SVM-Based Feature Selection for Differential Space Fusion and Its Application to Diabetic Fundus Image Classification

WEIPING DING1,2 (Senior Member, IEEE), WEI JING1, SHUAIRONG DING1, and KHAN MUHAMMAD3 (Member, IEEE), LONGJIE REN1

1 School of Information Science and Technology, Nantong University, Nantong 226019, China
2 Centre for Artificial Intelligence, Faculty of Engineering and Information Technology, University of Technology Sydney, Ultimo, NSW 2007, Australia
3 Intelligent Media Laboratory, Digital Contents Research Institute, Sejong University, Seoul 143-747, Republic of Korea

Corresponding author: Weiping Ding (e-mail: dwp9988@163.com).

This work was supported in part by the National Natural Science Foundation of China under Grant 61300167 and Grant 61976120, Natural Science Foundation of Jiangsu Province under Grant BK20151274 and Grant BK20191445, Six Talent Peaks Project of Jiangsu Province under Grant XYDXXJS-048, Jiangsu Provincial Government Scholarship Program under Grant JS-2016-065, and , sponsored by the Qing Lan Project of Jiangsu Province, China

ABSTRACT Data mining is one of the most important applications of machine learning. In machine learning algorithms, the fusion kernel principle component analysis (KPCA) and support vector machine (SVM) algorithm is used in complex data classification. To solve the problem that the fusion KPCA and SVM algorithm does not have promising classification performance, the SVM based on a feature selection algorithm for differential space fusion (DSF-FS) is proposed. First, the original data is processed to obtain differential space data by principle component analysis (PCA), and the KPCA algorithm is performed respectively on the original data and differential space data to get the differential space fusion features. Second, the ReliefF algorithm is used to get the weight of features, and the optimal feature combination is selected by a preliminary classification evaluation metric. Third, the SVM algorithm is used to classify the dimensionality reduction data. Finally, some experimental results on the five UCI datasets show that the proposed DSF-FS algorithm can not only improve the classification accuracy, but it can also reduce the computational complexity of the classification process. Moreover, the DSF-FS algorithm can be successfully applied in diabetic fundus image classification, and the encouraging results further demonstrate its strong feasibility and applicability.

INDEX TERMS Kernel principle component analysis, SVM, differential space data, ReliefF algorithm, Diabetic fundus image classification

I. INTRODUCTION
Machine learning enables machines to improve performance automatically as experience data accumulates. It is widely used in data mining to extract some meaningful features and to discover the implicit rules in the big data [1][2][3]. Data classification is a common problem in data mining, and for small samples, the commonly used classification algorithm is support vector machine (SVM). It is less affected by the dimension, and only related to the support vectors (SVs) of the classification margin. The SVM algorithm has been widely used in text and image classification [4]. Since the data often has redundant and useless features which increase the computation complexity of classification and decrease the classification accuracy, the dimensionality reduction algorithm is often used to preprocess the complex data. Kernel principle component analysis (KPCA), an improvement of principle component analysis (PCA), is a method of dimension reduction [5]. The algorithm can extract the principle components (PCs) with larger variance from the nonlinear data and extract the hidden classification information in the dataset. Until now, it has been used in anomaly detection [6], image denoising [7], as well as other applications. These two algorithms of
SVM and KPCA can be effectively integrated to handle the complex data classification problem.

Some work has been done to use the original and improved fusion KPCA and SVM algorithms in image analysis [8][9], fault analysis [10], and network intrusion detection [11], but the algorithms have some shortcomings of information loss in KPCA, which cause the lower classification accuracy and higher computation complexity when classifying complex data. SVM based on feature selection of differential space fusion (DSF-FS) is proposed to get better classification performance, where differential space data is used to compensate for information loss in data classification. The ReliefF algorithm, as a feature selection algorithm with a simple structure and low computation complexity [12], has been used to filter the useless features in fusion data.

In this paper, the SVM based on the feature selection algorithm for DSF-FS is proposed. First, the original data is processed to obtain differential space data by PCA, and the KPCA algorithm is performed respectively on the original data and differential space data to get the differential space fusion features. Second, the ReliefF algorithm is used to get the weight of features, and the optimal feature combination is selected by a preliminary classification evaluation metric. Third, the SVM algorithm is used to classify the dimensionality reduction data. The experiment on five UCI datasets shows that the proposed DSF-FS algorithm can effectively improve the accuracy of the classification algorithm and decrease the computation complexity. Meanwhile, the application to diabetic image classification also shows the feasibility and applicability of DSF-FS.

The rest of the paper is organized as follows: section II describes some related work. Section III describes the SVM and KPCA algorithms. A new DSF-FS algorithm is presented in Section IV, where the differential space data fusion algorithm and improved ReliefF feature selection algorithm are described in detail. The experimental results and the application to diabetic fundus image classification are provided in Section V. Finally, some conclusions and future directions are given in Section VI.

