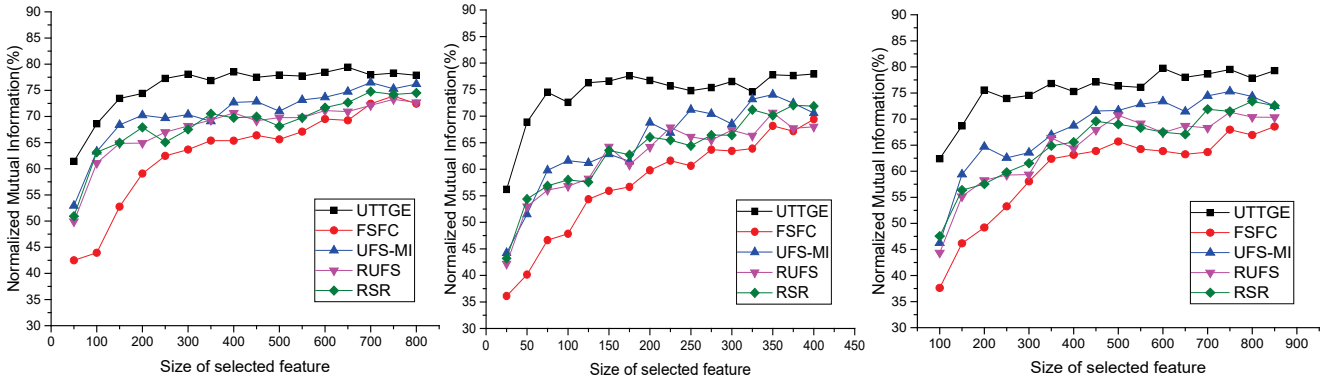
**Table 3** Performance comparison of feature selection methods (results shown as %)

| Features | RSR | | FSFC | | UFS-MI | | RUFS | | tp-bnb | | UTTGE | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $Macro\_\bar{p}$ | $\overline{Macro\_F1}$ | $Macro\_\bar{p}$ | $\overline{Macro\_F1}$ | $Macro\_\bar{p}$ | $\overline{Macro\_F1}$ | $Macro\_\bar{p}$ | $\overline{Macro\_F1}$ | $Macro\_\bar{p}$ | $\overline{Macro\_F1}$ | $Macro\_\bar{p}$ | $\overline{Macro\_F1}$ |
| 50 | 67.34 | 63.25 | 65.2 | 60.86 | 70.84 | 65.38 | 65.28 | 59.26 | 71.09 | 64.39 | **76.01** | **70.28** |
| 100 | 78 | 72.99 | 77.15 | 72.72 | 81.57 | 75.52 | 77.94 | 72.24 | 78.51 | 75.26 | **84.88** | **78.12** |
| 200 | 79.75 | 73.65 | 78.44 | 73.51 | **83.83** | 77.32 | 79.91 | 73.45 | 83.46 | **78.94** | 83.14 | 78.61 |
| 300 | 80.03 | 74.71 | 78.2 | 73.28 | 83.48 | 77.26 | 79.87 | 73.99 | 80.36 | 75.07 | **87.2** | **81.2** |
| 500 | 80 | 75.09 | 78.96 | 74.56 | 82.57 | 76.87 | 81.03 | 74.97 | 83.09 | **81.01** | **86.14** | 80.71 |
| 1000 | 79.36 | 74.38 | 78.19 | 72.94 | 82.15 | 77.1 | 78.73 | 73.18 | 83.22 | 78.8 | **86.48** | **80.67** |

Tab. 4 shows the experimental results of the RSR, FSFC, UFS-MI, RUFS, tp-bnb and UTTGE methods are obtained by the naïve Bayes classifier on the Sohu News dataset (SogouCS) 20151022. According to Tab. 4, UTTGE has a higher macro average accuracy rate than the other four methods except with 100 features, when it has a slightly lower macro average $F1$ value than tp-bnb. With 100 features, RSR, FSFC, UFS-MI and RUFS have macro average accuracy rates less than 60%, while UKGE-MS has a 73.12% macro average accuracy rate. Only tp-bnb has a result close to that of UTTGE, which shows that when fewer features are selected, UTTGE performs better than the other five methods. With increases in feature number, for UTTGE, the classification accuracy is stable after reaching the optimal value and has a macro average accuracy rate more than 80%, while the other four methods are less than 80% and also have lower macro average $F1$ values.

