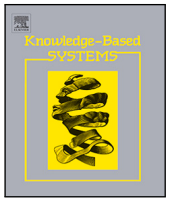
Global Subclass Discriminant Analysis

Knowledge-Based Systems 280 (2023) 111010

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys



Global subclass discriminant analysis

Huan Wan [a], Hui Wang [b,*], Bryan W. Scotney [c], Jun Liu [c], Xin Wei [d]

[a] School of Computer and Information Engineering, Jiangxi Normal University, NanChang, 330022, China
[b] School of Electronics, Electrical Engineering and Computer Science, Queen's University, Belfast, BT9 5BN, UK
[c] School of Computing, Ulster University, Belfast, BT37 0QB, UK
[d] School of Software, Nanchang University, Nanchang, 330047, China

ARTICLE INFO

Keywords:
Supervised discriminant reduction
Linear discriminant analysis
Local structure
Subclass discriminant analysis
Global subclass discriminant analysis
Face recognition
Pattern recognition
Machine learning

ABSTRACT

Linear discriminant analysis (LDA) is a powerful supervised dimensionality reduction method for analysing high-dimensional data. However, LDA cannot use locality information in data, which makes LDA degrade dramatically in performance on multimodal data. A number of LDA variants have been proposed to exploit locality information in data, including subclass-based LDAs. We discover a problem with these variants, which is that subclasses are selected on a within-class basis without considering other classes. This causes the loss of important information at class boundaries. In this paper, we present a novel variant of subclass-based LDA, *Global Subclass Discriminant Analysis* (GSDA). Unlike other subclass-based LDAs, GSDA selects subclasses from *global clusters* that may cross class boundaries, thus utilising within-class information and between-class information. More specifically, GSDA applies an effective clustering algorithm to the whole data to construct global clusters. It then utilises the *local structure refining strategy* on these global clusters to construct subclasses. Finally, GSDA learns a representative data subspace by maximising inter-subclass distance and minimising intra-subclass distance simultaneously. GSDA is extensively evaluated on a wide range of public datasets through comparison with the state-of-the-art LDA algorithms. Experimental results demonstrate its superiority in terms of accuracy and run times.

1. Introduction

Dimensionality reduction is a significant approach for analysing high-dimensional data. The main idea of dimensionality reduction is to transform the high-dimensional data into a low-dimensional subspace and preserve the discriminative information of high-dimensional data. Many dimensionality reduction algorithms have been proposed in the literature. Among these algorithms, principal component analysis (PCA) [1] and linear discriminant analysis (LDA) are the most representative and commonly used methods. PCA is an unsupervised method in which the label information is not taken into account. By contrast, LDA is a supervised dimensionality reduction method. It is well-known that LDA exceeds PCA in the majority of classification tasks. Thus, we mainly concentrate on LDA in this paper.

As a popular supervised dimensionality reduction method, LDA is widely used in many applications, for instance, computer vision [2,3], pattern recognition [4,5] and document classification [6,7]. It was originally proposed by Fisher for binary classification in [8,9]. Then, LDA was generalised by Rao [10] for multiclass classification. The goal of LDA is to find a subspace where the data of different classes are far from each other and simultaneously the data of the same class are close. To achieve this, LDA attempts to obtain the optimal transformation matrix by maximising the ratio of the between-class scatter matrix to the within-class scatter matrix, where the between-class scatter matrix and within-class scatter matrix are used to measure the inter-class distance and intra-class compactness, respectively. The optimal transformation matrix can be obtained analytically only under the assumption that all classes of data have equal covariance matrices, implying that the data are Gaussian distributed. However, real-world data are often multimodal, which is more complex than Gaussian distribution. For example, the face images from the same person are typically multimodal due to different illumination conditions or head poses; similarly, cat images of different breeds are also multimodal. Thus, LDA is unable to achieve satisfactory performance on multimodal data. To address the *multimodal problem*, i.e., LDA is ineffective in processing multimodal data, many LDA variants have been developed. The goal of these variants is to make use of the local structure in multimodal data to improve LDA. According to the way in which local structure is utilised, these LDA variants can be grouped into two categories: locality-based

* Corresponding author.
E-mail addresses: HuanWan@jxnu.edu.cn (H. Wan), h.wang@qub.ac.uk (H. Wang), bw.scotney@ulster.ac.uk (B.W. Scotney), j.liu@ulster.ac.uk (J. Liu), xinwei@ncu.edu.cn (X. Wei).

https://doi.org/10.1016/j.knosys.2023.111010
Received 20 December 2022; Received in revised form 13 June 2023; Accepted 13 September 2023
Available online 20 September 2023

discriminant analysis methods and subclass-based discriminant analysis methods.

Locality-based discriminant analysis methods typically utilise similarity matrix to extract local structure information from multimodal data. The first step is to construct the similarity matrix, which can be done in different ways. The second step is to calculate two Laplacian matrices based on the similarity matrix, namely, the inter-class Laplacian matrix and the intra-class Laplacian matrix, which are used to measure the inter-class distance and intra-class compactness separately. The final step is to obtain the optimal transformation matrix based on the two Laplacian matrices. For example, Local Fisher Discriminant Analysis (LFDA) [11] combines k-nearest neighbour and Gaussian kernel function to construct a similarity matrix. Locality Sensitive Discriminant Analysis (LSDA) [12] utilises the nearest neighbour graph to construct a similarity matrix. Unlike LFDA and LSDA, which both use the pre-defined similarity measurements, Nie et al. recently developed adaptive similarity measurements to construct similarity matrix in Locality Adaptive Discriminant Analysis (LADA) [13] and Adaptive Local Linear Discriminant Analysis (ALLDA) [14], respectively. Other locality-based discriminant analysis methods include Margin Fisher Analysis (MFA) [15], Laplacian Linear Discriminant Analysis (LapLDA) [16], Local Linear Discriminant Analysis (LLDA) [17], and so on.

By contrast, subclass-based discriminant analysis methods capture local structure by finding subclasses within each class, where the found subclasses are the local structure of multimodal data. The main idea of subclass-based discriminant analysis is to partition a class into several subclasses and seek to maximise inter-class distance and minimise intra-class distance based on subclasses. Fig. 1(a) shows an example with two classes. It is clear that Class One (data in green) and Class Two (data in red) are multimodal. Each class is comprised of two Gaussian distributions. Suppose the two Gaussian distributions of each class are captured as subclasses, such as *Subclass-One1*, *Subclass-One2*, *Subclass-Two1* and *Subclass-Two2* shown in Fig. 1(b), then subclass-based discriminant analysis naturally obtains the local structure of the multimodal data. To accurately find subclasses, Zhu and Martinez [18] proposed *subclass discriminant analysis* (SDA). SDA utilises a nearest neighbour-based clustering algorithm and a stability criterion to partition every class into the same number of subclasses. Then, SDA measures the inter-class distance using a between-subclass scatter matrix and the intra-class distance using a sample covariance matrix. Finally, SDA finds the subspace that maximises inter-class distance and minimises intra-class distance through the LDA optimisation mechanism. *Mixture subclass discriminant analysis* (MSDA) [19] partitions a class into subclasses only when this class does not have a Gaussian distribution according to the nongaussianity criterion the authors proposed, where the number of subclasses is determined according to the same stability criterion as in SDA [18]. As a result, different classes may have different numbers of subclasses. MSDA uses the same between-subclass scatter matrix as in SDA, and a new within-subclass scatter matrix to measure the intra-class distance. Unlike SDA and MSDA, *separability-oriented subclass discriminant analysis* (SSDA) [20] employs a separability criterion to partition every class into a number of non-overlapping subclasses. Based on these non-overlapping subclasses, SSDA defines a new between-subclass scatter matrix and uses the LDA optimisation mechanism to find a subspace that simultaneously maximises between-class distance and between-subclass distance, and minimises within-class distance.