II. RELATED WORK

Much practical work has been done to get good classification performance in the fusion KPCA and SVM algorithm. For example, Alam et al. [13] applied the fusion algorithm to diagnose Alzheimer’s disease. The KPCA dimensionality reduction data was projected onto a more efficient feature space by linear discriminant analysis (LDA) which is beneficial for data classification, and the multi-kernel learning SVM was used to classify data. This method achieved good classification performance, but it lost the classification information in the data which caused computation complexity to increase. In regards to complex distribution of the process variables, Xu et al. [14] used KPCA to extract the hidden information from the data, and the sparse SVM was used to identify the fault. The proposed method was applied to the simulation of the Tennessee Eastman (TE) chemical process, and this method identified various types of faults used in this specific domain. Kuang et al. [15] used the fusion KPCA and SVM algorithms for network intrusion detection, in which an insensitive cost function was added to the SVM algorithm and the data features in the Gaussian kernel function of the SVM algorithm were normalized. The resulting algorithm achieved good classification accuracy in the network intrusion data, but the algorithm process was very complex.

The improvement of the fusion KPCA and SVM algorithm can be a combination of feature processing methods. For example, Wu et al. [16] used self-organizing map (SOM) and KPCA algorithms to extract the features in the patent data, and the SVM algorithm was used to derive the classification results. The SOM algorithm was used to cluster the patient data, which was beneficial to the process of data classification. Cui et al. [17] used the graph-based substructure pattern mining (gSpan), graph kernel principal component analysis (graph kernel PCA), and subnetwork selection to extract features, and then the SVM algorithm was used to classify the data. The method was used to produce the brain function network, when diagnosing Alzheimer’s disease and its early stage mild cognitive impairment (MCI).

The improvement of the fusion KPCA and SVM algorithm can also be an improved SVM algorithm. For example, because the generalization error bound depends on radius and margin in the SVM algorithm, Wu et al. [18] proposed a convex radius-margin-based SVM model for joint learning of feature transformation and the SVM classifier, and the KPCA algorithm was used to extend the algorithm to joint learning of nonlinear transformation and the SVM classifier. This method effectively improved the classification effect of the fusion algorithm. Alaíz et al. [19] reinterpreted the KPCA algorithm as the solution to the convex optimization problem which optimized semi-supervised classification. The algorithm designed can be seen as a least squares SVM problem with a regularization parameter multiplied by a negative sign, combined with a variational principle for KPCA.

The main work in this paper is to get good features for classification in fusion KPCA and SVM algorithms. In order to get good features, differential space data was used to fuse the data and the improved ReliefF algorithm was used to filter the useless data in the fusion data. This method was used for complex data which was hard to classify.

III. BASIS ALGORITHM

A. SVM ALGORITHM

The SVM algorithm is a machine learning algorithm developed from statistical learning theory. It is a supervised classification method based on structural risk minimization.
The SVM algorithm can process nonlinear, high-dimensional, unbalanced small sample data [20] and results in good generalization.

The SVM algorithm primarily handles the binary classification problem, where the sample data can be expressed as

\[(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n) \in \mathbb{R}^d \times Y, Y = \{-1, 1\}\]

where \(x_i\) is the sample data for classification, \(y_i\) is the category label of \(x_i\). For the linear separable data, SVM algorithm satisfies \(y_i(w^T x_i + b) \geq 1, 1 \leq i \leq n\), in which \(w^T x + b = 0\) represents a hyperplane, where parameters \(w\) and \(b\) respectively represent the coefficient and bias. The maximum classification interval algorithm can be defined as Eq. (2).

\[
\begin{align*}
\min & \quad \frac{1}{2} ||w||^2 , \quad \text{subject to} y_i (w^T x_i + b) \geq 1, 1 \leq i \leq n
\end{align*}
\]

In order to achieve good classification in nonlinear inseparable data, the soft-margin optimal problem is defined as Eq. (3) by introducing slack variables.

\[
\begin{align*}
\min & \quad \frac{1}{2} ||w||^2 + C \sum_{i=1}^{n} \xi_i , \\
\text{s.t.} & \quad y_i (w^T \phi(x_i) + b) \geq 1 - \xi_i , 1 \leq i \leq n
\end{align*}
\]

where \(\xi_i\) is the \(i\) slack variable, \(C\) is a penalty parameter. \(\phi\) is a high-dimensional feature projection function related to the kernel function \(k(x_i, x_j)\).

B. KPCA ALGORITHM

The aim of the PCA algorithm is to define the most meaningful base, which can filter noise and expose the implicit structure of data. However, the PCA algorithm is a linear data process method, which can’t fit the nonlinear data [21]. FIGURE 2 is a schematic diagram of two-dimensional Gaussian linearly distribution data.