**Table 4** Performance comparison of feature selection methods (%)

| Features | RSR | | FSFC | | UFS-MI | | RUFS | | tp-bnb | | UTTGE | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $Macro\_\bar{p}$ | $\overline{Macro\_F1}$ | $Macro\_\bar{p}$ | $\overline{Macro\_F1}$ | $Macro\_\bar{p}$ | $\overline{Macro\_F1}$ | $Macro\_\bar{p}$ | $\overline{Macro\_F1}$ | $Macro\_\bar{p}$ | $\overline{Macro\_F1}$ | $Macro\_\bar{p}$ | $\overline{Macro\_F1}$ |
| 100 | 56.65 | 52.37 | 55.51 | 47.45 | 53.58 | 47.85 | 56.2 | 51.44 | 72.28 | **66.12** | **73.12** | 65.37 |
| 300 | 76.88 | 71.73 | 71.65 | 65.67 | 77 | 71.21 | 73.7 | 69.25 | 77 | 72.04 | **82.4** | **75.43** |
| 500 | 77.76 | 73.11 | 73.6 | 66.3 | 79.02 | 72.92 | 76.66 | 70.37 | 74.69 | 69.9 | **81.76** | **76.58** |
| 1000 | 78.42 | 73.58 | 74.43 | 67.03 | 78.88 | 72.86 | 75.11 | 69.48 | 75.77 | 71.91 | **84.02** | **77.55** |
| 2000 | 77.83 | 73.5 | 74.06 | 66.57 | 79.36 | 74.59 | 76.61 | 70.96 | 77.76 | 73.31 | **83.64** | **76.97** |




(a) 20 Newsgroups corpus (b) Reuter-21578 corpus (c) Sohu News data (SogouCS) 20151022 corpus

**Figure 4** Normalized mutual information of all algorithms according to feature number

Tab. 5 shows the macro average recall ratio peak of the different features selected from the three datasets from different sources by the KNN classifier. According to Tab. 5, with Datasets A and C, UTTGE has a significantly higher macro average recall ratio than the other five methods. With dataset B, UTTGE has a slightly lower macro average recall ratio than tp-bnb. However, it can be seen that when the Chinese unbalanced dataset is processed, the performance of UTTGE declines slightly compared with that of the English dataset, mainly because there are more factors impacting Chinese text than English text during processing, and the semantic conductive influences of the words are more significant.

**Table 5** Macro average recall ratios of the six methods (%)

| | UTTGE | RSR | FSFC | UFS-MI | RUFS | tp-bnb |
|---|---|---|---|---|---|---|
| Dataset A | **78.99** | 75.95 | 70.34 | 78.19 | 73.31 | 78.03 |
| Dataset B | 76.17 | 70.75 | 70.62 | 72.63 | 69.75 | **79.03** |
| Dataset C | **73.51** | 70.3 | 61.45 | 67.69 | 66.09 | 69.32 |

Tab. 6 shows that the experimental results of the RSR, FSFC, UFS-MI, RUFS, tp-bnb and UTTGE are obtained

on Datasets D by using LIBSVM classifier (RBF kernel function, the dimension of the feature space vector is set to 300).

**Table 6** The overall performance comparison of six methods

| Methods | $Macro\_\overline{p}$ (%) | $Macro\_\overline{F1}$ (%) |
|---|---|---|
| RSR | 78.31 | 73.28 |
| FSFC | 74.95 | 67.26 |
| UFS-MI | 80.69 | 73.38 |
| RUFS | 76.73 | 70.97 |
| tp-bnb | 79.55 | 74.02 |
| UTTGE | **84.32** | **77.18** |

From Tab. 6, it can be seen that UTTGE method significantly improves the macro average accuracy and macro average $F1$ value obtained by LIBSVM classifier compared with other methods. Based on the analysis of the experimental results in Tabs. 3-6, it can be known that the UTTGE method can accurately select the optimal feature subset that comprehensively and truly reflects texts' topic information, which can effectively improve the classification and recognition of texts.

