IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 10, 2018, accepted December 27, 2018, date of publication January 1, 2019, date of current version January 23, 2019.

Digital Object Identifier 10.1109/ACCESS.2018.2890549

Feature Selection With Ensemble Learning Based on Improved Dempster-Shafer Evidence Fusion

YIFENG ZHENG[1,2,3,4,5], GUOHE LI[1,2], WENJIE ZHANG[3,4,5], YING LI[1,2], AND BAOYA WEI[3,4,5]

[1]College of Information Science and Engineering, China University of Petroleum, Beijing 102249, China
[2]Beijing Key Laboratory of Data Mining for Petroleum Data, China University of Petroleum, Beijing 102249, China
[3]School of Computer Science, Minnan Normal University, Zhangzhou 363000, China
[4]Key Laboratory of Data Science and Intelligence Application, Minnan Normal University, Zhangzhou 363000, China
[5]Laboratory of Granular Computing, Minnan Normal University, Zhangzhou 363000, China

Corresponding author: Yifeng Zheng (zyf@mnnu.edu.cn)

This work was supported in part by the Nature Science Foundation of China under Grant 60473125 and Grant 61701213, in part by the Innovation Foundation of CNPC under Grant 05E7013, in part by the National Key Project Foundation of Science under Grant G5800-08-ZS-WX, in part by the Science Foundation of the China University of Petroleum-Beijing at Karamay under Grant RCYJ2016B-03-001, in part by the Cooperative Education Project of the Nation Education Ministry under Grant 201702098015, in part by the Research Fund for the Educational Department of Fujian Province under Grant JA15300 and in part by the Fujian Province Natural Science Funds under Grant 2018J01545 and Grant 2018J01546.

ABSTRACT Feature selection or attribute reduction is an important data preprocessing technique for dimensionality reduction in machine learning and data mining. In this paper, a novel feature selection ensemble learning algorithm is proposed based on Tsallis entropy and Dempster–Shafer evidence theory (TSDS). First, an improved correlation criterion is used to obtain the relevant feature based on Tsallis entropy. A forward sequential approximate Markov blanket is then defined to eliminate the redundant feature. An ensemble learning is employed to achieve approximately optimal global feature selection, which can acquire the feature subsets from different perspectives. Finally, by fusing all the feature subsets, the improved evidence theory approach is utilized to gain the final feature subset. To verify the effectiveness of TSDS, nine datasets from two different domains are used in the experimental analysis. The experimental results demonstrate that the proposed algorithm can select feature subset more effectively and enhance the classification performance significantly.

INDEX TERMS Feature selection, Tsallis entropy, Dempster-Shafer theory, feature selection ensemble, approximate Markov blanket.

I. INTRODUCTION

Machine learning aims to acquire knowledge from data. In practical application, the data increases both in scales of samples and features. These large-scale data may include hundreds even thousands of features, which result in the curse of dimensionality. Because of containing a mass of redundant attributes, the performance of classification has a great impact on machine learning. Therefore, elimination of insignificant information is increasingly being recognized as a key element in extracting potential useful information. Feature selection or attribute reduction is an important data optimization technique for dimensionality reduction, which focuses on eliminating irrelevant and redundant attributes from a dataset. The main purpose of feature selection is not only to find a suitable subset from original feature space, but also to retain high classification precision and preserve original meaning of those features after reduction. A typical feature selection process (called selector) can be divided into three steps: subset generation, subset evaluation, and stopping criterion [1], [2], as shown in Fig.1.

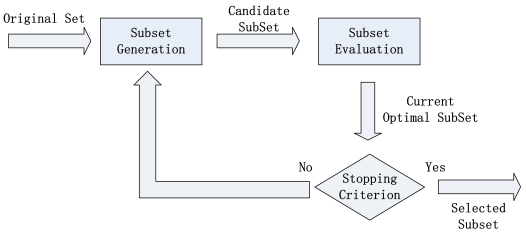


FIGURE 1. The typical feature selection process.

A candidate feature subset is constructed from the original feature space according to feature searching strategy in the subset generation process, whose efficiency is evaluated and compared with the previous one according to the

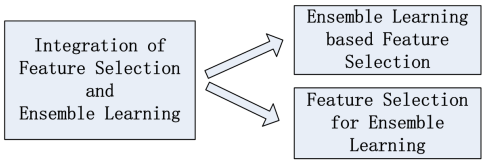


FIGURE 2. Integration approach of feature selection and ensemble learning.

evaluation criterion. Subset generation and subset evaluation are repeated until a given stopping criterion is satisfied. Fig.1 shows that the evaluation criterion is one of the key factors in feature selection. According to the evaluation criterion, feature selection algorithms are divided into wrapper, embedded, and filter methods. Wrapper feature selection approaches utilize a predefined classification model to evaluate the selected feature subset. Embedded feature selection models focus on embedding the feature selection process into the classifier construction [2]. A general drawback of these schemes is high computational complexity in wrapper models and embedded models. Besides, the selected feature subset obtained by these models is associated with the learning algorithm as well. Compared with these two methods, filter feature selection algorithms analyze the characteristics of data and evaluate features independently of any specific classifier. Features are ranked based on given criteria, and then the features with the highest ranking are employed to construct classification models [3], [4]. In filter models, the evaluation functions are of four broad categories: consistency [5], [6], distance criterion [7], [8], dependency criterion [9]–[11], and information metrics [12]–[15].

Over the last few years, many researchers pay more attention to ensemble learning in classification tasks, which combines the consequence of multiple base classifiers. Similarly, the principle of ensemble learning can also be utilized for feature selection. It can effectively incorporate ensemble learning into feature selection. There are two different integration strategies, as shown in Fig.2. One is to employ Ensemble Learning for Feature Selection (ELFS) [16]–[18]. It obtains an approximate optimal feature subset by combining multiple feature subsets based on the nature of ensemble learning [19]. The other one is to utilize Feature Selection for Ensemble Learning (FSEL) [20]–[22]. It utilizes different feature subsets to construct an ensemble of a diverse-based classifier [19]. At present, the output of the feature selector is partitioned into two general types: feature weighting, feature subset. For the former one, a weight is assigned to each feature after feature selection process. And then, the average weight of each feature is calculated in all base feature selectors. For the latter one, the cumulative number of each feature appeared in the output of all base selectors is listed in descending order, just as maximum majority voting.

The aforementioned ensemble strategy does not embody the uncertain of features. Therefore, evidence fusion [23]–[25] is utilized to integrate the output of all base feature selectors. In this paper, a novel feature selection ensemble algorithm is proposed based on Tsallis entropy and Dempster-Shafer evidence theory (TSDS). The study has been started with the view to obtain potentially informative features, to retain the nature of the original data.

The basic idea is as follows: 1) We utilize the modified symmetrical uncertainty and the forward searching approximate Markov blanket to measure the distinguishing ability of each feature with class label based on Tsallis entropy, which can effectively eliminate irrelevant and redundant feature to preserve optimal feature subset approximately. 2) We employ FSEL to gain the feature subset from different aspects. 3) By combining the credibility degree and Tsallis information entropy, the Dempster combination rule is used to realize information fusion to gain the final feature subset.

In summary, the key contributions of this paper are summarized as follows:

- *Correlation criterion*: The improved symmetrical uncertainty and forward sequential searching approximate Markov blanket are proposed to obtain optimal feature subset by the forward sequential selection.
- *Ensemble fusion strategy*: The novel evidence fusion approach based on Dempster-Shafer evidence theory is proposed to combine the consequence of each feature selection.
- *Effectiveness*: To prove the effectiveness of our proposed algorithm, a series of experiments have been performed to compare with DISR [26], CMIM [27], JMI [28], MRMR [29], SPFS-LAR [30], QIS [31] and HANDI [32].

The rest of the paper is organized as follows. Some related works are briefly reviewed in Section II. The proposed novel feature selection ensemble algorithm is presented in Section III. Experimental results and evaluations are given in Section IV. Finally, the conclusion is given in Section V.

II. RELATED WORKS

Feature selection has attracted a lot of attention in data preprocessing, for example, data optimization in machine learning, pattern recognition, and so on. In general, prior works on feature selection can mainly be categorized into three classes: wrapper, embedded, and filter methods.

In wrapper models, the feature subset with the highest assessment value will be chosen as the final subset by a predefined classifier. For example, Bermejo *et al*. [33] proposed a NB-based embedded incremental wrapper feature subset feature algorithm (IWSS). It can be updated when new features are added gradually. Chen and Chen [34] utilized the cosine distance to support vector machine to eliminate irrelevant or redundant features, namely cosine similarity measure SVM (CSMSVM). In the CSMSVM framework, feature selection, SVM parameter learning and low relevance features removing are accomplished together by optimizing the shape of an anisotropic radial basis function kernel in feature space. Wrapper models perform the classifiers many times in the feature selection process, which leads to the inadequacy of computationally prohibitive. Meanwhile, the selected feature

subset is inevitably biased to the preassigned classification model.

Embedded approaches incorporate the process of feature selection as part of the model learning. It is divided two major kinds: pruning method, and build-in mechanism model. The pruning method utilizes all features to train a classifier model, and then eliminates some features according to pruning strategy while maintaining the classification performance. The latter method is classifier with a build-in mechanism for feature selection. For example, Maldonado and López [35] utilized the embedded strategy to penalize the cardinality of feature set by using a concave approximation scaling factors technique. It can effectively enhance performance in high-dimensionality under a class-imbalance condition. Tao *et al*. [36] proposed a robust multi-source adaptation embedding framework by employing the correlation information which combines with joint $L_{2,1}$-*norm* and *trace-norm* regularization. Chung *et al*. [37] proposed a new learning scheme based on fuzzy rule to select useful features with controlled redundancy. In addition, it can discard derogatory and indifferent features.