![FIGURE 2](image)

**FIGURE 1.** 2D Gaussian linearly distribution data schematic diagram.

Compared to the PCA algorithm, the KPCA algorithm can handle the nonlinear data. The key method of KPCA is to transform the input data into a high-dimensional feature space and perform PCA. The Kernel method is widely used in the nonlinear data process, where the main step is to project the training set \(\{(x_i, y_i)\}_{i=1}^{n}\) onto a high-dimensional feature space to get \(\{\phi(x_i), y_i\}_{i=1}^{n}\), in which \(\phi(\cdot) : \mathbb{R}^d \rightarrow H\) is a high-dimensional nonlinear map, \(\mathbb{R}^d\) is Euclidean space, \(H\) is Hilbert space where the point multiplication of high-dimensional sample vectors is satisfied with the equation \(\phi(x_i)^T \phi(x_j) = k(x_i, x_j)\), in which \(k(x_i, x_j)\) is the kernel function.

The specific procedure of KPCA algorithm are as follows.

**Algorithm 1:** KPCA algorithm

1. Original data \(X = \{x_1, x_2, \ldots, x_n\}, x_i \in \mathbb{R}^d\).
2. \(K = k(x_i, x_j), i, j = 1, 2, \ldots, n\).
3. The centered kernel matrix is defined as \(K_c = K - I_n K - KL + I_n L\) (4)
4. Calculate the eigenvalue \(\lambda_i\) and eigenvector \(a_j, i = 1, 2, \ldots, n\) of \(K_c\).
5. Sort the eigenvalue \(\lambda_i\) from large to small, the corresponding eigenvector \(a_j\) is arranged in the order of eigenvalues.
6. Select corresponding eigenvector of the first \(m\) eigenvalues.
7. The \(p^{th}\) kernel principle feature of KPCA obtained by reducing the dimension of data \(X: \)

\[
t_p = \frac{1}{\sqrt{\lambda_p}} \sum_{i=1}^{n} a_j[k(x_i, x), k(x_i, x), \ldots, k(x_i, x)]^T, \quad (p = 1, 2, \ldots, m).
\]

C. EVALUATION METRICS OF CLASSIFICATION

The classification performance can be evaluated by computing the confusion matrix, which is defined in TABLE I.

| TABLE I | CONFUSION MATRIX
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Predicted positive</td>
</tr>
<tr>
<td>Actual Positive</td>
<td>TP</td>
</tr>
<tr>
<td>Actual Negative</td>
<td>FP</td>
</tr>
</tbody>
</table>

The column represents the predicted class, the row represents the actual class. In the confusion matrix, \(TP\) is the number of positive samples correctly classified, \(FN\) is the number of positive samples incorrectly classified into a negative class. \(FP\) is the number of negative samples incorrectly classified into a positive class, \(TN\) is the number of negative samples that are correctly classified [22].

The classification evaluation metrics can be obtained by computing the confusion matrix. The simplest evaluation metric is Accuracy of prediction, which represents the ratio of the correct classification number in all the samples as given in Eq. (6):

\[
\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}.
\]
The other related classification evaluation metrics are shown as follows:

**Precision:**

\[
\text{Precision} = \frac{TP}{TP + FP}.
\]  

**Recall:**

\[
\text{Recall} = \frac{TP}{TP + FN}.
\]  

**F-value:**

\[
F\text{-value} = \frac{(1 + \beta^2) \times \text{Recall} \times \text{Precision}}{\beta^2 \times (\text{Recall} + \text{Precision})}.
\]  

**G-mean:**

\[
G\text{-mean} = \frac{TP \times TN}{(TP + FN) \times (TN + FP)}.
\]  

**IV. SVM BASED FEATURE SELECTION FOR DIFFERENTIAL SPACE FUSION**

In this section, a novel fusion KPCA and SVM algorithm called SVM-based feature selection for differential space fusion (DSF-FS) is proposed to solve the problem of poor classification accuracy of the fusion KPCA and SVM algorithm. In the DSF-FS algorithm, the method of differential space data fusion algorithm is first introduced, which can solve the problem of the information loss. The improved FeliFeF algorithm is then used to filter out the useless features in the fusion data, and the final feature data is classified by SVM classifier. The flowchart of DSF-FS is illustrated in FIGURE 2.

**A. DATA PREPROCESSING**

Before classifying the data, the original data should be preprocessed. The two key points in data preprocessing are as follows:

1) Filling in the missing value in data. The methods of filling the missing data are as follows: first, delete the corresponding observations or variables; and second, use the average of the corresponding variables. The first method can cause information loss in original data sets; the second method does not use the classification information in data sets. In this paper, the variable average of the same category is used to fill in the missing data, which helps to classify the data.