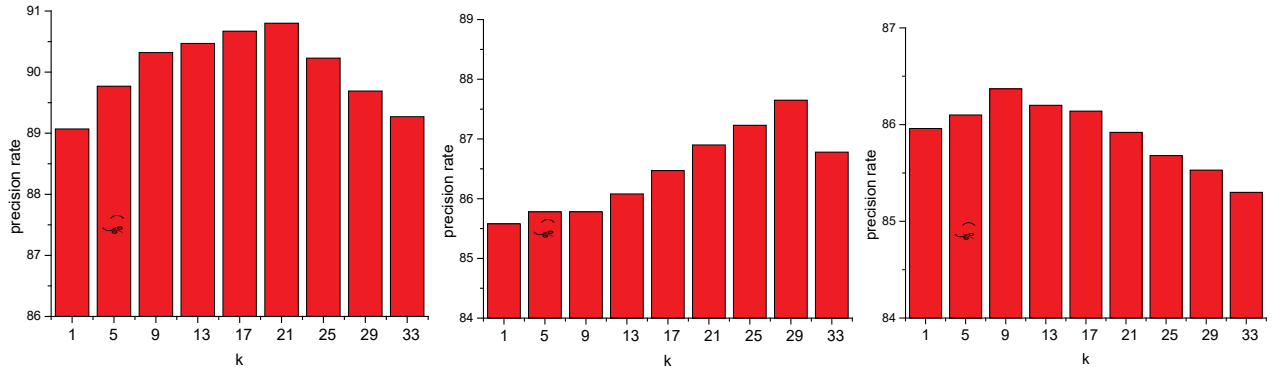
Referring to the idea of using a clustering algorithm to verify the validity of the classification algorithm proposed in literature [33], and the k-means clustering algorithm is used to analyze the clustering and original categories of the text datasets considered in this paper. The normalized mutual information value of the dataset is used to measure the effectiveness of the algorithm.

As it is necessary to clearly define a cluster number during $k$-means clustering to reduce the impact of the $k$-value section on the method, and the cluster number $k$ used for the proposed and comparison methods are set to the class number include in the data labels, i.e. 20, 10 and 12.

Fig. 4 shows the value corresponding to the normalized mutual information under different conditions. It also can be seen in Fig. 4 that the proposed UTTGE method has obvious advantages compared with the other four algorithms. As shown in Figs. 4(b) and (c), when the feature number is lower, UTTGE can rapidly achieve better results. Therefore, compared with the common unsupervised feature selection algorithm, the proposed algorithm performs better during unsupervised feature selection.

### 4.3 Parameter Analysis

Selection of the $k$ value has a major impact on the results of the KNN algorithm. If a smaller $k$ value is selected, only training samples that are close to the input sample affect the forecast result, which can cause overfitting. If a larger $k$ value is selected, its advantage is that it can reduce the learning estimation error, and it may also increase the learning approximation error. In this case, the difference between the training sample and the input sample will affect the prediction and cause the prediction error. Therefore, in practical application, the smaller $k$ value is generally selected, and the best $k$ value is selected by cross validation method. Fig. 5 shows the classification results of the KNN algorithm with three datasets, with the UTTGE method using various $k$ values. The plots of the experimental results from parabolic shapes. With the 20 Newsgroup corpus dataset, the KNN algorithm has the best classification effect when $k = 21$ (Fig. 5a). With the Reuter-21578 dataset, the optimal value is $k = 29$ (Fig. 5b). Meanwhile, with the Sohu News dataset (SogouCS), $k = 9$ is optimal (Fig. 5c).




(a) 20 Newsgroups corpus (b) Reuter-21578 corpus (c) Sohu News data (SogouCS) 20151022 corpus

**Figure 5** Accuracy of KNN algorithm at different $k$ value

## 5 CONCLUSIONS AND FUTURE WORK

This paper has explored the traditional feature-dimension reduction method with unbalance datasets, and proposes an unsupervised text topic-related gene extraction method based on the density peak, $\chi^2$ distribution matrix, and an independent component analysis approach. This method does not need large-scale training of marked samples or valid pre-definition of class relationships and relevant features, and overcomes disadvantages of poor generalization of models resulting from unbalanced distributions. On the basis of the rapid search and discovery peak text clustering method, the text feature distribution features of weighted $\chi^2$-values were determined by information entropy, which avoids changes in the class distribution of unbalanced datasets caused by oversampling and undersampling methods. The performance of the CHI statistical selection method is significantly improved by correcting the feature class distribution. Finally, the independent implicit information component of multi-dimensional data is extracted by the negentropy-based fast fixed-point algorithm (FastICA), and its feature subset has better generalisation performance than the RSR, FSFC, UFS-MI, RUFS and tp-bnb methods.

Feature dimension reduction is achieved under the condition of maintaining the identifiability of the dataset. Feature dimension reduction is a key step in the pre-processing of large industrial and social datasets [34, 35].

The genetic extraction approach proposed in this paper will play an important role in the field of "big data" processing. So future work will explore how to better meet the data processing requirements in this field.

### Compliance with Ethical Standards

This study was funded by Anhui province philosophy and social science planning project (No. AHSKY2018D 09). The authors would like to thank the anonymous revie wers for their helpful comments and suggestions.

## 6 REFERENCES

[1] Ma, L. & Liu, H. F. (2015). Study on the extraction of characteristics of Chinese text based on the LDA model. *Journal of Xi'an University of Posts and Telecommunications, 6*, 79-81.