One of the great advantages of the filter models over the wrapper and the embedded models is that none of classifier learning algorithms is taken into account in the feature selection process. Therefore, the selected features can represent the characteristics of the original data. Thus, many researchers pay attention to utilizing filter models in practical applications. For example, Nayak *et al*. [38] presented a filter feature selection algorithm by employing elitism based on multi-objective differential evolution. The objective function takes both linear and nonlinear dependency among features into account. Lyu *et al*. [39] advanced both the maximal information coefficient and gram-schmidt orthogonalization to address the irrelevant redundancy problem. A novel filter method based on multi-variable relative discrimination criterion is proposed for text classification in [40]. The document frequencies for each term are utilized to estimate their availability. Furthermore, Hancer *et al.* [41] elaborated a novel filter approach which can utilize information theory and evolutionary computation technique to extract the optimal feature subset. It proposed two different criterions to construct single-objective and multi-objective algorithms. Wang *et al*. [42] proposed a new feature selection approach to globally minimize the feature redundancy with maximizing the given feature ranking scores. In [43], F2DDLPP (fuzzy 2D discriminant locality preserving projections) is proposed for image feature selection. The fuzzy k-nearest neighbor is used to calculate the membership degree matrix. Then, F2DDLPP incorporates the membership degree matrix into the intra-class scatter matrix and inter-class scatter, respectively. It can extract discriminative features from overlapping samples effectively. Wan *et al*. [44] proposed a novel method 2DMED (two-dimensional maximum embedding difference) which combines graph embedding and difference criterion techniques for image feature extraction. It extracts the optimal projective vectors from 2D image matrices and does not convert the image matrix into a high-dimensional image vector. Wan *et al*. [45] proposed local graph embedding method based on maximum margin criterion for face recognition. Two novel fuzzy Laplacian scatter matrices are calculated using fuzzy k-nearest neighbor. In addition, maximum margin criterion is utilized to avoid the problem of small size sample.

Analysis of ensemble learning models demonstrates that their performance is better than the result of any single classifier. Similarly, although with diversity, a number of feature selection algorithms adopt a single feature selection process. Li *et al*. [16] proposed a diversity regularized ensemble feature weighting framework. The base feature selector is based on local learning. The feature weighting is converted directly to a ranking vector. Bolón-Canedo *et al*. [17] presented a new feature selection framework for an ensemble of filters and classifiers with different metrics. The outputs of these classifiers are combined by simple voting.

The filter approaches mentioned above do not take into account sample diversity in sampling the training set. They obtain only one feature subset in the feature selection process. Therefore, our work differs from the above-mentioned approaches in the evaluation criterion for feature selection and the ensemble fusion strategy based on Dempster-Shafer evidence theory aspects.

III. TSDS ALGORITHM

An information system can be written as $IS = \langle U, A, V \rangle$, where $U = \{u_1, u_2, \cdots u_n\}$ is a nonempty universe of objects ($|\ U\ |$ denotes the number of U). A is the attribute set, and V is the domain for attributes. For $IS = (U, A, V)$, V can be expressed as $\{V_\alpha \mid \alpha \in A\}$. $\forall u \in U, \alpha(u) \in V_\alpha$, V_α means the value domain with respect to $\alpha \in A$.

Let C denote the conditional attribute set, and D denote the decision attribute set, if $A = C \cup D$ and $C \cap D = \emptyset$, then an information system $IS = \langle U, A, V \rangle$ is called decision table, denoted as $DT = \langle U, C \cap D, V \rangle$.

$Div(U) = \{U_i \mid U_i \subseteq U\}$ denotes the division of U, if and only if $\forall U_1, U_2 \in Div(U)$, $U_1 \cap U_2 = \emptyset$, $U = \cup_{\forall U_i \in Div(U)} U_i$. For $A' \subseteq A$, U/A' denotes a division of U with respect to A'. $U_i \in U/A'$ can be called an equivalence class with respect to A', if it satisfies $\forall U_i \in U/A'$, $\forall u, v \in U_i$, $\forall \alpha \in A'$, $\alpha(u) = \alpha(v)$. Therefore, U/A' is also known as an equivalence class cluster with respect to A'. Especially, $X = U/C$, $Y = U/D$ denotes an equivalence class cluster with respect to C, D, respectively.

Furthermore, for $\forall U_i, U_j \subseteq U$, $p(U_i) = \frac{|U_i|}{|U|}$ represents the probability. Similarly, $p(U_i \mid U_j) = \frac{|U_i \cap U_j|}{|U_j|}$ denotes the conditional probability, and $p(U_i, U_j) = \frac{|U_i \cap U_j|}{|U|}$ expresses the joint probability.

The feature selection process is described as follow: for $DT = \langle U, C \cup D, V \rangle$, $F_s \subseteq C$, we can obtain $DT' = \langle U, F_s \cup D, V \rangle$ which includes the same information as DT. It means that the accuracy of classifier constructed by DT' is not lower than that obtained by DT. The feature selection

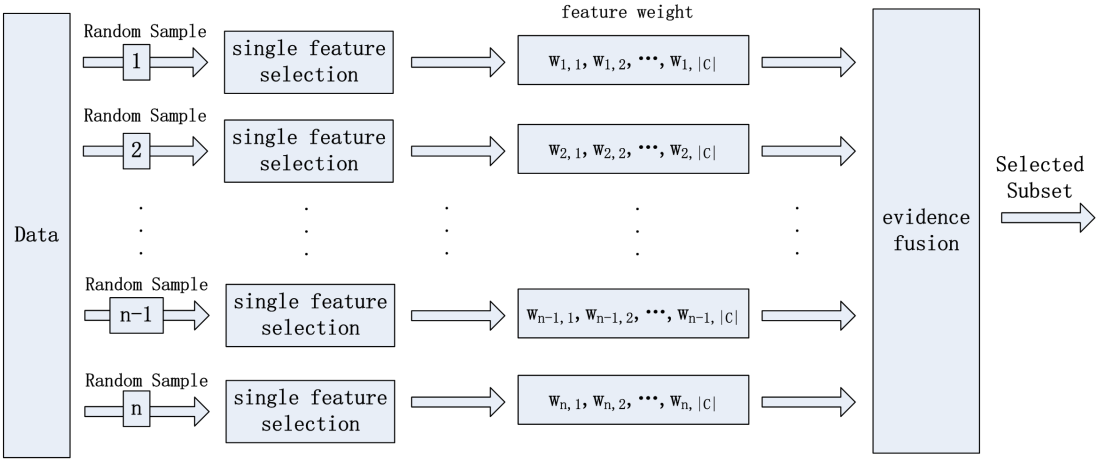


FIGURE 3. The process of TSDS algorithm.

process is related to two essential problems: how many features can be obtained and which features can be gained.

To retain the nature of the original data, we propose a novel feature selection ensemble learning algorithm based on TSDS. The process of TSDS is divided into two parts: 1) single feature selection model. 2) evidence fusion model, as shown in Fig. 3. In this section, we describe these two parts in detail.

A. SINGLE FEATURE SELECTION

For $DT = \langle U, C \cup D, V \rangle$, the feature set can be divided into three basic parts: 1) strong relevant features, 2) weak relevant features, and 3) irrelevant features. A feature is said to be relevant if it is predictive of the decision features; otherwise, it is irrelevant. A feature is considered to be redundant if it is highly correlated with other features. The characteristics of an informative feature are not only highly correlated with the decision features but also highly uncorrelated with other features. That means an optimal feature subset should include non-redundant features and strong relevant features. The correlation evaluation process of feature selection is shown in Fig.4.

1) CORRELATION MEASURE

In many feature selection approaches, Shannon Entropy [46] is employed to measure the degree of information uncertainty and quantify the amount of information contained in the dataset. With Shannon entropy, features with a high or low probability have equal weight in the entropy. Therefore, we utilize Tsallis entropy to evaluate feature importance which is an extension of the standard entropy [47]. The formula for Tsallis entropy is defined as

$$S_q = \frac{1}{q-1}\left(1 - \sum_{\forall U_i \subseteq Div(U)} p\left(U_i\right)^q\right), \quad (q \in R) \ and \ (q \neq 1) \tag{1}$$

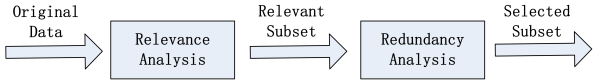


FIGURE 4. The feature correlation evaluation process.

where q is a positive parameter. When $q \to 1$, Tsallis entropy converges to Shannon entropy, and when $q = 2$, Tsallis entropy is equivalent to the Gini index [48].

With Tsallis entropy, for $q > 1$, features with high probability contribute more than that with low probability in the entropy value. Therefore, the higher is the value of q, the higher is the contribution of high probability events in the final result [49].

The main drawback of the approach based on information entropy is that biases towards the feature with more values. To making up the bias of information entropy, the symmetrical uncertainty is employed to estimate the degree of association between features and class labels [1].