2) According to the rule of the algorithm and distribution of the data, the question whether to normalize the data or not is raised. No normalization in the data which has different range in variables may cause the decrease of efficiency and classification accuracy. Data normalization, which is scaling the data in the range of [0,1], can improve efficiency of the classification algorithm, but the method cannot process the outliers in data. The method used in this paper is a standard method Z-score, which uses the average and standard deviation to normalize the data, fitting the common data into a Gaussian distribution. The processed data are subject to normal distribution, whose average and standard deviation are respectively 0 and 1. The function equation of standardization denotes as follow:

\[
x^* = \frac{x - \mu}{\sigma},
\]

where \(\mu\) is the average, and \(\sigma\) is the standard deviation.

**B. DIFFERENTIAL SPACE DATA FUSION ALGORITHM (DSF)**

It is assumed that PCs with big variance effectively represent data while PCs with small variance represent noise, so the PCA or KPCA algorithm gains eigenvectors of the first \(m\) biggest eigenvalues from original data and ignores eigenvectors of small eigenvalues. However, PCs with small variance also have information which is useful for data classification. As a result, the fusion data of the original space and the differential space used in face recognition is introduced.

The main ideas of the DSF algorithm are as follows: get differential space data; then perform the KPCA algorithm on the original data and differential space data; and finally, combine the dimensionality reduction data together. In these steps, differential space data is attained by subtracting dimensionality reduction data from original data. The detailed steps are shown as follows:
Algorithm 2: Differential Space Data Fusion Algorithm

Input data: \( x_i, i = 1, 2, ..., n \) is the input data in \( N \)-dimensional original data space. Suppose that the \( n \) input samples constitute original data space \( V = \text{span}\{x_1, x_2, ..., x_n\} \) and original sample matrix \( X = [x_1, x_2, ..., x_n] \cdot x_i, i = 1, 2, ..., n \)

Output data: differential space fusion data.

1. Perform PCA on the sample data in the original data space to get eigenvalues \( \lambda_i', i = 1, 2, ..., n \) and the corresponding eigenvectors \( a_i' \).
2. Sort the eigenvalues from large to small, and pick the first \( m(m < n) \) eigenvectors to constitute a feature coefficient matrix \( a' = [a_1', a_2', ..., a_m'] \) according to accumulative contribution rate threshold. Accumulative contribution rate threshold is represented by Eq. (12),
   \[
   A_k = \frac{\sum_{i=1}^{k} \lambda_i'}{\sum_{i=1}^{n} \lambda_i'} \times 100\% \leq T,
   \]
   where \( A_k \) represents the first \( k \) feature values’ accumulative contribution rate. \( T \) represents threshold.
3. Project original data onto subspace formed by the feature coefficient matrix \( a' \) in which subspace is \( V' = \text{span}\{a_1', a_2', ..., a_m'\} \), the subspace projection data \( Y \) takes the form as Eq. (13).
   \[
   Y = (a')^T X = [y_1, y_2, ..., y_m],
   \]
   where \( y_i \in \mathbb{R}^m \).
4. Multiply feature coefficient matrix with subspace projection data to get the reconstruction of dimensionality reduction data \( Y \) in the original space, the reconstructed sample vector is defined as \( X' \):
   \[
   X' = a' Y = [x'_1, x'_2, ..., x'_m],
   \]
   where \( x'_i \in \mathbb{R}^n \).
5. Suppse the differential space between original data space and subspace be \( D = V - V' \), sample vector in the differential space can be defined as \( Z = X - X' = [z_1, z_2, ..., z_n] \), differential space data is the sample vector \( Z \) of differential space.
6. Perform the KPCA algorithm on the original data \( X \) and differential space data \( Z \) to get dimensionality reduction data \( X^* \) and \( Z^* \), combine the PCs of the two dimensionality reduction data to get the fused features.

The DSF algorithm can extract more information from the data, so it decreases the effect of the information loss for classification. However, this algorithm only considers the variance information in data, which may add some noise to data.

C. IMPROVED RELIEFF FEATURE SELECTION ALGORITHM

In order to increase the correlation between data features and label information, the improved ReliefF feature selection algorithm is used for further processing of the data. The ReliefF algorithm insists that good features make the same class samples close and different class samples separated. However, the original ReliefF algorithm selects features by a fixed threshold, which makes classification accuracy uncertain for different types of data. Moreover, selecting features in accordance with classification performance provides the benefit of improving the effect of data classification. The major improvement of this algorithm is that SVM is used for preliminary classification while one feature is added into the data each time, and the optimal feature combination can be achieved by comparing classification evaluation metrics obtained from classification results.