We note that all subclass-based methods mentioned above restrict themselves to finding subclasses within a class, which neglects the local structure between different classes. Consequently, a boundary cluster (i.e. cluster comprising samples at a class boundary) may not form a cluster if viewed within any class, so it may not be found by existing subclass-based methods. It is well-known that boundary samples (i.e. samples at class boundaries) are important for classification. So, if we can take the boundary subclasses (i.e. subclasses comprising samples at class boundaries) into account, then a subclass-based method would push them apart (i.e. separate them) as much as possible, resulting in higher classification performance. This motivates us to search for a new method which globally clusters the whole data and uses these global clusters as the basis to find subclasses before we apply the LDA mechanism. This effort results in *Global Subclass Discriminant Analysis* (GSDA), the subject of this paper. Therefore, we distinguish two types of clusters in this paper, "local cluster" and "global cluster", according to their scopes. A local cluster is one when the clustering process is applied to one class of data, as is the case with the existing subclass-based LDAs such as SDA, MSDA, and SSDA. In contrast, a global cluster is one when the clustering process is applied to the whole data, as is the case with GSDA.

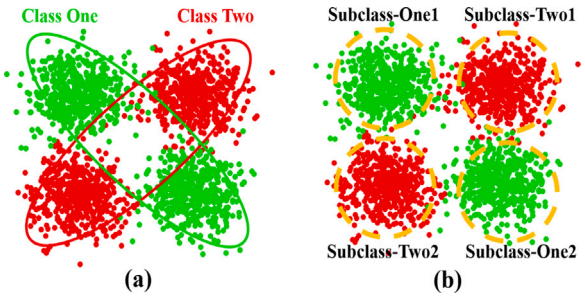


Fig. 1. An example of multimodal data. (a) A data set with two classes, in green and red, respectively, which are multimodal. (b) The data from the same class are partitioned into subclasses: subclass-One1 and subclass-One2 in Class One, subclass-Two1 and subclass-Two2 in Class Two. Each subclass is represented by an orange dashed circle.

Similar to SDA, MSDA and SSDA, GSDA also uses the idea of a subclass, but at a global scale rather than a local scale, which we call a *global separation* of the classes. GSDA seeks to capture boundary subclasses, so better classification performance is expected. Specifically, GSDA first finds *global clusters* across the whole data set rather than locally within every class. A global cluster may consist of data from different classes, suggesting the global cluster is the overlap of class-specific clusters, which must be then separated. Then a *local boundary correction strategy* is utilised to construct global subclasses from global clusters. The main idea of the local boundary correction strategy is to split every global cluster containing multiple classes of data into subsets, each of which contains only one class of data. Based on these global subclasses, a new between-class scatter matrix and a new within-class scatter matrix are defined to measure inter-class distance and intra-class distance, respectively. Finally, the LDA optimisation mechanism is used to find a subspace spanned by a set of new features, which maximises the inter-class distance and minimises the intra-class distance simultaneously. Comparing the subspace found by GSDA (GSDA subspace) with the subspaces found by SDA, MSDA and SSDA, we show in Section 4.2 that the subclasses in the GSDA subspace are not only separable between classes, but also separable within classes.

The contributions of this paper are the following:

(1) A novel subclass-based discriminant analysis, *Global Subclass Discriminant Analysis*, is developed to address the problem of LDA being ineffective in processing multimodal data. GSDA captures the local structure from both within and between classes to enhance the classification performance on multimodal data.

(2) A *local structure refining strategy* is proposed to obtain subclasses in GSDA, including boundary subclasses which are essential for separating different classes. Thus, GSDA can capture the local structure information between different classes.

(3) Extensive experiments on artificial and real-world data sets demonstrate that the proposed method outperforms the state-of-the-art locality-based and subclass-based discriminant analysis methods in terms of accuracy and run times.

The rest of this paper is organised as follows. Section 2 provides an overview of closely related work, including SDA, MSDA and SSDA. Section 3 presents details of GSDA. Sections 4 and 5 present experimental results. Section 6 concludes the paper with an outlook for future work.

2. Related work

In this section, we present an overview of subclass-based discriminant analysis methods, which are closely related to GSDA, including SDA, MSDA and SSDA. Furthermore, the context for this work and the necessary technical notations are provided.

The classic LDA and its variants are based on maximising class discriminability defined by *Fisher–Rao's criterion* [9,21], also known as the *LDA objective function*, as follows:

$$J(W) = \frac{tr(WAW^T)}{tr(WBW^T)}, \tag{1}$$

where $tr()$ denotes the trace of a matrix, A and B are matrices for inter-class difference and intra-class difference, respectively, and W is a transformation matrix that maps samples in the original space into a discriminant subspace called *LDA space*. A and B are both assumed to be symmetric and positive-definite. The optimal transformation is W^* that maximises $J(W)$, i.e.

$$W^* = \arg\max_W J(W).$$

Classically, the inter-class difference (A) and intra-class difference (B) are measured by a between-class scatter matrix S_b and a within-class scatter matrix S_w, respectively:

$$S_b = \sum_{i=1}^{C} p_i(\mu_i - \mu)(\mu_i - \mu)^T,$$

$$S_w = \frac{1}{N}\sum_{i=1}^{C}\sum_{j=1}^{N_i}(x_{ij} - \mu_i)(x_{ij} - \mu_i)^T,$$

where C is the number of classes, N is the number of samples, N_i is the number of samples in class i, $p_i = N_i/N$ is the prior probability of class i, μ_i is the mean of class i, μ is the mean of all samples and x_{ij} is the jth sample in class i. It is well known that W^* is the analytic solution to the generalised eigenvalue decomposition equation $S_w^{-1}S_bW^* = W^*\Lambda$, where Λ is a diagonal eigenvalue matrix of $S_w^{-1}S_b$. Thus, the columns of W^* are the eigenvectors of $S_w^{-1}S_b$.

2.1. Subclass discriminant analysis

Subclass Discriminant Analysis [18] is a variant of LDA that solves the multimodal problem of LDA by partitioning every class into the same number of subclasses. Thus, SDA can capture the local structure from the multimodal data based on these subclasses. The two matrices A and B in the objective function are re-defined based on subclasses, with $A = S_{bsb}^{SDA}$ being the between-subclass scatter matrix and $B = \Sigma_X$ being the sample covariance matrix:

$$S_{bsb}^{SDA} = \sum_{i=1}^{C-1}\sum_{j=1}^{H_i}\sum_{l=i+1}^{C}\sum_{n=1}^{H_l} p_{ij}p_{ln}(\mu_{ij} - \mu_{ln})(\mu_{ij} - \mu_{ln})^T, \tag{2}$$

$$\Sigma_X = \frac{1}{N}\sum_{j=1}^{N}(x_j - \mu)(x_j - \mu)^T, \tag{3}$$

where C denotes the number of classes, H_i denotes the number of subclasses in class i, μ_{ij} denotes the mean of the jth subclass in class i, $p_{ij} = \frac{N_{ij}}{N}$ ($p_{ln} = \frac{N_{ln}}{N}$) denotes the prior probability of the jth (nth) subclass of class i (l), N_{ij} is the number of samples in the jth subclass of class i, N is the number of samples, x_j is the jth sample of the data set, and μ is the overall mean over all samples. These notations will be used in the remainder of the paper.