In our proposed approach, the formula of symmetrical uncertainty based on Tsallis entropy is defined as

$$\begin{aligned} SU_q\left(X, Y\right) &= \frac{2 \times \left[S_q\left(X\right) + S_q\left(Y\right) - S_q\left(X, Y\right)\right]}{S_q\left(X\right) + S_q\left(Y\right)} \\ &= \frac{2 \times I_q\left(X; Y\right)}{S_q\left(X\right) + S_q\left(Y\right)}, \ SU_q \in [0, 1] \end{aligned} \tag{2}$$

where the formula for Tsallis mutual entropy $I_q\left(X; Y\right)$ [50]–[52] is defined as

$$\begin{aligned} I_q\left(X; Y\right) &= S_q\left(X\right) - S_q\left(X \mid Y\right) \\ &= S_q\left(X\right) + S_q\left(Y\right) - S_q\left(X, Y\right) \end{aligned} \tag{3}$$

where the $S_q\left(X \mid Y\right)$ and the $S_q\left(X, Y\right)$ are given by

$$S_q\left(X \mid Y\right) = -\sum_{\forall x \in X, \forall y \in Y} p\left(x, y\right)^q \log_q p\left(x \mid y\right), \quad q \neq 1 \tag{4}$$

$$S_q\left(X, Y\right) = -\sum_{\forall x \in X, \forall y \in Y} p\left(x, y\right)^q \log_q p\left(x, y\right), \quad q \neq 1 \tag{5}$$

When $SU_q = 0$, it indicates that X is independent of Y, in other words, X is irrelevant with respect to Y.

2) REDUNDANCY ANALYSIS

For $DT = \langle U, C \cup D, V \rangle$, $C_1, C_2 \subset C$ are conditionally independent for given class variable D, if $P(D \mid (C_1 \cup C_2)) \approx P(D \mid C_1)$. That means there is no additional information for C_1 when C_2 is added. Thus C_2 is redundancy.

Let $C' \subseteq C$, then a feature $c_i \in C'$ is redundancy if and only if c_i has a Markov blanket [53] in $C' - \{c_i\}$.

Definition 1 (Markov Blanket): Given a feature $c_i \in C$, let $C' \subseteq C$ denote a feature subset, then C' is said to be a Markov blanket for c_i if and only if $P\left(D \mid \left(C' \cup \{c_i\}\right)\right) = P\left(D \mid C'\right)$ with respect to $C - C' - \{c_i\}$.

According to Markov blanket, it is easy to obtain the redundant feature in the feature space. However, in case of high dimension and minuscule sample, the cardinality of the Markov blanket gives rise to overfitting [54]–[58].

In the proposed approach, to solve this problem, the Approximate Markov Blanket based on Tsallis entropy is utilized in forwarding sequential selection.

Definition 2 (Approximate Markov Blanket): Given two features $c_i, c_j \in C$ $(i \neq j)$, and a feature $d \in D$, c_i is an approximate Markov blanket for c_j, if and only if $SU_q(U/\{c_i\}, U/\{d\}) > SU_q\left(U/\left\{c_j\right\}, U/\{d\}\right)$ and $I_q\left(U/\{c_i\}; U/\left\{c_j\right\}\right) > I_q\left(U/\left\{c_j\right\}; U/\{d\}\right)$.

B. EVIDENCE FUSION

Dempster-Shafer evidence theory was firstly presented by Dempster [59], [60]. It is an effective uncertainty reasoning method to combine multiple information sources. The researches indicate that the synthetic consequence of conventional Dempster's combination rule is frequently contrary to the reality in the practical applications [61], [62]. Two major approaches are presented to enhance the accuracy of synthetic consequence. One is to amend the combination rule. The other is to alter the original evident resource. In this paper, we focus on the latter one.

1) CONFLICT MATRIX

The Minkowski distance (also called $l_p - norm$) [63] is utilized to construct the conflict matrix between evidence. According to the intension of Minkowski distance, the formula is redefined based on the information system.

Definition 3 (Minkowski Distance): For $IS = \langle U, A, V \rangle$, $z, w \in U$. $\forall \alpha \in A$, $\alpha(u) \in \mathbb{R}$, the Minkowski distance is defined as

$$Dis(z, w) = \left(\sum_{z,w \in U, \forall \alpha \in A} |\alpha(z) - \alpha(w)|^{\varsigma} \right)^{\frac{1}{\varsigma}} \tag{6}$$

when $\varsigma = 1$ or $\varsigma = 2$, the Minkovski distance corresponds to the Manhattan distance(also called $l_1 - norm$) [63] or the Euclidean distance(also called $l_2 - norm$) [63], respectively.

Let n denote the number of evidence (also represents the execution times for single feature selection), m_i and m_j $(1 \leq i, j \leq n)$ represent two evidence. Integrated with feature selection, let $\eta = |\ C\ |$, each evidence is denoted by $m_i = \left(m_{i1}, m_{i2}, \cdots, m_{i\eta}\right)$. The Minkowski distance can be utilized to calculate the conflict distance between m_i and m_j as

$$mc_{ij} = \left(\sum_{\gamma=1}^{\gamma=|\eta|} \left| m_{i\gamma} - m_{j\gamma} \right|^{\varsigma} \right)^{\frac{1}{\varsigma}} \tag{7}$$

Then the normalization conflict matrix is defined as (or simply MC for short)

$$Matrix_{conflict} = \begin{bmatrix} 0 & mc_{12} & \cdots & mc_{1j} & \cdots & mc_{1n} \\ mc_{21} & 0 & \cdots & mc_{2j} & \cdots & mc_{2n} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ mc_{j1} & mc_{j2} & \cdots & 0 & \cdots & mc_{jn} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ mc_{n1} & mc_{n2} & \cdots & mc_{nj} & \cdots & 0 \end{bmatrix} \tag{8}$$

2) SUPPORT DEGREE

Evidence support degree indicates the support degree of an evidence which is supported by other evidence. The higher similarity with other evidence, the higher support degree it is, and vice versa. According to $Matrix_{conflict}$, the following formula is utilized to calculate the similarity degree between m_i and m_j.

$$s_{ij} = 1 - mc_{ij}, \quad i, j = 1, 2, \cdots, n \tag{9}$$

As a result, we can obtain the following similarity matrix of all evidence (or simply MS for short)

$$Matrix_{similarity} = \begin{bmatrix} 1 & s_{12} & \cdots & s_{1j} & \cdots & s_{1n} \\ s_{21} & 1 & \cdots & s_{2j} & \cdots & s_{2n} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ s_{j1} & s_{j2} & \cdots & 1 & \cdots & s_{jn} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ s_{n1} & s_{n2} & \cdots & s_{nj} & \cdots & 1 \end{bmatrix} \tag{10}$$

And then, the support degree of each evident is calculated as

$$Sup(m_i) = \sum_{j=1, j \neq i}^{j=n} s_{ij} \tag{11}$$

3) EVIDENT WEIGHT

Credibility degree indicates the credibility of an evidence. It can be calculated by following formula.

$$CR(m_i) = \frac{Sup(m_i)}{\sum_{j=1}^{n} Sup\left(m_j\right)}, \quad i, j = 1, 2, \cdots, n \tag{12}$$

Information entropy can be utilized to measure the informative quantity of evidence in the information fusion process. Integrated with Dempster-Shafer theory, given an evidence

$m_i = (m_{i1}, m_{i2}, \cdots, m_{i\eta})$, and $\sum_{l=1}^{\eta} m_{il} = 1$. The information quantity of the *ith* evidence is defined as

$$Info_{S_q}(m_i) = \frac{1}{q-1}\left(1 - \sum_{l=1}^{\eta} m_{il}^q\right) \tag{13}$$

For information entropy, the larger the uncertainty, the smaller the weight it is. On the other hand, the smaller the information entropy, the larger the weight it is. The method mentioned above can be used to reduce the weight ratio of the evidence with higher indeterminacy in the fusion process. Therefore, the weight of each evidence is defined as

$$weight(m_i) = \frac{CR(m_i)}{Normalization\left(Info_{S_q}(m_i)\right)}, \quad 1 \le i \le n \tag{14}$$

$$m_i = weight(m_i) \times m_i, \quad 1 \le i \le n \tag{15}$$

4) EVIDENCE COMBINATION RULE

Suppose that the feature subsets generated in the previous chapter are independent, TSDS allows the fusion of information coming from different feature subsets. Therefore, the evidence combination rule is utilized to combine different weighted feature subsets in a manner that is both accurate and robustness.

Given $DT = \langle U, C \cup D, V \rangle$, $m_i (i = 1, 2, \cdots n)$, for $\forall c \in C$, the combination rule is redefined as

$$(m_1 \oplus m_2 \oplus \cdots \oplus m_n)(c) = \frac{1}{K}\prod_{i=1}^{n} m_i(c) \tag{16}$$

where K means the conflict between different pieces of evidence, is given by

$$K = \sum_{i=1}^{\eta=|C|} \prod_{j=1}^{n} m_j(c_i) \tag{17}$$

C. TSALLIS ENTROPY AND DEMPSTER-SHAFER (TSDS) ALGORITHM

Feature selection focus on obtaining the approximate optimal feature subset which can preserve the discrimination ability of the original data. The pseudo-code of our proposed algorithm of feature selection with ensemble learning is elaborated in Algorithm1, which is divided into two major parts. One is to select the feature subset. Each execution of the body of the feature selection is an iteration. It adjusts the parameter q of Tsallis entropy automatically in each iteration. Because each iteration is independent, it guarantees that each feature subset is diversity (the detailed is shown in Algorithm2). The other is to fuse $F_{evidence}$ which obtained from the previous task (the detailed is shown in Algorithm3). We can fuse each evidence according to the equation (16) after the weight of each evidence has been calculated. Furthermore, normalization of feature weight can ensure computational efficiency.