The detailed steps are listed as follows:

Algorithm 3: Improved ReliefF Feature Selection Algorithm

Input: Fusion data from DSF algorithm, which is set to \( D = \{(x_i, y_i)\}_{i=1}^n \cdot x_i \in \mathbb{R}^N, y_i \in \{-1, 1\} \). The data attribute is normalized in the range of \([0, 1]\).

Output: SVM algorithm model

1. Define and initialize a weight vector \( \omega_0 \) as zero vector \( \omega_0(f) = 0, f = 1, 2, ..., n \).
2. Randomly take sample vector \( x_i \) in the data set \( D \), find \( k \) nearest vectors \( H_j \) in the same class of \( x_i \), at the same time, find \( k \) nearest vectors \( M_j \) in the different class of \( x_i \) in which \( j = 1, 2, ..., n, j \neq i \). Weight of the \( f \) th feature of data is computed by
   \[
   \omega_0(f) = \omega_0(f) - \frac{1}{l+k} \sum_{j=1}^{l+k} \text{diff}(f, x_i, H_j) + \frac{1}{l+k} \sum_{j=1}^{l+k} \text{diff}(f, x_i, M_j),
   \]
   where \( \text{diff}(f, x_i, x_j) = |x_i^f - x_j^f| \), \( x_i^f \) represents the \( f \) th feature of data \( x_i \), and \( l \) represents taking different sample vectors \( x_j \) for \( l \) times.
3. Sort the values in the feature weight vector from big to small, cross validation SVM algorithm is used to classify data when adding one feature to classification data by the order at a time. Recall is computed as a preliminary classification evaluation metric to get optimal feature combination.
4. Repeat the (3) step of the SVM algorithm to get \( p \) optimal feature combinations, and take new feature combinations made up of features which have appeared for \( (p / 2) \) times or more than \( (p / 2) \) times.
5. The data is classified by the nonlinear soft-margin SVM classifier, the nonlinear kernel function uses the Gaussian kernel function, and the penalty parameter \( C \) is 1.

The above algorithm is used to select features that are more effective for classification. It not only decreases the computational complexity of classification, but also improves the classification accuracy.

V. EXPERIMENTAL RESULTS
In order to evaluate the performance of the algorithm, five UCI datasets [23] were first selected for experiments. Next, the DSF-FS algorithm was also applied to diabetic fundus image classification. The UCI datasets are shown in TABLE II.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>No. of samples</th>
<th>No. of dimensions</th>
<th>No. of categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iris</td>
<td>150</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>Ionosphere</td>
<td>351</td>
<td>34</td>
<td>2</td>
</tr>
<tr>
<td>Wine</td>
<td>178</td>
<td>13</td>
<td>3</td>
</tr>
<tr>
<td>Breast Cancer</td>
<td>569</td>
<td>30</td>
<td>2</td>
</tr>
<tr>
<td>Heart Disease</td>
<td>303</td>
<td>13</td>
<td>5</td>
</tr>
</tbody>
</table>

In these datasets, the two categories were selected from the data, such as the Iris and Wine datasets. The categories were randomly selected as positive and negative class when computing the Recall and Precision. In the Heart Disease dataset, the data with class label ‘0’ is defined as positive class samples while the other data was defined as negative class samples. The average computed by the same class data was used to fill in the missing data.

A. CLASSIFICATION COMPARISONS OF DIFFERENT DIMENSIONALITY REDUCTION METHODS

The problem of data standardization was analyzed in the case of the fusion KPCA and SVM algorithm [24] (KPCA+SVM). The percentage of classification accuracy obtained by using k-fold cross-validation is shown in TABLE III below, in which the abbreviation “stdn.” represents standardization. The four standardization cases are shown in the order: “No stdn.” represents no standardization of data; “Original data stdn.” represents standardization of original data; “Dimensionality data stdn.” represents standardization of dimensionality reduction data; “Both stdn.” represents standardization of the original data and dimensionality reduction data.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Iris</td>
<td>95.05</td>
<td>93.16</td>
<td>96.73</td>
<td>94.06</td>
</tr>
<tr>
<td>Ionosphere</td>
<td>94.01</td>
<td>94.41</td>
<td>96.02</td>
<td>96.13</td>
</tr>
<tr>
<td>Wine</td>
<td>63.28</td>
<td>98.93</td>
<td>68.04</td>
<td>99.62</td>
</tr>
<tr>
<td>Breast Cancer</td>
<td>93.00</td>
<td>96.99</td>
<td>93.51</td>
<td>96.52</td>
</tr>
<tr>
<td>Heart Disease</td>
<td>61.01</td>
<td>82.94</td>
<td>60.14</td>
<td>81.16</td>
</tr>
</tbody>
</table>

The experimental results of the Wine and Heart Disease datasets showed that the classification performance of “Original data stdn.” was better than those of “No stdn.” and “Dimensionality data stdn.”. At the same time, the classification performance of “Original data stdn.” was almost the same as that of “Both stdn.”. As a result, the data should have been standardized before dimensionality reduction, which is beneficial for the data to keep the original information. After standardizing the original data, the data doesn’t need to be standardized again.