Fisher–Rao's criterion of SDA is then the following:

$$J(W)^{SDA} = \frac{tr(WS_{bsb}^{SDA}W^T)}{tr(W\Sigma_X W^T)}. \tag{4}$$

The number of subclasses, H_i, is a key parameter in SDA. To determine H_i, the *leave-one-out-test* (LOOT) criterion [18] or the *stability criterion* [22] are used together with a nearest neighbour based clustering algorithm [18]. For details of the algorithms please see [18].

2.2. Mixture subclass discriminant analysis

Mixture Subclass Discriminant Analysis [19] is an extension of SDA, which captures the local structure of data based on the subclass allowing different numbers of subclasses within the classes. While SDA partitions every class into the same number of subclasses, MSDA partitions a specific class (or subclass) based on a nongaussianity criterion resulting in possibly different numbers of subclasses for different classes. The nongaussianity criterion is defined as the sum of the skewness and kurtosis. Every time a class (or subclass) with the largest nongaussianity is chosen, MSDA will partition it, repeat and stop partitioning when no subclass can be found with large enough nongaussianity or the specified maximum number of subclasses is reached. Finally, MSDA obtains the number of subclasses for each class based on *LOOT or stability criterion*. In MSDA the Fisher–Rao's criterion is defined as follows, adopting SDA's between-subclass scatter matrix S_{bsb}^{SDA} but introducing a new within-subclass scatter matrix $\breve{\Sigma}_X$:

$$J(W)^{MSDA} = \frac{tr(WS_{bsb}^{SDA}W^T)}{tr(W\breve{\Sigma}_X W^T)}, \tag{5}$$

where $\breve{\Sigma}_X$ is defined as

$$\breve{\Sigma}_X = S_{bsb}^{SDA} + S_{wbs}^{MSDA}, \tag{6}$$

and

$$S_{wbs}^{MSDA} = \sum_{i=1}^{C}\sum_{j=1}^{H_i} p_{ij}(x_{ij} - \mu_{ij})(x_{ij} - \mu_{ij})^T. \tag{7}$$

2.3. Separability-oriented subclass discriminant analysis

Separability-oriented Subclass Discriminant Analysis [20] is another extension of SDA. Similar to SDA and MSDA, SSDA deals with the multimodal problem by partitioning each class into several subclasses. Instead of using the *stability criterion* to determine the number of subclasses as adopted in SDA and MSDA, SSDA employs a new criterion, *separability criterion* [20], to determine the number of subclasses for each class. SSDA seeks to find non-overlapping or lightly overlapping subclasses for each class using the agglomerative hierarchical clustering with the *separability criterion* (HC-SC). In SSDA, the Fisher–Rao criterion is defined using a new between-subclass scatter matrix S_{bsb}^{SSDA} and LDA's within-class scatter matrix S_w as

$$J(W)^{SSDA} = \frac{tr(WS_{bsb}^{SSDA}W^T)}{tr(WS_wW^T)}, \tag{8}$$

where S_{bsb}^{SSDA} is defined as follows:

$$S_{bsb}^{SSDA} = \frac{N_{ij}}{N}\sum_{i=1}^{C}\sum_{j=1}^{H_i}(\mu_{ij} - \mu)(\mu_{ij} - \mu)^T. \tag{9}$$

3. Global subclass discriminant analysis

All subclass-based LDA methods seek to utilise local structure information via subclasses. Typically, SDA, MSDA and SSDA generate subclasses by clustering each class into clusters without considering other classes and taking these clusters as subclasses. In contrast, GSDA generates subclasses by clustering the whole data into clusters, and

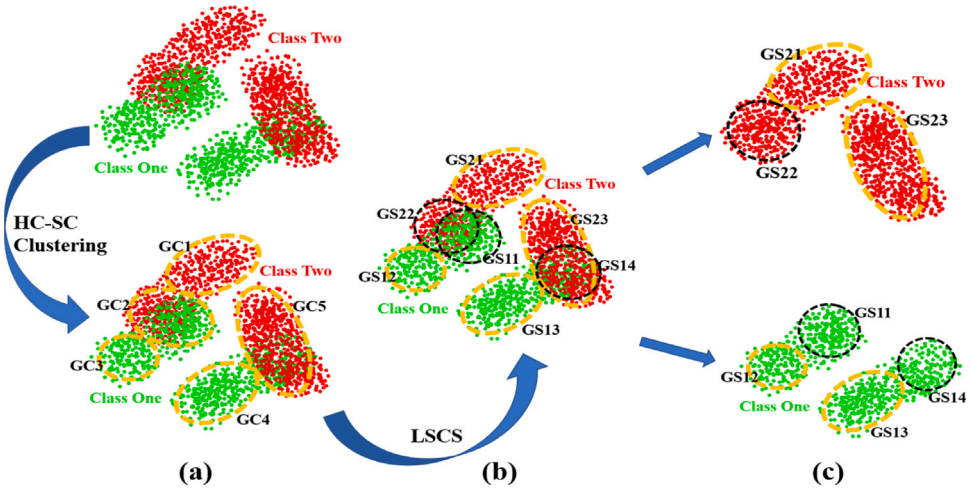


Fig. 2. An illustration of how GSDA finds G-Subclasses. (a) A data set with two classes in green and red, and five non-overlapping global clusters (i.e. GC1, GC2, GC3, GC4 and GC5) by the HC-SC clustering, denoted by dashed orange circles. (b) Seven G-Subclasses (i.e. GS11, GS12, GS13, GS14, GS21, GS22 and GS23) found using the local structure refining strategy, denoted by dashed orange and black circles, where dashed black circles represent boundary subclasses. (c) Seven G-Subclasses listed separately – 4 G-Subclasses (i.e. GS11, GS12, GS13 and GS14) for Class One, and 3 G-Subclasses (i.e. GS21, GS22 and GS23) for Class Two.

we call these clusters as *global clusters*. In GSDA, it utilises the HC-SC clustering method [20] to obtain the global clusters. Note that each global cluster may consist of data from different classes. Then a *local boundary correction strategy* is applied to further separate each global cluster into possibly multiple class-specific sub-clusters, i.e., each sub-cluster being a subset of its parent cluster and consisting of data samples from a single class. Each class-specific sub-cluster is a *GSDA subclass*.

Those subclasses with the same parent cluster are *boundary subclasses* as they belong to the same global cluster but are from different classes. So, boundary subclasses contain boundary data. It is well known that boundary data are hard to separate by their class memberships, and the performance of a classifier is to a large extent dependent on how it handles boundary data [23]. GSDA seeks to separate boundary data explicitly by maximising the distance between different boundary subclasses and then improves classification performance. We call subclasses in GSDA as *global subclasses* or G-Subclasses for short. Similarly, subclasses in SDA/MSDA/SSDA are called *local subclasses* or L-Subclasses for short. In the rest of this section, we present details of the *local boundary correction strategy* and our way of composing Fisher–Rao's criterion.

3.1. Local structure refining strategy

The local structure refining strategy (LSRS) is designed to construct global subclasses from global clusters for GSDA. LSRS splits a global cluster containing multiple classes of data into sub-clusters, or subsets, each of which contains only one class of data. More specifically, let GC_j, $j = 1, 2, \ldots, J$, be the set of global clusters; C_i, $i = 1, 2, \ldots, I$, be the set of all data samples from class i. For a global cluster GC_j, we can construct one or more global subclasses as follows:

$$GS_{ij} = \{x \in GC_j : x \in C_i\} \tag{10}$$

If all data samples in GC_j are from one class i, then we construct one and only one subclass GS_{ij} from GC_j; otherwise, we construct multiple subclasses from GC_j, which is then called *boundary cluster*.