Algorithm 1 TSDS_Feature_Selection

Input: U: a set of samples, φ: the times of single feature selection
Output: F_{final}: the final feature subset

for $i = 1$ to φ **do**
  $U_i \leftarrow$random sample from U
  $C_i \leftarrow GetFeatues(U_i)$
  $D_i \leftarrow GetClass(U_i)$
  $F_{single} \leftarrow single\_feature\_selection(C_i, D_i)$
  $F_{single}.weight \leftarrow Normalization\left(F_{single}.weight\right)$
  $F_{evidence} \leftarrow F_{evidence} \cup \{F_{single}\}$
end for
$F_{final} \leftarrow evidence\_fusion(F_{evidence})$
$\delta \leftarrow$ calculate the average weight of F_{final}
for each feature f in F_{final} **do**
  if $f_{weight} < \delta$ **then**
    $F_{final} \leftarrow F_{final} - \{f\}$
  end if
end for
return F_{final}

Algorithm 2 Single_Feature_Selection

Input: C: the original feature set; D:the class label set
Output: F_{single}: the feature set of single feature selection

$F_{set} \leftarrow \{\emptyset\}, F_{single} \leftarrow \{\emptyset\}$
for each feature c in C **do**
  $Weight_{su} \leftarrow SU_q(\{c\}, D)$, according to equation (2)
  $f.weight \leftarrow Weight_{su}$
  $F_{set} \leftarrow F_{set} \cup \{c\}$
end for
$\sigma \leftarrow$ calculate the average weight of F_{set}
for each feature f in F_{set} **do**
  if $f.weight < \sigma$ **then**
    $F_{set} \leftarrow F_{set} - \{f\}$
  end if
end for
order F_{set} in descending by weight
$f_{first} \leftarrow GetFirst(F_{set})$
$F_{single} \leftarrow F_{single} \cup \{f_{first}\}$
for each feature f_{set} in F_{set} **do**
  $flag \leftarrow 0$
  for each feature f_{single} in F_{single} **do**
    if $I_q(\{f_{single}\}, \{f_{set}\}) > I_q(\{f_{set}\}, D)$ **then**
      $flag \leftarrow 1$
      break;
    end if
  end for
  if $flag \neq 1$ **then**
    $F_{single} \leftarrow F_{single} \cup \{f_{set}\}$
  end if
end for
return F_{single}

Algorithm 3 Evidence_Fusion

Input: $F_{evidence}$: the feature union, each element is a feature set

Output: F_{fusion}: the final feature set of evidence fusion

$mass_{set} \leftarrow F_{evidence}$
$n_{mass} \leftarrow$ the element number of $F_{evidence}$
$MC \leftarrow$ according to equation (7), use $mass_{set}$ to calculate conflict matrix
$MS \leftarrow$ according to equation (9), use MC to calculate similarity matrix
for each mass m in $mass_{set}$ **do**
    $Sup_m \leftarrow$ according to equation (11), use MC to calculate support degree
    $Sup_{set} \leftarrow Sup_{set} \cup \{Sup_m\}$
end for
$Sup_{total} \leftarrow \sum_{i=1}^{n} Sup_{set}(m_i)$
for each mass m in $mass_{set}$ **do**
    $CR_m \leftarrow \frac{Sup_{set}(m)}{Sup_{total}}$
    $q_m \leftarrow Normalization\left(Info_{S_q}(m)\right)$
    $weight_m \leftarrow \frac{CR_m}{q_m}$
    $m.weight \leftarrow m.weight \times weight_m$
end for
$F_{fusion} \leftarrow$ utilize equation (16) to combination $mass_{set}$
return F_{fusion}

TABLE 1. Data description.

Data Set	Number of Instances	Number of Attributes	Data Sources
Chess	3196	36	UCI
Connect	67557	42	UCI
Lung Cancer	32	56	UCI
Lymphography	148	18	UCI
Promoters	106	58	UCI
Spect	267	22	UCI
Splice	3190	61	UCI
Colon	62	2000	ASU
Leukemia	72	7070	ASU

IV. EXPERIMENTAL ANALYSIS

In this section, we make a comparison of the proposed method with the other existing approaches. Three different classification algorithms are employed to evaluate the performance of all feature selection methods. The classifiers include Support Vector Machine (SVM), Decision Tree (CART), and Bayes. Nine datasets are used in the experimental analysis. These datasets are divided into two classes: seven standard datasets from the University of California Irvine (UCI) Machine Learning Repository and two gene expression datasets with high dimension and minuscule sample from Arizona State University (ASU), as shown in Table 1. The goodness of given approaches cannot be only measured in terms of the improvement for the average classification accuracy. Therefore, we utilize the Friedman test [64], [65] and Contrast Estimation [66], [67] to evaluate the significant differences between different algorithms. All algorithms are executed in Python and run in the hardware environment with Core i7-7500 2.7GHz and 32.0GB RAM.

A. PARAMETER SETTING

To validate the effectiveness of TSDS algorithm, we compare our algorithm with the following seven feature selection approaches:

1) DISR [26]: It relies on a measure of variable complementarity to evaluate the additional information. DISR criterion combines two properties of feature selection. One is feature complementarity, which means that a combination of feature can obtain more information than the sum returned by each feature individually. The other is the computation of a lower-bound on the information of a feature subset expressed as the average of information of all its subaggregate.
2) CMIM [27]: It mainly utilizes Maximization conditional mutual information. The feature that can obtain additional information about the predicted class is selected. In other words, in the process of CMIM, it does not select a feature similar to each one which has been picked to the selected feature subset. CMIM takes the tradeoff between independence and discrimination into account.
3) JMI [28]: It is the model-independent approach for feature selection based on joint mutual information. It is found to be better in eliminating redundancy than simple mutual information.
4) MRMR [29]: It obtains feature subset by utilizing minimal redundancy and maximal relevance measure criterion. This scheme avoids the difficult multivariate density estimation in maximizing dependency. Meanwhile, MRMR can be combined with other evaluation criterion such as wrapper to obtain a very compact feature subset at a lower cost.
5) SPFS-LAR [30]: SPFS-LAR utilizes similarity preserving feature selection framework to preserve feature. The regularized sparse multiple-output regression formulation is used in this framework to enhance its effectiveness. The advantage of SPFS-LAR is that it does not require parameter tuning in the feature selection process.
6) QIS [31]: QIS pay attention to the distinguish ability of each feature which can be used to distinguish a given sample with other samples. The maximum-nearest-neighbor is employed to discriminate the nearest neighbors of samples. To address the problem of neighborhood parameter selection, the margin of the sample is utilized to set the neighborhood parameter value.
7) HANDI [32]: When adding a new feature to the current feature subset, HANDI utilizes the conditional discrimination index to calculate the increment of distinguishing information. The proposed discrimination index is computed by the cardinality of a neighborhood relation.

To compare with the above algorithms, there are some parameters to be predefined in terms of original papers. The first four algorithms are related to information entropy, thus

TABLE 2. Classification accuracies (%) of classifiers with TSDS in different δ value.

Data Set	CART			SVM			Bayes		
	Third Quartile	First Quartile	Mean Value	Third Quartile	First Quartile	Mean Value	Third Quartile	First Quartile	Mean Value
Chess	0.9412	**0.9621**	0.9531	0.9393	0.9496	**0.9513**	0.6728	0.7381	**0.8325**
Promoters	0.7636	0.7454	**0.7980**	0.7909	0.7727	**0.8283**	0.8272	0.7818	**0.8727**
Spect	0.7851	0.7925	**0.8222**	0.8037	0.8	**0.8287**	**0.7962**	0.7889	0.7889
Splice	0.8267	0.8670	**0.9173**	0.8229	0.8811	**0.9213**	0.8253	0.8686	**0.9314**
Average	0.8292	0.8418	**0.8727**	0.8392	0.8509	**0.8824**	0.7804	0.7944	**0.8564**

TABLE 3. Classification accuracies (%) of CART with different feature selection algorithms.

Data Set	TSDS	DISR	CMIM	JMI	MRMR	SPFS-LAR	QIS	HANDI
Chess	0.9531	0.9678	0.9584	0.9603	0.9563	0.9656	0.9415	**0.9884**
Connect	**0.6854**	0.6753	0.6726	0.6726	0.6689	0.6737	-	-
Lung Cancer	0.7500	0.7250	**0.8000**	0.7000	0.7250	0.7500	0.6500	0.7000
Lymphography	**0.7733**	0.7500	0.7067	0.7467	0.7666	0.7266	0.7067	0.7067
Promoters	0.7980	0.7455	0.7636	0.7545	0.7636	0.7455	**0.8636**	0.8364
Spect	**0.8222**	0.7519	0.7704	0.7741	0.7889	0.7704	0.7889	0.8148
Splice	0.9173	0.9148	**0.918**	0.9182	0.9173	0.9151	**0.9182**	0.7371

no additional parameter setting is needed. Since SPFS-LAR is related to similarity preserving framework, it does not need additional parameters. For QIS, the neighborhood size is set as 0.1. According to the original paper, the parameter is set as 0.001 for low-dimensional data and 0.01 for high-dimensional data for HANDI.

In the experiment process, TSDS only reserves those features whose weights are greater than δ. Table 2 shows the comparison results of different thresholds δ in the accuracy. The experiments indicate that each classifier can gain the highest average classification accuracy when the threshold δ is set as the mean value. Thus, the mean value is adopted as δ.