The DSF-FS algorithm was compared for classification performance with the SVM algorithm [1], PCA+SVM algorithm [25] and KPCA+SVM algorithm in the experiment. The cross validation was used in all the algorithms. For the algorithms that use PCA or KPCA to reduce dimension, the dimension of the original data was decided by an accumulative contribution rate threshold. In order to evaluate the algorithm more accurately, only the original data was standardized in all algorithms. The ReliefF algorithm uses all samples to calculate the weights, so was the number of samples, and the number of the same class samples and different class samples chosen was 10. When selecting the optimal feature combination, the SVM algorithm was repeated five times, that is, is 5. The overall algorithms were repeated five times to get the confusion matrix and the corresponding classification evaluation metrics, the results of which were averaged in the cross-validation model.

The classification evaluation metrics of the different algorithms are shown in FIGURES 3, 4, and 5.
The above classification results showed that the classification performances of the DSF-FS algorithm for the Iris, Wine, and Breast Cancer datasets were higher than those of other compared algorithms, which indicated the effectiveness of the DSF-FS algorithm. Specially, classification Accuracy in FIGURE 3 represents the classification performance of the algorithms, and only the classification effect of Ionosphere data is lower than that of other datasets. Because the Ionosphere data set had too many features, using cumulative contribution rate threshold to select the feature vectors caused information loss in the differential space data fusion algorithm. The classification evaluation metrics of Precision and Recall in FIGURE 4 represent the classification evaluation metric of a certain category. The classification Precision was greatly influenced by the distribution of the data category, while the Recall was not. The Ionosphere data set was imbalanced, so its classification Precision was low. The influence of imbalanced data could also be seen by the F-value classification evaluation metric in FIGURE 5 which indicated the harmonic mean of classification Precision and Recall for a certain category. Once either Precision or Recall was small, the F-value will be small, so the F-value of Ionosphere was smaller than that of other datasets. The G-mean represents the geometric mean of the classification Recall of the two categories.

From FIGURES 3, 4, and 5, it can be seen that the classification effect of the KPCA+SVM algorithm was not better than that of the PCA+SVM algorithm in most cases. The main reason was that the kernel function in SVM classifier has been used to transform data into high-dimensional space and reusing the kernel function in the KPCA algorithm did not lead to better classification results. However, the DSF-FS algorithm solved this problem in the pie chart, because it uses the improved ReliefF algorithm to filter the fusion features.

### TABLE IV

<table>
<thead>
<tr>
<th>Data Set</th>
<th>DSF-FS Accuracy/%</th>
<th>Time/s</th>
<th>DSF-SVM Accuracy/%</th>
<th>Time/s</th>
<th>FS-SVM Accuracy/%</th>
<th>Time/s</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iris</td>
<td>96.53</td>
<td>20.07</td>
<td>96.47</td>
<td>19.04</td>
<td>93.48</td>
<td>20.30</td>
</tr>
<tr>
<td>Ionosphere</td>
<td>93.39</td>
<td>27.68</td>
<td>93.64</td>
<td>30.70</td>
<td>94.60</td>
<td>27.89</td>
</tr>
<tr>
<td>Wine</td>
<td>99.38</td>
<td>12.29</td>
<td>99.11</td>
<td>13.09</td>
<td>82.25</td>
<td>24.44</td>
</tr>
<tr>
<td>Breast Cancer</td>
<td>80.53</td>
<td>24.44</td>
<td>79.42</td>
<td>34.32</td>
<td>97.57</td>
<td>34.32</td>
</tr>
<tr>
<td>Heart Disease</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### A. CLASSIFICATION COMPARISONS OF FEATURE EXTRACTION AND SELECTION PERFORMANCE

In this experiment, the DSF-FS algorithm is compared with the SVM algorithm based on differential space data fusion [11] (DSF-SVM) and the SVM algorithm-based improved ReliefF feature selection [26] (FS-SVM) for the classification accuracy of the algorithm (Accuracy) and the running time of the classification process (Time). The comparison results are shown in Table 4 below. The DSF-FS algorithm shows the advantage of classification Time on Ionosphere data, Wine data, and Breast Cancer data, and improved classification Accuracy on Iris, Wine, and Heart Disease data. The classification Accuracy of the algorithm on Ionosphere data and Breast Cancer data is inferior to that of the FS-SVM algorithm. The possible reason is that the dimension of Ionosphere data and Breast Cancer data are...
higher and the correlation of features is smaller, so the algorithm does not need feature extraction for abundant attribute sets. The classification Accuracy of both DSF-FS and DSF-SVM is similar, but the running time of DSF-FS is smaller than that of DSF-SVM, because the feature selection process decreases the dimension of the data. In the next phased study, the classification problems for higher dimensional data will be analyzed, and the algorithm will be improved to obtain better classification results.