The process of finding subclasses in GSDA is illustrated by an example in Fig. 2. There are two classes in red and green in Fig. 2. It is clear that the two classes have substantial overlap. GSDA employs the HC-SC clustering to partition the whole data set, resulting in five global clusters (GC) — see GC1–GC5 in Fig. 2(a), where GC2 and GC5 are boundary clusters. Then LSCS is applied to these five global clusters, resulting in seven G-Subclasses (GS). When we look at the classes separately, we have four GS for Class One (GS11, GS12, GS13, GS14) and three GS for Class Two (GS21, GS22, GS24). It is clear that GS11, GS14 and GS22 are boundary subclasses. If we partition Class One based on only Class One data (in green), we will get two local clusters (LC), LC1 and LC2 (Fig. 3(b), top), which are taken as two subclasses of Class One. Similarly, if we partition Class Two based on only Class Two data (in red), we will get two local clusters, LC3 and LC4 (Fig. 3(b), bottom), which are taken as two subclasses of Class Two. Thus, they completely neglect the boundary subclasses due to local subclasses being limited to within a class.

Comparing the G-Subclasses (GS) in Fig. 2(c) with L-Subclasses (LS) in Fig. 3(c), it is clear that our global approach, through the local structure refining strategy, can identify not only subclasses but also **boundary** subclasses at the intersection of different classes. For example, the dashed black circles in Fig. 2(c) are boundary subclasses. It is well known that boundary data are notoriously hard to separate correctly and are usually the culprits for incorrect classifications. Additionally, it is clear from this example that the number of G-Subclasses varies for different classes. This differs from SDA, which requires the same number of L-Subclasses for every class. This generalisation is vital as there is nothing to guarantee that every class comprises the same number of subclasses/clusters/distributions (i.e. Gaussian distributions).

Once G-Subclasses are identified, we seek to separate them in order to maximise the distance between these G-Subclasses and minimise the distance within these G-Subclasses. This is achieved through the LDA optimisation process with new scatter matrices, which are described in the next subsection.

3.2. The re-defined Fisher–Rao's criterion

A key component of the LDA optimisation is Fisher–Rao's Criterion in (1), where matrices A and B can be defined for different purposes. In GSDA, A is taken to be a scatter matrix between different G-Subclasses, S_{bGsb}, and B is taken to be a scatter matrix within G-Subclasses, S_{wGsb}. These two matrices are defined below:

$$S_{bGsb} = \sum_{k=1}^{K} (\mu_k - \mu)(\mu_k - \mu)^T, \tag{11}$$

$$S_{wGsb} = \frac{1}{N} \sum_{k=1}^{K} \sum_{l=1}^{N_k} (x_{kl} - \mu_k)(x_{kl} - \mu_k)^T, \tag{12}$$

where K is the total number of G-Subclasses in a data set, μ_k is the mean of G-Subclass k and μ is the global mean of the data set, N_k

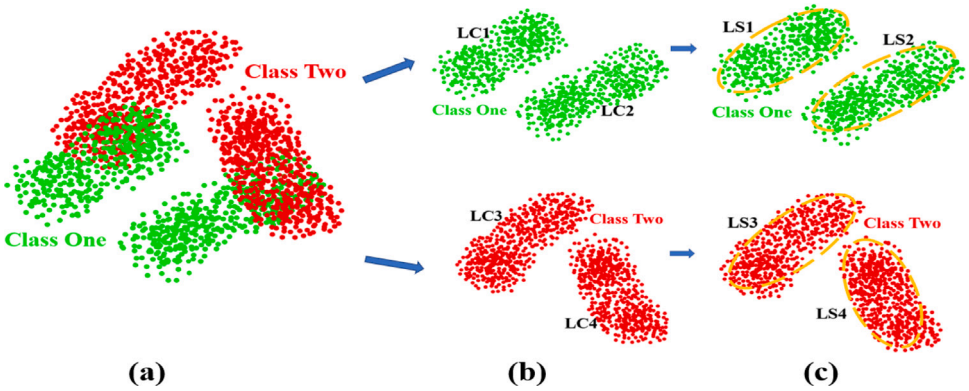


Fig. 3. An illustration of how to get L-Subclasses. (a) The same data set as in Fig. 2(a), in which two classes are in green and red, and each class consists of two clusters. (b) Two local clusters (LC) found in each class separately. (c) L-Subclasses denoted by the dashed orange circles, each corresponding to a cluster in (b).

is the number of samples in G-Subclass k and x_{kl} is the lth sample in G-Subclass k.

The Fisher–Rao's criterion, i.e. the objective function of GSDA, is re-defined as:

$$J(W)^{GSDA} = \frac{tr(W S_{bGsb} W^T)}{tr(W S_{wGsb} W^T)}. \tag{13}$$

The matrix W^* that maximises the GSDA objective is obtained by solving the generalised eigenvalue decomposition equation[1]

$$S_{wGsb}^{-1} S_{bGsb} W = \lambda W.$$

W^* is a transformation matrix, which transforms data from the original space to GSDA space, which is spanned by the eigenvectors of matrix $S_{wGsb}^{-1} S_{bGsb}$.

According to the definition of S_{bGsb} and S_{wGsb} in (11) and (12), GSDA aims to maximise between-class separation and within-class compactness at the subclass level rather than the class level. If this is achieved, separation and compactness should also be optimal at the class level. Additionally, instead of trying only to compact G-Subclasses, GSDA also tries to separate G-Subclasses. Thus, boundary data in the boundary subclasses are well separated, and high classification performance is achieved. The GSDA algorithm is summarised in Algorithm 1.

Algorithm 1 *GSDA*. In this algorithm, k_{max} is the maximum number of subclasses, which is a parameter in the HC-SC clustering method; K is the number of G-Subclasses; N_k is the number of data samples in the k^{th} G-Subclass; W^* is the final transformation matrix. The columns of W^* are the eigenvectors corresponding to the largest eigenvalues of $S_{wGsb}^{-1} S_{bGsb}$, where S_{wGsb} and S_{bGsb} are the between G-Subclass scatter matrix and within G-Subclass scatter matrix, respectively.

Input: A set of training data samples T_{set} and k_{max}.
Output: W^*.
1: Apply HC-SC clustering method on T_{set} with k_{max} to obtain the global clusters $\{GC_i\}$.
2: Apply local structure refining strategy (LSRS) on $\{GC_i\}$ to obtain K G-Subclasses.
3: **for** $k = 1$ to K **do**
4: Calculate S_{bGsb}.
5: **for** $l = 1$ to N_k **do**
6: Calculate S_{wGsb}.
7: **end for**
8: **end for**
9: Solve $S_{wGsb}^{-1} S_{bGsb} W = \lambda W$.
10: **return** W^*.

4. Evaluation using artificial data

To evaluate GSDA, we first use an artificial data set. We compare GSDA with SDA/MSDA/SSDA by measuring the separability of the transformed data in different feature spaces and visualising data distribution in the original data space as well as the different feature spaces. For clarity, we visualise data in two-dimensional space. Since the data are high-dimensional, we use the T-distributed Stochastic Neighbour Embedding (t-SNE) algorithm [24] to reduce dimensions. t-SNE is a nonlinear dimensionality reduction technique and is widely used to visualise high-dimensional data in a low-dimensional space of two or three dimensions.

The artificial data were created to contain two classes. Each class has 300 samples, in three subclasses with 100 samples per subclass. Samples in each subclass are generated by a 5-variate normal distribution. Thus every sample is a vector of 5 feature values. So, this artificial data set is a 600×5 matrix, named as *artifi-600*. The sample distribution in the original space using two t-SNE dimensions is shown in Fig. 4.