[2] Cao, Y., Jia, L., Chen, Y. et al. (2018). Recent advances of generative adversarial networks in computer vision. *IEEE Access, 7*, 14985-15006. https://doi.org/10.1109/ACCESS.2018.2886814

[3] Zhang, X., Cheng, J., Yuan, T. et al. (2013). TopRec: domain-specific recommendation through community topic mining in social network. https://doi.org/10.1145/2488388.2488519

[4] Bakar, N. H., Kasirun, Z. M., & Salleh, N. (2015). Feature extraction approaches from natural language requirements for reuse in software product lines: *A systematic literature review. Journal of Systems & Software, 106*(C), 132-149. https://doi.org/10.1016/j.jss.2015.05.006

[5] Abualigah, L. M. Q. & Hanandeh, E. S. (2015). Applying Genetic Algorithms to Information Retrieval Using Vector Space Model. *International Journal of Computer Science, Engineering and Applications, 5*(1), 19-28. https://doi.org/10.5121/ijcsea.2015.5102

[6] Xu, X., He, X., Ai, Q. et al. (2015). A Correlation Analysis Method for Power Systems Based on Random Matrix Theory. *IEEE Transactions on Smart Grid, PP*(99), 1-10.

[7] Chen, Q., Hu, L., Xu, J. et al. (2015). Document similarity analysis via involving both explicit and implicit semantic couplings. *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA), IEEE, 2015*, 1-10. https://doi.org/10.1109/DSAA.2015.7344832

[8] Verma, N. & Branson, K. (2015). Sample complexity of learning mahalanobis distance metrics. *Computer Science, 2015*, 2584-2592.

[9] Xu, J. & Jiang, H. (2015). An Improved Information Gain Feature Selection Algorithm for SVM Text Classifier. *2015 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, IEEE, 2015*, 273-276. https://doi.org/10.1109/CyberC.2015.53

[10] Assunção, M. D., Calheiros, R. N., Bianchi, S. et al. (2015). *Big Data computing and clouds: Trends and future directions. Journal of Parallel & Distributed Computing, s 79-80*, 3-15. https://doi.org/10.1016/j.jpdc.2014.08.003

[11] Jiang, L. Q, Zhang, M. X, Zheng, J. L. et al. (2016). Optimization of clustering by fast search and find of density peaks. *Application Research of Computers, 33*(11), 3251-3254.

[12] Xie, L., Li, G., Xiao, M. et al. (2016). Novel classification method for remote sensing images based on information entropy discretization algorithm and vector space model. *Computers & Geosciences, 89*(C), 252-259. https://doi.org/10.1016/j.cageo.2015.12.015

[13] Wang, B., Yan, X., & Jiang, Q. (2016). Independent component analysis model utilizing de-mixing information for improved non-Gaussian process monitoring. *Computers & Industrial Engineering, 94*, 188-200. https://doi.org/10.1016/j.cie.2016.01.021

[14] Chawla, N. V., Bowyer, K. W., Hall, L. O. et al. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*(1), 321-357. https://doi.org/10.1613/jair.953

[15] Chen, R., Zhang, L., Yang, J. et al. (2014). Classification algorithm for imbalanced data sets based on combination of BSMOTE and inverse under sampling. Application Research of Computers, 2014(11), 3299-3303.

[16] Fang, X. N, Zhang, H. X, & Gao, S. (2013). Web spam detection based on SMOTE and random forests. *Journal of Shandong University (Engineering Science), 2013*(01), 22-27.

[17] Li, J. L. & Fu, H. G. (2010). Improved KDE-based data classification algorithm. *CONTROL AND DECISION, 25*(04), 507-514.

[18] Zhang, W. S. & Yu, T. Z. (2016). Research on Boosting theory and its applications. *Journal of University of Science and Technology of China*, (3), 222-230.

[19] Yin, S. A. (2013). Classification method for imbalanced data sets based on FCM-KFDA discriminant. *Journal of Huazhong Normal University, 52*(6), 776.