B. EXPERIMENTS ON LOW-DIMENSIONAL DATASETS

To prove the effectiveness of feature selection, the experiments have been performed on the seven UCI datasets to compare the performance of classifiers constructed by dataset with TSDS, DISR, CMIM, JMI, MRMR, SPFS-LAR, QIS, and HANDI. Need to pay attention to it, neither QIS nor HANDI can process the Connect dataset with running out of memory. Therefore, these algorithms are divided into two parts: 1) DISR, CMIM, JMI, MRMR, and SPFS-LAR. 2) QIS, and HANDI. For each dataset, the 10-fold cross-validation is utilized to estimate the classification accuracy. Meanwhile, to make the comparative analysis between different algorithms more balanced, this process is repeated ten times for each dataset. Furthermore, to evaluate the performance of different feature selection algorithms in an experiment, we introduce three metrics including classification accuracy, Contrast Estimation, and Friedman test.

A procedure for Contrast Estimation assumes that the expected differences between performance of different algorithms are the same across datasets [66], [67]. In this paper, for every pair of eight algorithms in experiment, the formula which is utilized to calculate the difference between the performances of the two algorithms in each of the seven datasets is given as

$$D_{i_{(\mu\nu)}} = \kappa_{i\mu} - \kappa_{i\nu} \tag{18}$$

where $i = 1, \cdots, N_{data}$ represent the index of the datasets, $\mu, \nu = 1, \cdots, N_{algorithm}$ represent the index of the algorithms. In this paper, we set $N_{data} = 7$ and $N_{algorithm} = 8$, respectively.

The difference between two algorithms is computed by $\omega_\mu - \omega_\nu$, ω_μ is given as

$$\omega_\mu = \frac{\sum_{j=1}^{|N_{algorithm}|} Z_{\mu j}}{\mid N_{algorithm} \mid} \tag{19}$$

where $Z_{\mu j}$ denotes the median of each set of differences.

Friedman test aims to detect significant differences between the performances of two or more algorithms [64], [65]. It calculates the ranking of the observed results for the given algorithms, and then order by descending order. The Friedman test is distributed according to λ_F^2 with $k-1$ degrees of freedom (k denote the algorithms), is given as

$$\Lambda_F^2 = \frac{12 \times N}{k \times (k+1)} \left[\sum_j R_j^2 - \frac{k \times (k+1)^2}{4} \right] \tag{20}$$

where N represents the number of datasets, $R_j = \sum_i r_i^j$ (r_i^j denote the algorithm j with k algorithms).

To facilitate experimental comparison, the number of features of other approaches is uniformly set to that obtained by TSDS. Some experiments have been conducted to verify the TSDS algorithm. The detail of experiments in different algorithms are shown in Table 3-5 (boldface represents the highest accuracy for each dataset in the classification algorithm). The details illustrate that TSDS can obtain the highest accuracy for 3,4,3 datasets in CART, SVM, Bayes, respectively. Fig.5 presents TSDS can achieve the highest average classification accuracy among the former ones in the CART, SVM, and Bayes, respectively.

TABLE 4. Classification accuracies (%) of SVM with different feature selection algorithms.

Data Set	TSDS	DISR	CMIM	JMI	MRMR	SPFS-LAR	QIS	HANDI
Chess	0.9513	0.9541	0.9519	0.9516	0.9469	0.9518	0.9453	**0.9728**
Connect	**0.6835**	0.6753	0.6726	0.6726	0.6707	0.6721	-	-
Lung Cancer	**0.8250**	**0.8250**	0.6750	0.4750	0.7750	0.6125	0.5250	0.5750
Lymphography	**0.8267**	0.7667	0.7800	0.7667	0.7933	0.7867	0.7733	0.8266
Promoters	0.8283	0.8455	0.8273	0.7909	0.8000	0.8091	**0.8636**	0.8455
Spect	**0.8287**	0.8074	0.8259	0.8185	0.8148	0.7815	0.8148	0.7777
Splice	0.9213	0.9126	0.9179	0.9119	0.9142	0.9035	**0.9233**	0.7607

TABLE 5. Classification accuracies (%) of Bayes with different feature selection algorithms.

Data Set	TSDS	DISR	CMIM	JMI	MRMR	SPFS-LAR	QIS	HANDI
Chess	**0.8325**	0.6284	0.6650	0.8038	0.6300	0.7847	0.6316	0.6716
Connect	0.6503	0.6360	0.6571	0.6571	**0.6630**	0.6613	-	-
Lung Cancer	**0.7250**	0.7000	0.6500	**0.7250**	0.7000	0.5750	0.6000	0.7000
Lymphography	0.7933	0.7067	0.6733	0.7333	0.7733	**0.8067**	0.7600	0.7600
Promoters	0.8727	0.8727	0.8727	**0.8909**	0.8727	0.8727	0.8364	0.8273
Spect	**0.7889**	0.5074	0.4778	0.6222	0.4963	0.7814	0.7148	0.7519
Splice	0.9314	0.9311	0.9340	**0.9434**	0.9318	0.9320	0.9091	0.7406

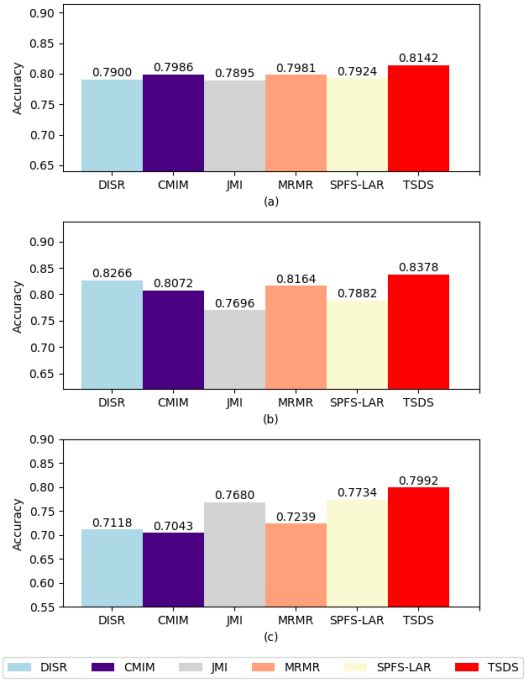


FIGURE 5. Average accuracy comparison with different feature selection algorithms for (a) CART, (b) SVM, (c) Bayes.

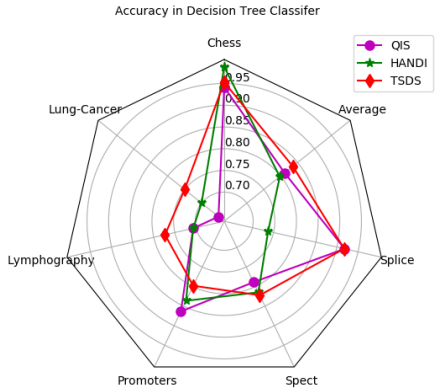


FIGURE 6. Classification performance comparison with QIS and HANDI in CART.

TABLE 6. Contrast estimation of the algorithms with six datasets in CART.

	TSDS	DISR	CMIM	JMI	MRMR	SPFS-LAR
TSDS	0	0.02	0.017	0.021	0.016	0.018
DISR	-0.02	0	-0.003	0.001	-0.004	-0.002
CMIM	-0.017	0.003	0	0.004	-0.001	0.001
JMI	-0.021	-0.001	-0.004	0	-0.005	-0.003
MRMR	-0.016	0.004	0.001	0.005	0	0.002
SPFS-LAR	-0.018	0.002	-0.001	0.003	-0.002	0

The performance analysis of a new method is a crucial task to carry out in research. Furthermore, Contrast Estimation and Friedman Test are used to estimate the performance of the given algorithms. Contrast Estimation based on medians can be employed to evaluate the performance difference between two algorithms. Friedman test is a multiple comparison test approach which is employed to detect the significant differences between two or more algorithms. Thus, Contrast Estimation is employed to estimate the differences between TSDS and the former ones. The results in Table 6-8 demonstrate that TSDS can outperform the former ones in terms of classification performance. Then, in the following experiment, Friedman test is used to estimate TSDS and the latter ones. In Table 9, the statistical analysis results illustrate that TSDS would be more effective than QIS and HANDI. The detailed experimental comparisons are shown in Fig. 6-8. Meanwhile, from Fig. 6-8, it is obvious that TSDS can obtain better overall performance than the latter ones.

Conventional feature selection algorithms execute sampling only once, which cannot maintain the diversity of samples. On the contrary, from the above descriptions, with multiple ensemble learning based on random sampling,

TABLE 7. Contrast estimation of the algorithms with six datasets in SVM.

	TSDS	DISR	CMIM	JMI	MRMR	SPFS-LAR
TSDS	0	0.008	0.006	0.013	0.011	0.017
DISR	-0.008	0	-0.002	0.006	0.004	0.01
CMIM	-0.006	0.002	0	0.007	0.006	0.012
JMI	-0.013	-0.006	-0.007	0	-0.002	0.004
MRMR	-0.011	-0.004	-0.006	0.002	0	0.006
SPFS-LAR	-0.017	-0.01	-0.012	-0.004	-0.006	0

TABLE 8. Contrast estimation of the algorithms with six datasets in Bayes.