4.1. Comparison in the original space

We compare the subclasses found by different methods as they are shown in the original data space. We use SDA, MSDA, SSDA and GSDA on *artifi-600* to find subclasses, which are shown in Fig. 5. According to Fig. 4 and Fig. 5, we have the following observations.

- Qualitatively, the subclasses found by GSDA are more clearly separable than those found by other methods. We visually inspect and compare the subclasses found by different methods. The subclasses found by SSDA and GSDA overlap much less than those found by SDA and MSDA. In particular, the subclasses found by GSDA are the least overlapping.
- Quantitatively, the separability of subclasses found by GSDA is higher than that by other methods, which is consistent with the visual perception from Fig. 5. We use the Dunn Index [25] to measure the degree of separability between subclasses. Dunn Index (DI) is commonly used to evaluate clustering algorithms. A higher DI indicates better clustering in that clusters are compact

[1] Let $W S_{wGsb} W^T = \alpha$, where $\alpha > 0$ is any constant. Then, the LDA optimisation is equivalent to finding a projective matrix W to maximise $W S_{bGsb} W^T$ and under the constraint $W S_{wGsb} W^T = \alpha$. Thus, the Lagrangian multipliers method is introduced to find W^*. We define $L(W, \lambda) = W S_{bGsb} W^T - \lambda(W S_{wGsb} W^T - \alpha)$, where $\lambda \neq 0$ is Lagrange's multiplier. By setting the derivative of $L(W, \lambda)$ with respect to W to zero, we get

$$\frac{\partial L(W, \lambda)}{\partial W} = 2 S_{bGsb} W - 2\lambda S_{wGsb} W = 0$$

$$\Leftrightarrow S_{bGsb} W = \lambda S_{wGsb} W \Leftrightarrow S_{wGsb}^{-1} S_{bGsb} W = \lambda W$$

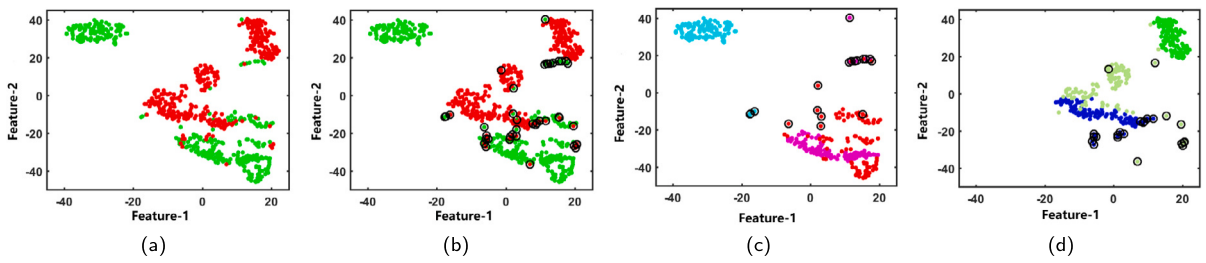


Fig. 4. Sample distribution in the original space using the first two t-SNE dimensions. (a) Two classes, green for Class One and red for Class Two. (b) Both classes with boundary points marked by black circles. (c) Class One, where dots in three different colours represent samples generated by three different 5-variate normal distributions. (d) Class Two, where dots in three different colours represent samples generated by three different 5-variate normal distributions.

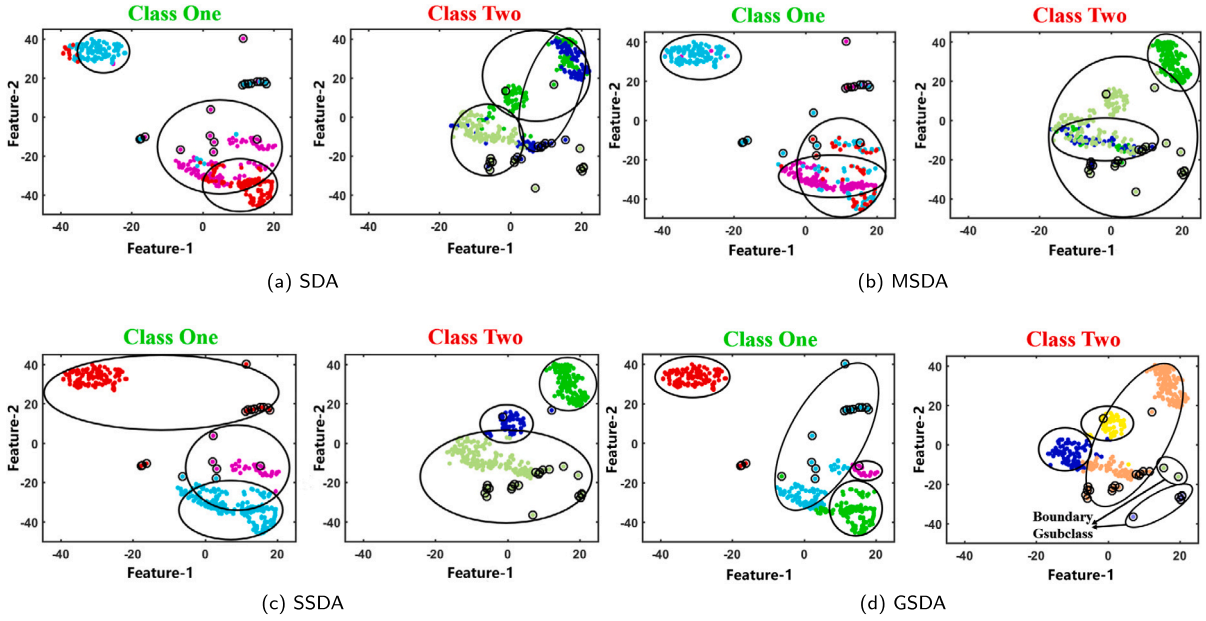


Fig. 5. Sample distribution in the original space with subclasses found by different methods. Subclasses are indicated by black circles and represented by sample dots in different colours.

and well-separated from each other. The DIs of the subclasses in the original space found by SDA, MSDA, SSDA and GSDA are 0.0046, 0.0059, 0.0188 and 0.0222, respectively.

- Again, quantitatively, the known subclasses in the original space are better separated in GSDA space than in other spaces. The DIs of the six known subclasses in SDA, MSDA, SSDA and GSDA spaces are 0.0016, 0.0141, 0.0072 and 0.0144, respectively.
- GSDA can find natural clusters of data as subclasses since it clusters the whole data set rather than one class of data at a time. One example is the cluster of data in the colour cyan at the top left corner of Fig. 4(c). This cluster is part of one subclass generated by one normal distribution, and is well separated from other clusters in the original space. This cluster has been correctly identified as a single subclass by GSDA, which is represented by red dots in Fig. 5(d), but not by any of the other methods.
- In terms of boundary subclasses, GSDA can find boundary data and place them in separate subclasses, while this is not the case with SDA, MSDA and SSDA. For example, in Fig. 5(d), two boundary subclasses are clearly marked, which contain Class Two samples that are mixed up with Class One samples, see Fig. 4(b).

4.2. Comparison in the LDA spaces

Now we compare the results of the different methods in different LDA spaces. These methods project data into respective subspaces spanned by a number of extracted features. SDA found 4 features, and MSDA, SSDA and GSDA all found 5 features. In order to visualise data in the subspace, we again use t-SNE to find two dimensions from each subspace and plot data against these two dimensions (see Fig. 6). Note that the classical LDA finds $C-1$ LDA features for a data set with C classes, so the LDA space in Fig. 6(b) has only one feature.