The above experimental results show that the DSF-FS algorithm can effectively improve the Accuracy of the classification algorithm and reduce the running Time of the classification algorithm in some datasets.

VI. APPLICATION TO DIABETIC FUNDUS IMAGE CLASSIFICATION

In this section, we evaluate the performance of our DSF-FS algorithm in the application to diabetic fundus image classification. The classification evaluation metrics are Sensitivity and Specificity. Sensitivity represents the Recall of positive samples and the Specificity represents the Recall of negative samples. The existence of hard exudates in the lesion fundus images is chosen as the label information, the images with the symptom are labeled as positive samples and the images without the symptom are labeled as negative samples. Hard exudates indicate that there is liquid that leaks out of the fragile blood vessels, which may cause more serious consequences of macular degeneration. Hard exudates in the color fundus images are shown as yellow-white spots, which are ring-shaped, fan-shaped, star-shaped or irregularly distributed. The schematic diagram is shown in FIGURE 6, in which hard exudates are indicated by lines.

The data used in the experiment is DIARETDB1 [27], which has 89 image samples and the corresponding ground truth images. The image samples are (1152*1250*3) RGB images. The ground truth consists of binary images labeling lesion information, such as location and size. All the images contain five symptoms of diabetic retinopathy, one case is completely healthy, and the other four cases have other symptoms. The diabetic fundus images with hard exudates are regarded as one classification class, and another class has no hard exudates. The process of the diabetic fundus image classification was as follows.

First, image preprocessing was used in order to easily extract the lesion information in the image. It contained three steps: 1) extracting the image background in green channel in original RGB images by median filtering, 2) extracting the blood vessels in the fundus image by image morphology, and 3) improving the contrast of the fundus image by Gaussian filtering. The preprocessing result is shown in FIGURE 7. The images shown from left to right are the green channel image, the background removal image, the blood vessel removal image, the Gaussian filtering image.

Second, the lesion feature segmentation was processed in order to extract the classification information in the fundus images. The main process was to move out the blood vessels in fundus image. The code used in this step can be found on the fundus image website.

The fundus image data had the characteristic of small samples and high dimension, which might have resulted in the low classification accuracy. So, methods were used to increase the number of the samples and reduce the dimensions of image data. One such method was increasing the number of image samples by flipping images and increasing random noise; another method was compressing the original images to decrease the dimensions of image data.
Finally, the image data was transformed to the vectors and classified by the DSF-FS algorithm. The classification result of the DSF-FS algorithm was compared with those of the other algorithms mentioned. The experiment results are showed below in FIGURE 8.

![Image](image.png)

**FIGURE 8.** Diabetic fundus image classification results of different algorithms.

From the experiment results shown on the histogram, the DSF-FS algorithm achieved the highest Sensitivity among all algorithms, indicating that it could correctly classify the positive samples with hard exudates. The DSF-FS algorithm didn’t lose lesion information in the original images, because it extracted the differential space data features and fused the features in original space and differential space. The DSF-FS algorithm had relatively larger Specificity compared with the DSF-SVM, KPCA+SVM algorithm, because the improved ReliefF algorithm was fused to extract the useful information in data. The SVM and PCA+SVM algorithm could get larger Specificity than DSF-FS algorithm because the feature processing methods used by the DSF-FS algorithm may cause the overfitting problem in the algorithm. The Sensitivity of all algorithms was somewhat larger than the Specificity of all the algorithms, which indicated more Sensitivity and Specificity research should be done to extract the lesion features in the diabetic fundus images segmentation step.

The change of PCs has different influence for different algorithms in the classification result. In the experiment, we used the first 50 PCs to compare the KPCA+SVM and DSF-FS algorithms. The differential space data and fusion data were obtained by changing PCs in the DSF-FS algorithm, and the kernel PCs were gained by the changing PCs in the KPCA+SVM algorithm. The line chart below showed the classification results of different algorithms with a different number of PCs.

![Image](image.png)

**FIGURE 9.** Line chart of classification sensitivity with the increasing number of PCs.

**FIGURE 10.** Line chart of classification specificity with the increasing number of PCs.

In these experiment results, the DSF-FS algorithm had changing results, while the KPCA+SVM algorithm did not, because DSF-FS could get the more meaningful features from the data, and the KPCA+SVM algorithm could not extract useful fundus image features in the first 50 PCs in the processed data. Therefore, the DSF-FS algorithm was more effective than the KPCA+SVM algorithm when they were employed to classify the diabetic fundus image data in a small number of PCs.