	TSDS	DISR	CMIM	JMI	MRMR	SPFS-LAR
TSDS	0	0.033	0.044	0.001	0.025	0.017
DISR	-0.033	0	0.012	-0.032	-0.007	-0.016
CMIM	-0.044	-0.012	0	-0.043	-0.019	-0.027
JMI	-0.001	0.032	0.043	0	0.024	0.016
MRMR	-0.025	0.007	0.019	-0.024	0	-0.008
SPFS-LAR	-0.017	0.016	0.027	-0.016	0.008	0

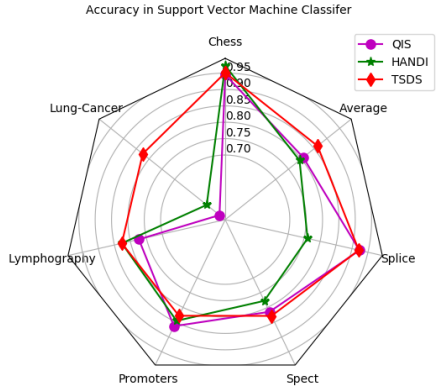


FIGURE 7. Classification performance comparison with QIS and HANDI in SVM.

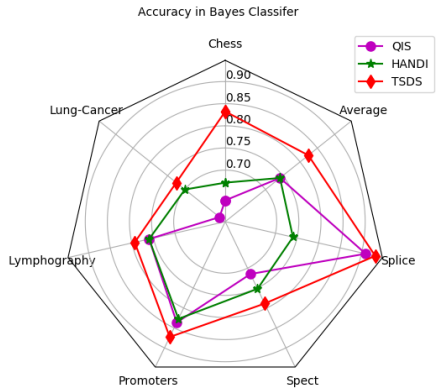


FIGURE 8. Classification performance comparison with QIS and HANDI in Bayes.

TSDS can maintain the diversity of samples effectively. After feature subsets generated, it utilizes the ensemble fusion strategy based on Dempster-Shafer evidence theory to ensemble all feature subsets. The above experiments prove that TSDS can enhance the classification performance effectively in low-dimensional datasets.

TABLE 9. Average rankings by applying the Friedman procedure.

Algorithm	CART Ranking	SVM Ranking	Bayes Ranking
TSDS	1.6667	1.6667	1
HANDI	2.0833	2.1667	2.4167
QIS	2.25	2.1667	2.5833

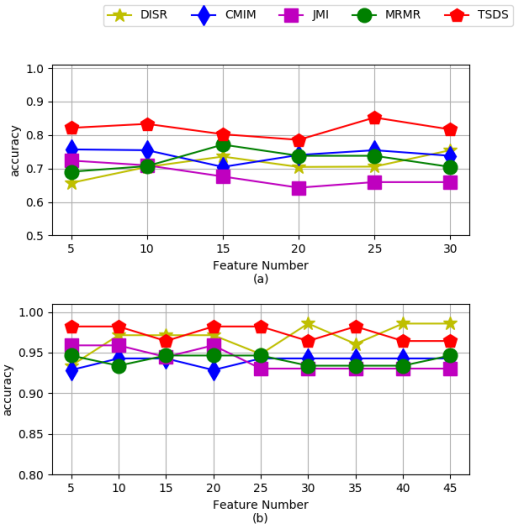


FIGURE 9. Classification accuracy vs. number of the selected feature subset in CART for (a) Colon DataSet, (b) Leukemia DataSet.

C. EXPERIMENTS ON HIGH-DIMENSIONAL DATASETS

To illustrate the scalability of TSDS, another series of experiments are performed. Two datasets, Colon and Leukemia from different application domains are employed in this part. In the experiments, these two datasets contain thousands of features, and many of them could be highly correlated with others, as shown in Table 1. Similarly, we employ the 10-fold cross-validation to evaluate the classification accuracy in all datasets, and the process is repeated ten times for each dataset to ensure the comparability. According to the intension of feature selection approaches, approaches for comparison are divided into two categories: 1) DISR, CMIM, JMI, MRMR. 2)SPFS-LAR, QIS, and HANDI. Then, the TSDS algorithm will compare with these two categories of feature selection algorithms in detail, respectively.

To evaluate the classification performance of different feature selection algorithms, a comparison of the feature selection algorithms for Colon and Leukemia in the CART, SVM and Bayes are shown in Fig.9-11. The numbers of the selected feature subset are taken as {5, 10, 15, 20, 25, 30} for Colon. The numbers of the selected feature subset are taken as {5, 10, 15, 20, 25, 30, 35, 40, 45} for Leukemia. We can see from Fig. 9-11 that the classification results of different classifiers based on TSDS are in general better than the former ones.

Fig.12-14 compare the performances of TSDS with HANDI, QIS, and SPFS-LAR in different classifiers. The results illustrate that TSDS can achieve superior performance with the same feature numbers given by HANDI, QIS,

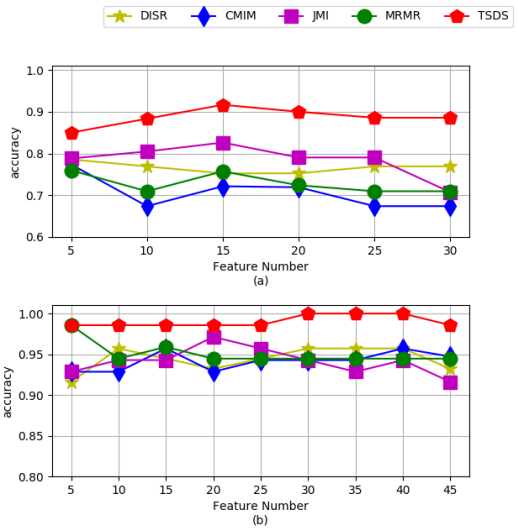


FIGURE 10. Classification accuracy vs. number of the selected feature subset in SVM for (a) Colon DataSet, (b) Leukemia DataSet.

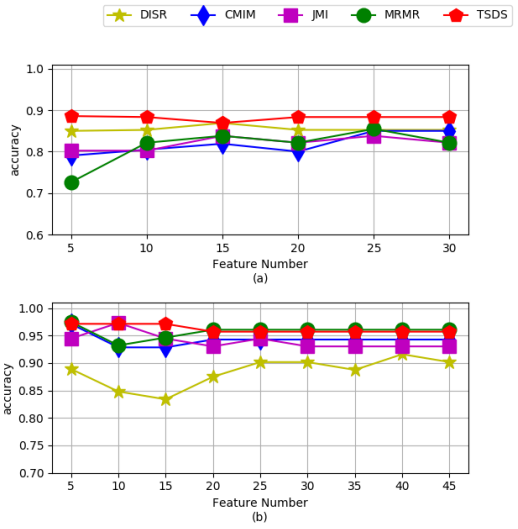


FIGURE 11. Classification accuracy vs. number of the selected feature subset in Bayes for (a) Colon DataSet, (b)Leukemia DataSet.

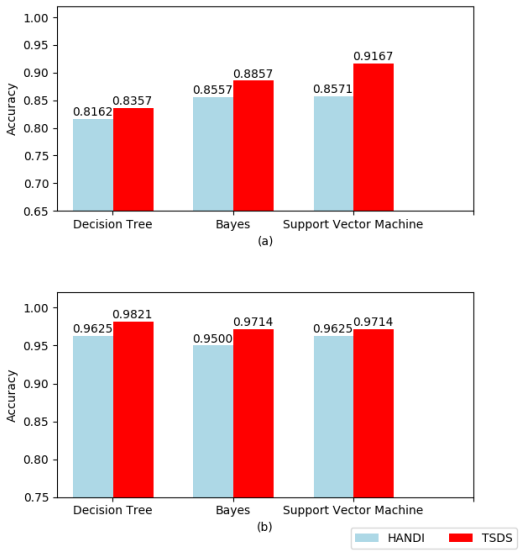


FIGURE 12. Classification accuracy compare with HANDI high-dimensional datasets for (a) Colon DataSet, (b) Leukemia DataSet.

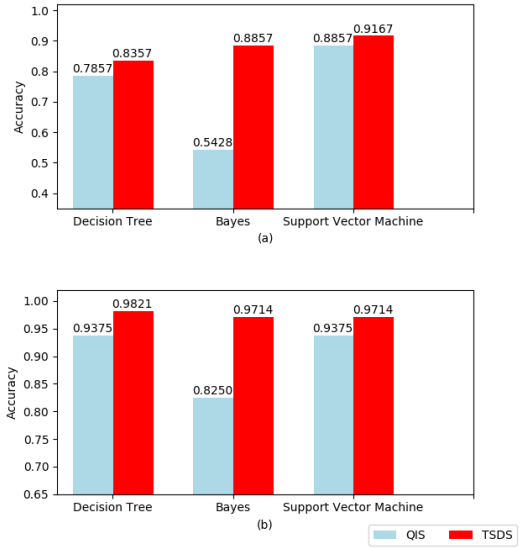


FIGURE 13. Classification accuracy compare with QIS high-dimensional datasets for (a) Colon DataSet, (b) Leukemia DataSet.

and SPFS-LAR. From the above three figures, we have the following observations. For Colon dataset, in comparison to HANDI, our proposed TSDS algorithm can improve the classification accuracy by 0.019, 0.03 and 0.059 for CART, Bayes and SVM, respectively. Comparing with QIS, the accuracy can be enhanced by 0.05, 0.342 and 0.03 for CART, Bayes and SVM, respectively. While comparing with SPFS-LAR, the improvement of the accuracy is 0.035, 0.085 and 0.102 for CART, Bayes and SVM, respectively. Similarly, for the case of Leukemia dataset, TSDS can outperform HANDI, QIS and SPFS-LAR in terms of classification accuracy.