In Fig. 6, the data samples are plotted in different spaces, with green or red representing different classes. Comparing the original space with the classical LDA space, it is clear that the two classes are completely joined up in the classical LDA space, which confirms the problem that the classical LDA is unable to process the multimodal data. In the SDA space, the class separability does not improve much. However, the separability clearly improves substantially in MSDA, SSDA and GSDA spaces (see Fig. 6(d-f)). Furthermore, three clusters in each class are clearly observable in these three spaces, which correspond to the three multivariate normal distributions in each class. In particular, six clusters in the GSDA space are more apparent than in the other spaces. Moreover, it is also clear that the two classes are better separated in the GSDA space than in MSDA and SSDA spaces — only a few green samples are mixed into the red class in GSDA, whereas some red and green samples are still mutually mixed in both the MSDA and SSDA spaces.

4.3. Summary

The above comparative evaluations support the following conclusions.

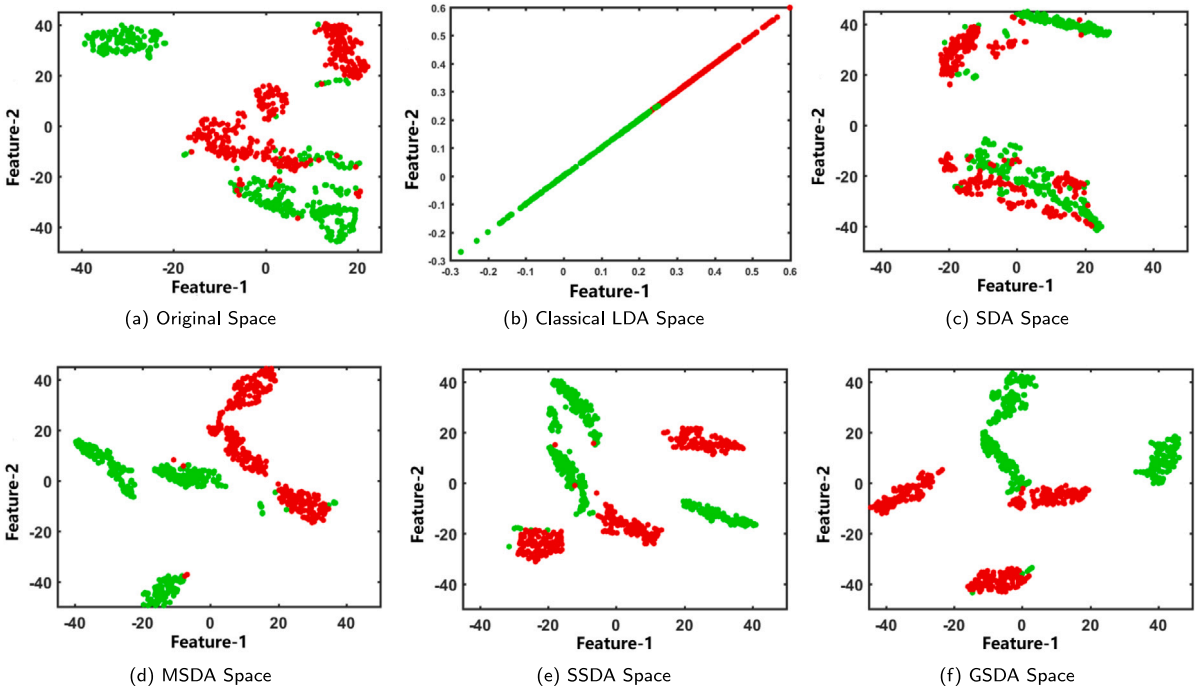


Fig. 6. Sample distribution in different spaces, where different colours represent different classes.

- GSDA can find some boundary subclasses at the class boundary, such as the two boundary subclasses shown in Fig. 5(d), which will be the anchor points to separate different classes.
- Using the LDA optimisation process, coupled with the newly defined between-subclass and within-subclass scatter matrices, GSDA can more effectively separate different classes and also separate subclasses.

5. Evaluations using real data

In this section, we use real data to evaluate the proposed GSDA through a series of experiments. We compare GSDA with LDA and its subclass-based variants that are closely related: LDA, SDA, MSDA, SSDA, Kernel SDA (KSDA) [26] and Kernel MSDA (KMSDA) [27]. Furthermore, we also compare GSDA with locality-based discriminant analysis methods: Locality Sensitive Discriminant Analysis (LSDA) [12], Adaptive Local Linear Discriminant Analysis (ALLDA) [14] and Dynamic Maximum Entropy Graph (DMEG).For KSDA and KMSDA, we employ commonly used kernels: Gaussian radial basis (RBF) kernel, Gaussian kernel, Polynomial (Poly) kernel, PolyPlus kernel and Linear kernel. In our experiments, we consider a range of classification tasks: imbalanced classification, general classification and face recognition. Five data sets are selected from the KEEL [28] repository for imbalanced classification; eleven data sets from the UCI Data Repository [29] for general classification; and YouTube faces database [30] for face recognition.

5.1. Data sets and notation

Five imbalanced data sets and eleven UCI data sets are selected in the experiments. The data sets are all numerical due to the need to compute the mean and distance. General information about these data sets is shown in Table 1 and Table 2, respectively.

We also use the YouTube faces database for the face recognition task. It contains 3425 videos of 1595 different people collected from YouTube. The average length of each video clip is 181.3 frames, and there are large variations in expression, pose and illumination in each video. In our experiments, we use the aligned images database, which contains aligned face frames broken from videos, and we use CenterSymmetric LBP (CSLBP) descriptor [31] provided by YouTube faces database website to represent each face frame.

Table 1
General information about the five imbalanced data sets used in the experiments. Here #Class denotes the number of classes, #Attribute denotes the number of attributes, #Instance is the number of instances and IR is short for imbalance ratio.

Name of dataset	#Class	#Attribute	#Instance	IR
Dermatology	6	34	366	5.55
Glass1	2	9	214	1.82
Hayes-roth	3	4	132	1.7
New-thyroid1	2	5	215	5.15
Wisconsin	2	9	683	1.86

Table 2
General information about the eleven UCI datasets used in the experiments. FTM and WDBC denote forest type mapping and Wisconsin diagnostic breast cancer, respectively. #Class denotes the number of classes, #Attribute denotes the number of attributes and #Instance is the number of instances.

Name of dataset (Acronym)	#Class	#Attribute	#Instance
Diabetic	2	19	1151
FTM	4	27	523
Glass	6	9	214
Haberman	2	3	306
Leaf	30	14	340
Letter	26	16	20000
Pageblock	2	10	5472
Penbased	10	16	1100
Pima	2	8	768
Seeds	3	7	210
WDBC	2	30	569

5.2. Experimental results

In our experiments, every DA method mentioned above is applied to find a subspace where different classes are most separated based on the method's criteria. We project a data set to this subspace, then use k-Nearest Neighbour (kNN, $k = 1$) as the classifier and ten-fold cross-validation as the evaluation framework. The evaluation metrics are

Table 3
EMA±SEM of LDA, GSDA and other subclass-based DA methods on five imbalanced data sets.