**VII. CONCLUSION AND FUTURE DIRECTION**

This paper proposed a novel KPCA and SVM classification algorithm-based feature selection for difference space fusion (DSF-FS). The algorithm obtained the differential space data through a PCA algorithm and performed the KPCA algorithm on the differential space data and the original data, and merged the resulting features. Then, the improved ReliefF feature selection algorithm was used to filter the features, in which the preliminary classification
results were used to select the optimal feature combination. Then, the data was classified by the SVM algorithm. This algorithm could achieve better accuracy and greatly reduce the computational complexity as indicated by experimental results of UCI datasets. Moreover, the application to diabetic fundus image classification demonstrated the effectiveness of the proposed DSF-FS algorithm.

In the era of big data, the size of large data usually dynamically increases, including the current changing and interconnected datasets [28]. In the future, more research should be carried out allowing the DSF-FS algorithm to classify above-mentioned datasets. On the other hand, we will also explore more image processing methods to improve the collaborative classification performance of the DSF-FS algorithm in the diabetic fundus image classification, so as to minimize the influence of increasing complex noise in big data sets, which is important for supporting information integration necessary for diabetic fundus disease.

REFERENCES


This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2944899, IEEE Access
WEIPING DING (M’16-SM’19) received the Ph.D. degree in Computation Application, Nanjing University of Aeronautics and Astronautics (NUAA), Nanjing, China, in 2013, and was awarded the Outstanding Doctoral Dissertation by NUAA. He was a Visiting Scholar at University of Lethbridge(UL), Alberta, Canada, in 2011. From 2014 to 2015, He is a Postdoctoral Researcher at the Brain Research Center, National Chiao Tung University (NCTU), Hsinchu, Taiwan. In 2016, He was a Visiting Scholar at National University of Singapore (NUS), Singapore. From 2017 to 2018, he was a Visiting Professor at University of Technology Sydney (UTS), Ultimo, NSW, Australia. Now, Dr. Ding is the Chair of IEEE CIS Task Force on Granular Data Mining for Big Data. He is a member of Technical Committee on Soft Computing of IEEE SMCS, a member of Technical Committee on Granular Computing of IEEE SMCS, a member of Technical Committee on Data Mining and Big Data Analytics Technical Committee of IEEE CIS. His main research directions involve deep learning, data mining, evolutionary computing, granular computing, machine learning and big data analytics. He has published over 50 papers in flagship journals and conference proceedings as the first author. To date, he has held 13 approved invention patents in total over 20 issued patents. Dr. Ding currently serves on the Editorial Advisory Board of Knowledge-Based Systems, and Editorial Board of Information Fusion. He serves/served as an Associate Editor of several prestigious journals, including IEEE Transactions on Fuzzy Systems, Information Sciences, Swarm and Evolutionary Computation, IEEE Access and Journal of Intelligent & Fuzzy Systems, Co-Editor-in-Chief of Journal of Artificial Intelligence and Systems, as well as the lead guest editor in several international journals. He serves/served as a program committee member for several international conferences and workshops.

WEI JING is an employee of Internet Operations Center in Yangzhou Branch of China United Network Communications Co., Ltd. He got the Master degree from Nantong University, China in 2019. His main research interests include data mining and medical image processing.

KHAN MUHAMMAD (S’16-M’18) received the Ph.D degree in Digital Contents from Sejong University, South Korea. He is currently working as an Assistant Professor at Department of Software and lead researcher of Intelligent Media Laboratory, Sejong University, Seoul. His research interests include medical image analysis (brain MRI, diagnostic hysteroscopy and wireless capsule endoscopy), information security (steganography, encryption, watermarking and image hashing), video summarization, computer vision, fire/smoke scene analysis, and video surveillance. He has published over 60 papers in peer-reviewed international journals and conferences in these research areas with target venues as IEEE COMMAG, TII, TIE, TSMC-Systems, IoTJ, Access, TSC, Elsevier INS, Neurocomputing, PRL, FGCS, ASOC, IJIM, SWEVO, COMCOM, COMIND, JPDC, PMC, BSCP, CAEE, Springer NCA, MTAP, JOMS, and RTIP, etc. He is also serving as a professional reviewer for over 40 well-reputed journals and conferences. He is currently involved in editing of several special issues as GE/LGE. He is a member of the IEEE and ACM.

SHUAIRONG DING is a postgraduate student of School of Information Science and Technology, Nantong University, his main research interest is data mining and machine learning.

SHUAI RONG DING is a postgraduate student of School of Information Science and Technology, Nantong University, his main research interest is data mining and machine learning.

LONGJIE REN is a postgraduate student of School of Information Science and Technology, Nantong University. His main research direction is data mining and image recognition.