A goodness of feature selection algorithm may enhance the performance of classification tasks in that it can eliminate the redundant features and irrelevant features effectively, and highlight the relevant informative features. However, high-dimensional datasets include excessive redundant and irrelevant features. The above experiments demonstrate that TSDS can outperform the compared feature selection algorithms by utilizing the improved symmetrical uncertainty based on Tsallis entropy and forward sequential with approximate Markov blanket. Meanwhile, these consequences indicate that TSDS can obtain relative optimal feature subset more stable than other comparing feature selection algorithms on high-dimensional datasets by adopting ensemble fusion strategy with Dempster-Shafer evidence theory. All in all, TSDS can achieve competitive performance compared to existing state-of-the-art feature selection approaches mentioned.

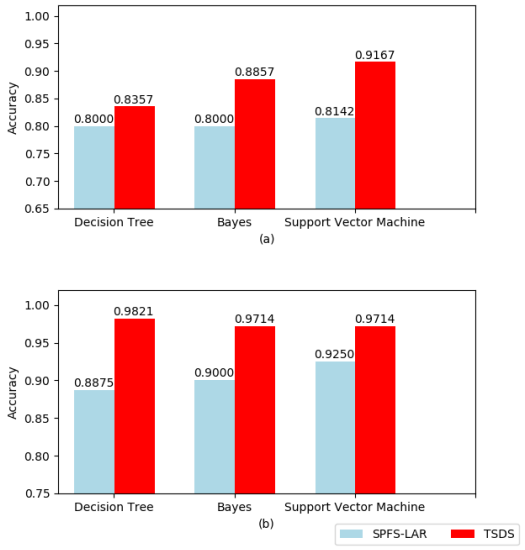


FIGURE 14. Classification accuracy compare with SFPS-LAR high-dimensional datasets for (a) Colon DataSet, (b) Leukemia DataSet.

V. CONCLUSIONS

Feature selection is one of the data optimization techniques in machine learning and data mining. If the dataset can be compressed effectively by utilizing feature selection, we can obtain potential valuable information, and further improve the performance of the classifier models. In this paper, a novel feature selection ensemble learning algorithm is proposed based on evidence theory. The key contribution of TSDS is to improve correlation criterion and ensemble fusion strategy. The improved symmetrical uncertainty based on Tsallis entropy and forward sequential approximate Markov blanket help us to gain more relevant informative features in the original feature space. Ensemble fusion strategy makes an algorithm better to obtain optimal global feature selection approximately. A study of all experiments demonstrates that TSDS can obtain relative optimal feature subset to construct a classifier, which can gain a higher classified accuracy than other comparing feature selection algorithms. To select the threshold to control the number of selected feature effectively and automatically and to enhance fusion efficiency will be the focus of further research.

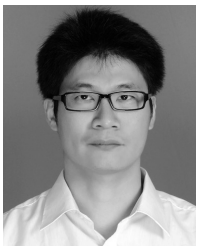
ACKNOWLEDGMENT

The authors would like to thank colleagues at the Data and Knowledge Engineering Center, School of Information Technology and Electrical Engineering, University of Queensland. They would also like to thank Prof. X. Zhou for his special suggestions and many interesting discussions.

REFERENCES

[1] Y. Li, T. Li, and H. Liu, "Recent advances in feature selection and its applications," *Knowl. Inf. Syst.*, vol. 53, no. 3, pp. 551–577, Dec. 2017.
[2] J. Tang, S. Alelyani, and H. Liu, "Feature selection for classification: A review," in *Data Classification: Algorithms and Applications*. Boca Raton, FL, USA: CRC Press, 2014.
[3] J. Miao and L. Niu, "A survey on feature selection," *Procedia Comput. Sci.*, vol. 91, pp. 919–926, 2016.
[4] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Comput. Elect. Eng.*, vol. 40, no. 1, pp. 16–28, Jan. 2014.
[5] H. Almuallim and T. G. Dietterich, "Learning with many irrelevant features," in *Proc. Nat. Conf. Artif. Intell.*, 1991, vol. 66, no. 4, pp. 547–552.
[6] F. Yue and W. Zuo, "Consistency analysis on orientation features for fast and accurate palmprint identification," *Inf. Sci.*, vol. 268, pp. 78–90, Jun. 2014.
[7] I. Kononenko, "Estimating attributes: Analysis and extensions of RELIEF," in *Proc. Eur. Conf. Mach. Learn.*, 1994, pp. 82–171.
[8] W. Zhu, G. Si, Y. Zhang, and J. Wang, "Neighborhood effective information ratio for hybrid feature subset evaluation and selection," *Neurocomputing*, vol. 99, pp. 25–37, Jan. 2013.
[9] A. Al-Ani, "A dependency-based search strategy for feature selection," *Expert Syst. Appl.*, vol. 36, no. 10, pp. 12392–12398, Dec. 2009.
[10] Y. Lin, J. Li, P. Lin, G. Lin, and J. Chen, "Feature selection via neighborhood multi-granulation fusion," *Knowl.-Based Syst.*, vol. 67, pp. 162–168, Sep. 2014.
[11] H. Zhao and K. Qin, "Mixed feature selection in incomplete decision table," *Knowl.-Based Syst.*, vol. 57, pp. 181–190, Feb. 2014.
[12] N. Hoque, D. K. Bhattacharyya, and J. K. Kalita, "MIFS-ND: A mutual information-based feature selection method," *Expert Syst. Appl.*, vol. 41, no. 14, pp. 6371–6385, 2014.
[13] Y. Lin, X. Hu, and X. Wu, "Quality of information-based source assessment and selection," *Neurocomputing*, vol. 133, pp. 95–102, Jun. 2014.
[14] W. Qian and W. Shu, "Mutual information criterion for feature selection from incomplete data," *Neurocomputing*, vol. 168, pp. 210–220, Nov. 2015.
[15] M. Wei, T. W. S. Chow, and R. H. M. Chan, "Heterogeneous feature subset selection using mutual information-based feature transformation," *Neurocomputing*, vol. 168, pp. 706–718, Nov. 2015.
[16] Y. Li, S. Y. Gao, and S. C. Chen, "Ensemble feature weighting based on local learning and diversity," in *Proc. AAAI Conf. Artif. Intell.*, Toronto, ON, Canada, 2012, pp. 1019–1025.
[17] V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos, "An ensemble of filters and classifiers for microarray data classification," *Pattern Recognit.*, vol. 45, no. 1, pp. 531–539, 2012.
[18] W. Awada, T. M. Khoshgoftaar, D. Dittman, and R. Wald, "The effect of number of iterations on ensemble gene selection," in *Proc. 11th Int. Conf. Mach. Learn. Appl. (ICMLA)*, Boca Raton, FL, USA, Dec. 2012, pp. 198–203.
[19] D. Guan, W. Yuan, Y.-K. Lee, K. Najeebullah, and M. K. Rasel, "A review of ensemble learning based feature selection," *IETE Tech. Rev.*, vol. 31, no. 3, pp. 190–198, 2014.
[20] D. W. Opitz, "Feature selection for ensembles," in *Proc. 16th Nat. Conf. Artif. Intell. (AAAI)*, Orlando, FL, USA, Jul. 1999, pp. 379–384.
[21] A. Tsymbal, S. Puuronen, and D. W. Patterson, "Ensemble feature selection with the simple Bayesian classification information fusion," *Inf. Fusion*, vol. 4, no. 2, pp. 87–100, 2003.
[22] A. Tsymbal, M. Pechenizkiy, and P. Cunningham, "Diversity in search strategies for ensemble feature selection," *Inf. Fusion*, vol. 6, no. 1, pp. 83–98, 2005.
[23] Y. Kessentini, T. Burger, and T. Paquet, "A Dempster–Shafer theory based combination of handwriting recognition systems with multiple rejection strategies," *Pattern Recognit.*, vol. 48, no. 2, pp. 534–544, Feb. 2015.
[24] H. Jiang, R. Wang, J. Gao, Z. Gao, and X. Gao, "Evidence fusion-based framework for condition evaluation of complex electromechanical system in process industry," *Knowl.-Based Syst.*, vol. 124, pp. 176–187, May 2017.
[25] M. Beynon, B. Curry, and P. Morgan, "The Dempster–Shafer theory of evidence: an alternative approach to multicriteria decision modelling," *Omega*, vol. 28, no. 1, pp. 37–50, Feb. 2000.
[26] P. E. Meyer, C. Schretter, and G. Bontempi, "Information-theoretic feature selection in microarray data using variable complementarity," *IEEE J. Sel. Topics Signal Process.*, vol. 2, no. 3, pp. 261–274, Jun. 2008.
[27] F. Fleuret, "Fast binary feature selection with conditional mutual information," *J. Mach. Learn. Res.*, vol. 5, pp. 1531–1555, Nov. 2004.
[28] H. H. Yang and M. Moody, "Data visualization and feature selection: New algorithms for nongaussian data," in *Proc. Adv. Neural Inf. Process. Syst.*, 2000, pp. 687–693.
[29] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, 2005.
[30] Z. Zhao, L. Wang, H. Liu, and J. Ye, "On similarity preserving feature selection," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 3, pp. 619–632, Mar. 2013.