Methods \ Datasets	Derma-tology	Glass1	Hayes-roth	New-thyroid1	Wiscon-sin
LDA	0.9645 ± 0.0134	0.6119 ± 0.0226	0.7725 ± 0.0499	0.9483 ± 0.0150	0.9635 ± 0.0090
GSDA	0.9726 ± 0.0072	**0.8459 ± 0.0197**	**0.8401 ± 0.0437**	**1.0000 ± 0.0000**	**0.9663 ± 0.0038**
SSDA	0.9673 ± 0.0068	0.7716 ± 0.0318	0.8099 ± 0.0478	0.9859 ± 0.0072	0.9649 ± 0.0082
SDA	0.9561 ± 0.0085	0.7708 ± 0.0317	0.7500 ± 0.0454	0.9810 ± 0.0105	0.9648 ± 0.0059
MSDA	0.9679 ± 0.0087	0.7721 ± 0.0347	0.8099 ± 0.0492	0.9952 ± 0.0048	0.9634 ± 0.0062
KSDA(RBF)	0.9589 ± 0.0103	0.7987 ± 0.0237	0.8176 ± 0.0366	0.9907 ± 0.0062	0.9663 ± 0.0062
KMSDA(Gaussian)	0.3135 ± 0.0288	0.7236 ± 0.0283	0.8104 ± 0.0349	0.8556 ± 0.0163	0.9194 ± 0.0074
KMSDA(Linear)	0.9617 ± 0.0109	0.7519 ± 0.0037	0.7571 ± 0.0340	0.9764 ± 0.01065	0.9531 ± 0.0079
KMSDA(Poly)	0.9673 ± 0.0055	0.7610 ± 0.0234	0.7879 ± 0.0407	0.9907 ± 0.0062	0.9643 ± 0.0059
KMSDA(PolyPlus)	**0.9727 ± 0.0081**	0.7284 ± 0.0303	0.8181 ± 0.0322	0.9907 ± 0.0062	0.9561 ± 0.0072

Table 4
EMA±SEM of GSDA and locality-based DA methods on five imbalanced data sets.

Methods \ Datasets	Derma-tology	Glass1	Hayes-roth	New-thyroid1	Wiscon-sin
GSDA	**0.9726 ± 0.0072**	**0.8459 ± 0.0197**	**0.8401 ± 0.0437**	**1.0000 ± 0.0000**	**0.9663 ± 0.0038**
DMEG	0.9178 ± 0.0142	0.8219 ± 0.0329	0.7505 ± 0.0494	0.9861 ± 0.0071	0.9561 ± 0.0069
ALLDA	0.8962 ± 0.0088	0.7143 ± 0.0255	0.7725 ± 0.0512	0.9671 ± 0.0142	0.9487 ± 0.0074
LSDA	0.9480 ± 0.0076	0.7701 ± 0.0214	0.7637 ± 0.0581	0.9952 ± 0.0048	0.9576 ± 0.0067

Estimated Mean Accuracy (EMA) and *Standard Error of the Mean* (SEM). Besides, EMA and SEM are calculated in (14) and (15), respectively.

$$EMA = \frac{\sum_{i=1}^{10} p_i}{10}, \tag{14}$$

$$SEM = \frac{\delta}{\sqrt{10}}, \delta = \sqrt{\frac{\sum_{i=1}^{10}(p_i - EMA)^2}{9}}, \tag{15}$$

where p_i denotes the percentage of correct classification in the ith fold validation. For the face recognition task, we use PCA to reduce data dimensionality and keep 95% of variance before the DA method is used.

5.2.1. Classification accuracy: All methods

Imbalanced Data: The classification accuracies of the subclass-based and locality-based DA methods on the five imbalanced data sets are shown in Table 3 and Table 4, respectively. We see from Table 3 that GSDA obtains the best on 4 out of 5 imbalanced data sets and second best on the remaining data set (only 0.01% inferior to the best). In particular, GSDA outperforms LDA on *Glass1* by over 20%. Moreover, compared with the locality-based DA methods, GSDA is the best on all imbalanced data sets according to Table 4.

UCI Data: Results on the 11 UCI data sets are shown in Tables 5 and 6. It can be observed from Table 5 that GSDA achieves better classification accuracy than SDA and MSDA on the majority of data sets. Again, it is no surprise that GSDA outperforms LDA on all ten data sets. Compared with SSDA, GSDA appears to be on par with it on these UCI data sets. Additionally, comparing GSDA with kernel DA methods, it is noted that GSDA is superior to KSDA, KMSDA (Gaussian), KMSDA (Linear), KMSDA (Poly) and KMSDA (PolyPlus) on 8 out of 11, 10 out of 11, 9 out of 11, 9 out of 11 and 7 out of 11 UCI data sets, respectively. Furthermore, GSDA outperforms the locality-based DA methods on 7 out of 11 UCI data sets and achieves second best on 3 based on Table 6.

Face data: Results for LDA, subclass-based DA and locality-based DA on YouTube are presented in Table 7 and Table 8, respectively. It can be readily seen that GSDA is superior to LDA, SDA, MSDA, SSDA and KSDA on the YouTube data set. Compared with the KMSDA, GSDA is also quite competitive. From Table 8, we can observe that GSDA obtains better classification accuracy than LSDA, ALLDA and DMEG on the YouTube data set.

5.2.2. Runtime performance

Runtime results for all DA methods used in our experiments are shown in Tables 9–11. It is not surprising that GSDA is slower than LDA

Table 5
EMA±SEM of LDA, GSDA and other subclass-based DA methods on eleven UCI data sets.

Methods \ Datasets	Diabetic	FTM	Glass	Haberman	Leaf	Letter	Pageblock	Penbased	Pima	Seeds	WDBC
LDA	0.6377 ± 0.0128	0.8468 ± 0.0142	0.5747 ± 0.0269	0.6603 ± 0.0196	0.6912 ± 0.0417	0.9551 ± 0.0014	0.9117 ± 0.0026	0.9582 ± 0.0043	0.6782 ± 0.0216	0.9524 ± 0.0142	0.9579 ± 0.0079
GSDA	0.6699 ± 0.0131	**0.8871** ± **0.0106**	0.6775 ± 0.0303	0.7025 ± 0.0216	**0.7765** ± **0.0197**	**0.9717** ± **0.0011**	0.9439 ± 0.0030	0.9800 ± 0.0052	0.6939 ± 0.0216	0.9619 ± 0.0119	0.9613 ± 0.0090
SSDA	**0.6907** ± **0.0103**	0.8603 ± 0.0175	0.6647 ± 0.0372	0.7029 ± 0.0229	0.7676 ± 0.0242	0.9676 ± 0.0013	0.9673 ± 0.0022	0.9809 ± 0.0039	**0.7004** ± **0.0157**	0.9524 ± 0.0142	0.9632 ± 0.0088
SDA	0.6508 ± 0.0144	0.8813 ± 0.0091	**0.7106** ± **0.0273**	0.6605 ± 0.0144	0.5971 ± 0.0186	0.9582 ± 0.0013	0.9318 ± 0.0034	0.9745 ± 0.0038	0.6873 ± 0.0190	0.9524 ± 0.0142	0.9297 ± 0.0070
MSDA	0.6846 ± 0.0080	0.8737 ± 0.0175	0.6595 ± 0.0400	0.6898 ± 0.0295	0.7265 ± 0.0260	0.9599 ± 0.0017	**0.9678** ± **0.0017**	0.9791 ± 0.0041	0.6965 ± 0.0150	0.9571 ± 0.0085	**0.9648** ± **0.0045**
KSDA(RBF)	0.6959 ± 0.0146	0.8870 ± 0.0127	0.6491 ± 0.0114	0.6762 ± 0.0206	0.7676 ± 0.0230	0.9708 ± 0.0010	0.9631 ± 0.0017	**0.9827** ± **0.0050**	0.6925 ± 0.0179	0.9381 ± 0.0201	0.9349 ± 0.0111
KMSDA(Gaussian)	0.5153 ± 0.0196	0.3725 ± 0.0188	0.6067 ± 0.0289	**0.7286** ± **0.0187**	0.7676 ± 0.0212	0.9492 ± 0.0019	0.7666 ± 0.1013	0.0764 ± 0.0049	0.6509 ± 0.0216	0.9048 ± 0.0188	0.5292 ± 0.0447
KMSDA(Linear)	0.6907 ± 0.0135	0.8735 ± 0.0216	0.6636 ± 0.0317	0.7061 ± 0.0289	0.7559 ± 0.0224	0.9611 ± 0.0012	0.9609 ± 0.0024	0.9773 ± 0.0039	0.6900 ± 0.0206	0.9333 ± 0.0177	0.9298 ± 0.0108
KMSDA(Poly)	0.6786 ± 0.0103	0.8755 ± 0.0164	0.6677 ± 0.0246	0.6924 ± 0.0295	0.7559 ± 0.0248	0.9653 ± 0.0017	0.9543 ± 0.0023	0.9791 ± 0.0027	0.6848 ± 0.0155	0.9524 ± 0.0159	0.9420 ± 0.0117
KMSDA(PolyPlus)	0.6716 ± 0.0097	0.8698 ± 0.0165	0.6677 ± 0.0248	0.6799 ± 0.0304	0.7529 ± 0.0253	0.9592 ± 0.0018	0.9576 ± 0.0022	0.9800 ± 0.0040	0.6861 ± 0.0148	**0.9714** ± **0.0127**	0.9367 ± 0.0126