[20] Ji, X., Zhou, L., & Wu, Q. (2015). A Novel Action Recognition Method Based on Improved Spatio-Temporal Features and AdaBoost-SVM Classifiers. *International Journal of Hybrid Information Technology, 8*(12), 5136-5145. https://doi.org/10.14257/ijhit.2015.8.5.19

[21] Zheng, Z., Wu, X., & Srihari, R. (2010). Feature selection for text categorization on imbalanced data. *Acm Sigkdd Explorations Newsletter, 6*(1), 80-89. https://doi.org/10.1145/1007730.1007741

[22] Khoshgoftaar, T. M., Gao, K., & Napolitano, A. (2012). Exploring an iterative feature selection technique for highly imbalanced data sets. *2012 IEEE 13th International Conference on Information Reuse & Integration (IRI), IEEE, 2012*, 101-108. https://doi.org/10.1109/IRI.2012.6302997

[23] Lin, Y. E., Li, J. Z., Liang, X. Z. et al. (2012). Direct Unsupervised Orthogonal Locality Preserving Method for Feature Extraction. *Opto-Electronic Engineering, 39*(3), 100-105

[24] Xu, J. L., Zhou, Y. M., Chen, L. et al. (2012). An Unsupervised Feature Selection Approach Based on Mutual Information. *Journal of Computer Research and Development, 49*(2), 372-382.

[25] Zhu, P., Zuo, W., Zhang, L. et al. (2015). Unsupervised feature selection by regularized self-representation. *Pattern Recognition, 48*(2), 438-446. https://doi.org/10.1016/j.patcog.2014.08.006

[26] Li, J., Hu, X., Wu, L. et al. (2016). Robust Unsupervised Feature Selection on Networked Data. *Proceedings of the 2016 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, 2016*, 387-395. https://doi.org/10.1137/1.9781611974348.44

[27] Alibeigi, M., Hashemi, S., & Hamzeh, A. (2011). Unsupervised Feature Selection Based on the Distribution of Features Attributed to Imbalanced Data Sets. *International Journal of Artificial Intelligence & Expert Systems, 2*(1), 2011-2014.

[28] Gupt, Y. & Sahay, S. (2015). Review of extended producer responsibility: A case study approach. *Waste Management & Research the Journal of the International Solid Wastes & Public Cleansing Association Iswa, 33*(7), 595. https://doi.org/10.1177/0734242X15592275

[29] Rodriguez, A. & Laio, A. (2014). Clustering by fast search and find of density peaks. *Science, 344*(6191), 1492. https://doi.org/10.1126/science.1242072

[30] Du, Y., Zhang, Y. D., Li, M. H. et al. (2016). Improved FastICA algorithm for data optimization processing in intrusion detection. *Journal on Communications, 37*(1), 42-

48.
[31] Wang, L. X. & Jiang, S. Y. (2015). Novel feature selection method based on feature clustering. *Application Research of Computers, 2015*(5), 1305-1308.
[32] Kim, H. K. & Kim, M. (2016). Model-induced term-weighting schemes for text classification. *Applied Intelligence, 45*(1), 30-43. https://doi.org/10.1007/s10489-015-0745-z
[33] Cai, D., Zhang, C., & He, X. (2010). Unsupervised feature selection for multi-cluster data. *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, 2010*, 333-342. https://doi.org/10.1145/1835804.1835848
[34] Li, Q., Qu, H., Liu, Z. et al. (2019). AF-DCGAN: Amplitude Feature Deep Convolutional GAN for Fingerprint Construction in Indoor Localization Systems. *IEEE Transactions on Emerging Topics in Computational Intelligence.* https://doi.org/10.1109/TETCI.2019.2948058
[35] Hu, H., Liu, Z., & An, J. (2020). Mining mobile intelligence for wireless systems: A deep neural network approach. *IEEE Computational Intelligence Magazine, 15*(1), 24-31. https://doi.org/10.1109/MCI.2019.2954641

**Contact information:**

**Li JING-MING,** Lecturer
School of Management Science and Engineering,
Anhui University of Finance and Economics,
962 Caoshan Road, 233030, Bengbu, Anhui Province, China
E-mail: 645675686@qq.com

**Sun JING-TAO,** Senior engineer
(Corresponding author)
School of Computer Science and Technology,
Shaanxi Key Laboratory of Network Data Analysis and Intelligent Processing,
Xi'an University of Posts and Telecommunications,
710021, Xi'an, Shaanxi Province, China
E-mail: sun2651@qq.com

**Huang WEN-HAN,** Lecturer
School of Mathematics and Computer Science,
Shaanxi University of Technology,
723001, Hanzhong Shaanxi, China

**Zhang QIU-YU,** Researcher/PhD supervisor
School of Computer and Communication,
Lanzhou University of Technology,
36 Pengjiaping Road, Qilihe District, Lanzhou, Gansu, China
E-mail: zhangqylz@163.com

**Tian ZHEN-ZHOU,** Lecturer
School of Computer Science and Technology,
Xi'an University of Posts and Telecommunications,
710021, Xi'an, Shaanxi Province, China

**Lu NING,** Lecturer
School of Computer Science and Technology,
Xi'an University of Posts and Telecommunications,
710021, Xi'an, Shaanxi Province, China