[31] J. Liu, Y. Lin, M. Lin, S. Wu, and J. Zhang, "Feature selection based on quality of information," *Neurocomputing*, vol. 225, pp. 11–22, Feb. 2017.
[32] C. Wang, Q. Hu, X. Wang, D. Chen, Y. Qian, and Z. Dong, "Feature selection based on neighborhood discrimination index," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 7, pp. 2986–2999, Jul. 2018.
[33] P. Bermejo, J. A. Gámez, and J. M. Puerta, "Speeding up incremental wrapper feature subset selection with Naive Bayes classifier," *Knowl.-Based Syst.*, vol. 55, pp. 140–147, Jan. 2014.
[34] G. Chen and J. Chen, "A novel wrapper method for feature selection and its applications," *Neurocomputing*, vol. 159, pp. 219–226, Jul. 2015.
[35] S. Maldonado and J. López, "Dealing with high-dimensional class-imbalanced datasets: Embedded feature selection for SVM classification," *Appl. Soft Comput.*, vol. 67, pp. 94–105, Jun. 2018.
[36] J. Tao, D. Zhou, and B. Zhu, "Multi-source adaptation embedding with feature selection by exploiting correlation information," *Knowl.-Based Syst.*, vol. 143, pp. 208–224, Mar. 2018.
[37] I.-F. Chung, Y.-C. Chen, and N. R. Pal, "Feature selection with controlled redundancy in a fuzzy rule based framework," *IEEE Trans. Fuzzy Syst.*, vol. 26, no. 2, pp. 734–748, Apr. 2018.
[38] S. K. Nayak, P. K. Rout, A. K. Jagadev, and T. Swarnkar, "Elitism based Multi-Objective Differential Evolution for feature selection: A filter approach with an efficient redundancy measure," *J. King Saud Univ. Comput. Inf. Sci.*, to be published.
[39] H. Lyu, M. Wan, J. Han, R. Liu, and C. Wang, "A filter feature selection method based on the maximal information coefficient and Gram-Schmidt orthogonalization for biomedical data mining," *Comput. Biol. Med.*, vol. 89, pp. 264–274, Oct. 2017.
[40] M. Labani, P. Moradi, F. Ahmadizar, and M. Jalili, "A novel multivariate filter method for feature selection in text classification problems," *Eng. Appl. Artif. Intell.*, vol. 70, pp. 25–37, Apr. 2018.
[41] E. Hancer, B. Xue, and M. Zhang, "Differential evolution for filter feature selection based on information theory and feature ranking," *Knowl.-Based Syst.*, vol. 140, pp. 103–119, Jan. 2018.
[42] D. Wang, F. Nie, and H. Huang, "Feature selection via global redundancy minimization," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 10, pp. 2743–2755, Oct. 2015.
[43] M. Wan, G. Yang, S. Gai, and Z. Yang, "Two-dimensional discriminant locality preserving projections (2DDLPP) and its application to feature extraction via fuzzy set," *Multimedia Tools Appl.*, vol. 76, no. 1, pp. 355–371, Jan. 2017.
[44] M. Wan, M. Li, G. Yang, S. Gai, and Z. Jin, "Feature extraction using two-dimensional maximum embedding difference," *Inf. Sci.*, vol. 274, pp. 55–69, Aug. 2014.
[45] M. Wan, Z. Lai, G. Yang, Z. Yang, F. Zhang, and H. Zheng, "Local graph embedding based on maximum margin criterion via fuzzy set," *Fuzzy Sets Syst.*, vol. 318, pp. 120–131, Jul. 2017.
[46] C. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 623–666, Jul. 1948.
[47] C. Tsallis, "Possible generalization of Boltzmann-Gibbs statistics," *J. Statist. Phys.*, vol. 52, nos. 1–2, pp. 479–487, 1988.
[48] C. Gini, *On the Measure of Concentration With Special Reference to Income and Statistics* (General Series No. 208). Colorado Springs, CO, USA: Colorado College Publication, 1936, pp. 73–79.
[49] K. Gajowniczek, A. Orłowski, and T. Ząbkowski, "Simulation study on the application of the generalized entropy concept in artificial neural networks," *Entropy*, vol. 20, no. 4, p. 249, 2018.
[50] S. Furuichi, "Information theoretical properties of Tsallis entropies," *J. Math. Phys.*, vol. 47, no. 2, pp. 479–561, 2006.
[51] Z. Ye, J. Yang, M. Wang, X. Zong, L. Yan, and W. Liu, "2D Tsallis entropy for image segmentation based on modified chaotic bat algorithm," *Entropy*, vol. 20, no. 4, pp. 239–267, 2018.
[52] M. Vila, A. Bardera, M. Feixas, and M. Sbert, "Tsallis mutual information for document classification," *Entropy*, vol. 13, no. 9, pp. 1694–1707, 2011.
[53] D. Koller and M. Sahami, "Toward optimal feature selection," in *Proc. 13th Int. Conf. Mach. Learn.*, 1996, pp. 284–292.
[54] R. Arias-Michel, M. García-Torres, C. Schaerer, and F. Divina, "Feature selection using approximate multivariate Markov blankets," in *Hybrid Artificial Intelligent Systems* (Lecture Notes in Computer Science), vol. 9648, Cham, Switzerland: Springer, 2016, pp. 114–125.
[55] Y. Xu, X. D. Wang, Y. X. Zhang, and W. Quan, "Feature selection algorithm-based approximate Markov blanket and dynamic mutual information," (in Chinese), *Comput. Sci.*, vol. 39, no. 8, pp. 220–223, 2012.
[56] M. Han and X. Liu, "Forward feature selection based on approximate Markov blanket," in *Advances in Neural Networks* (Lecture Notes in Computer Science), vol. 7368. Berlin, Germany: Springer, 2012, pp. 64–72.
[57] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," *J. Mach. Learn. Res.*, vol. 5, no. 10, pp. 1205–1224, 2004.
[58] R. Arias-Michel, M. García-Torres, C. E. Schaerer, and F. Divina, "Feature selection via approximated Markov blankets using the CFS method," in *Proc. Int. Workshop Data Mining Ind. Appl. (DMIA)*, Sep. 2015, pp. 38–43.
[59] A. P. Dempster, "Upper and lower probabilities induced by a multivalued mapping," *Ann. Math. Statist.*, vol. 38, no. 2, pp. 325–339, 1967.
[60] G. Shafer, *A Mathematical Theory of Evidence*. Princeton, NJ, USA: Princeton Univ. Press, 1976.
[61] J. Xiao, M. Tong, C. Zhu, and Q. Fan, "Improved combination rule of evidence based on pignistic probability distance," *J. Shanghai Jiaotong Univ.*, vol. 46, no. 4, pp. 636–641 and 645, 2012.
[62] Y. Deng, W. Shi, and Z. Zhu, "Efficient combination approach of conflict evidence," (in Chinese), *J. Infr. Millim. Waves*, vol. 23, no. 1, pp. 27–32, 2004.
[63] E. B. Paul, "Dictionary of algorithms and data structures," Nat. Inst. Standards Technol., Gaithersburg, MD, USA, Tech. Rep., Oct. 1998.
[64] M. Friedman, "The use of ranks to avoid the assumption of normality implicit in the analysis of variance," *J. Amer. Statist. Assoc.*, vol. 32, no. 200, pp. 675–701, 1937.
[65] M. Friedman, "A comparison of alternative tests of significance for the problem of m rankings," *Ann. Math. Statist.*, vol. 11, no. 1, pp. 86–92, 1940.
[66] K. Doksum, "Robust procedures for some linear models with one observation per cell," *Ann. Math. Statist.*, vol. 38, no. 3, pp. 878–883, 1967.
[67] S. García, A. Fernández, J. Luengo, and F. Herrera, "Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power," *Inf. Sci.*, vol. 180, no. 10, pp. 2044–2064, 2010.



YIFENG ZHENG received the B.E. degree in computer science and technology from Minnan Normal University, Zhangzhou, China, in 2004, and the M.E. degree in computer technology from the China University of Petroleum, Beijing, China, in 2016, where he is currently pursuing the Ph.D. degree in computer technology. In 2004, he joined the Faculty of the School of Computer Science, Minnan Normal University. His research interests include artificial intelligence, machine learning, and network communications.

GUOHE LI received the Ph.D. degree. He is currently a Professor and a Ph.D. Supervisor with the College of Information Science and Engineering, China University of Petroleum, Beijing. His research interests include artificial intelligence, machine learning, and knowledge discovery.

WENJIE ZHANG received the B.E. degree in applied mathematics from the University of Electronic Science and Technology of China, Chengdu, China, in 2008, and the Ph.D. degree in computer engineering from Nanyang Technological University, Singapore, in 2014. From 2013 to 2014, he was a Postdoctoral Research Fellow with the Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong. In 2014, he joined the Faculty of the School of Computer Science, Minnan Normal University, Zhangzhou, China. His research interests include cognitive radio networks, TV white spaces, and wireless communications.

YING LI received the B.E. and M.E. degrees from the College of Geophysics, Northeast Petroleum University, Heilongjiang, China, in 2008 and 2012, respectively. She is currently pursuing the Ph.D. degree in computer technology with the China University of Petroleum, Beijing. Her research interests include artificial intelligence and machine learning.

BAOYA WEI received the B.E. degree in computer science and technology from Fuzhou University, Fuzhou, China, in 2003, and the M.E. degree in control science and engineering from Xiamen University, Xiamen, China, in 2008. In 2003, she joined the Faculty of the School of Computer Science, Minnan Normal University, Zhangzhou, China. Her research interests include network communications, artificial intelligence, and machine learning.