Table 6
EMA±SEM of GSDA and locality-based DA methods on eleven UCI data sets.

Methods \ Datasets	Diabetic	FTM	Glass	Haberman	Leaf	Letter	Pageblock	Penbased	Pima	Seeds	WDBC
GSDA	**0.6699** ± **0.0131**	**0.8871** ± **0.0106**	0.6775 ± 0.0303	**0.7025** ± **0.0216**	0.7765 ± 0.0197	0.9717 ± 0.0011	0.9439 ± 0.003	**0.9800** ± **0.0052**	**0.6939** ± **0.0216**	**0.9619** ± **0.0119**	**0.9613** ± **0.0090**
DMEG	0.6143 ± 0.0134	0.8449 ± 0.0130	**0.6818** ± **0.0301**	0.6535 ± 0.0329	0.6206 ± 0.0311	0.9538 ± 0.0024	0.9618 ± 0.0023	0.9700 ± 0.0049	0.6757 ± 0.0167	0.9048 ± 0.0159	0.8944 ± 0.0150
ALLDA	0.6282 ± 0.0111	0.6942 ± 0.0136	0.6736 ± 0.0249	0.6602 ± 0.0282	**0.8000** ± **0.0174**	0.9454 ± 0.0016	0.9680 ± 0.0019	0.9773 ± 0.0047	0.6900 ± 0.0146	0.8905 ± 0.0175	0.8383 ± 0.0122
LSDA	0.6272 ± 0.0182	0.8622 ± 0.0135	0.6732 ± 0.0204	0.6995 ± 0.0184	0.6882 ± 0.0278	**0.9731** ± **0.0014**	**0.9649** ± **0.0024**	0.9773 ± 0.0058	0.6653 ± 0.0163	0.9524 ± 0.0100	0.9490 ± 0.0130

Table 7
EMA±SEM of LDA, GSDA and other subclass-based DA methods on YouTube data set.

Methods \ Datasets	YouTube
LDA	0.9790± 0.0043
GSDA	0.9820 ± 0.0035
SSDA	0.9790 ± 0.0050
SDA	0.9760 ± 0.0050
MSDA	0.9790 ± 0.0043
KSDA(RBF)	0.9750 ± 0.0050
KMSDA(Gaussian)	**0.9850 ± 0.0040**
KMSDA(Linear)	0.9780 ± 0.0039
KMSDA(Poly)	0.9810 ± 0.0043
KMSDA(Polyplus)	0.9840 ± 0.0034

Table 8
EMA±SEM of GSDA and locality-based DA methods on YouTube data set.

Methods \ Datasets	YouTube
GSDA	**0.9820 ± 0.0035**
DMEG	0.9770 ± 0.0040
ALLDA	0.9410 ± 0.0288
LSDA	0.9730 ± 0.0045

on all data sets. However, GSDA is faster than SDA, MSDA and SSDA on most data sets. Moreover, GSDA is faster than KSDA and all KMSDAs on all data sets. In particular, GSDA is much faster than them on data sets that have large numbers of samples, such as *Diabetic*, *Letter*, *Pageblock*, *Penbased* and *YouTube*. This is because constructing the Gram matrix, needed in these nonlinear DA methods, is time-consuming with time complexity of $O(N^2)$, where N is the number of samples. In addition, GSDA is much faster than LSDA, ALLDA and DMEG on all data sets.

6. Conclusions

This paper has presented a new subclass-based variant of LDA, *global subclass discriminant analysis (GSDA)*, to deal with the problem

Table 9
Running time,in seconds, of the DA methods on five imbalanced data sets.

Methods \ Datasets	Dermatology	Glass1	Hayes-roth	New-thyroid1	Wisconsin
LDA	1.65	0.23	0.13	0.12	0.15
SDA	2.77	1.15	0.29	0.36	0.92
MSDA	52.47	5.26	0.96	0.39	0.61
SSDA	2.26	0.32	0.65	0.22	0.33
GSDA	2.16	0.75	0.46	0.22	0.60
DMEG	168.72	369.62	21.17	103.45	2358.83
ALLDA	525.52	328.02	53.16	369.88	1010.86
LSDA	6.44	4.41	3.53	4.79	13.47
KSDA(RBF)	19.08	7.46	15.60	4.75	594.88
KMSDA(Gaussian)	141.02	20.32	13.22	4.04	150.73
KMSDA(Linear)	131.76	17.74	31.53	7.59	123.11
KMSDA(Poly)	66.04	24.44	31.43	9.77	75.22
KMSDA(PolyPlus)	86.81	26.29	7.81	13.10	165.44

Table 10
Running time,in seconds, of the DA methods on eleven UCI data sets.

Methods \ Datasets	Diabetic	FTM	Glass	Haberman	Leaf	Letter	Pageblock	Penbased	Pima	Seeds	WDBC
LDA	0.19	0.16	0.14	0.12	0.38	2.55	0.42	0.19	0.13	0.11	0.15
SDA	2.62	0.82	0.34	0.42	1.59	211.29	27.05	1.33	2.29	0.36	1.31
MSDA	3.33	8.19	2.67	1.50	212.46	378.81	11.75	3.77	2.31	0.28	0.92
SSDA	3.02	0.63	0.45	0.43	1.73	534.01	23.96	1.07	1.01	0.27	0.85
GSDA	1.71	0.32	0.71	0.22	0.41	5099.17	29.91	3.11	0.88	0.24	1.21
DMEG	8747.86	287.63	6.14	143.20	8.92	2605909.38	386651.37	829.58	1853.98	67.58	1123.95
ALLDA	10321.52	311.97	7.66	199.85	7.30	885633.00	490167.10	1149.97	2364.29	42.15	1103.36
LSDA	65.62	21.12	12.84	11.95	16.29	12044.72	943.86	47.26	27.97	11.51	26.19
KSDA(RBF)	213.28	23.65	6.05	3.61	42.76	430820.91	8093.85	209.93	46.03	1.36	181.99
KMSDA(Gaussian)	349.31	10.52	7.05	37.73	222.65	160333.85	60107.48	410.15	249.09	86.48	149.79
KMSDA(Linear)	297.07	102.67	12.43	31.29	213.60	113362.91	23502.76	362.47	205.71	26.22	107.47
KMSDA(Poly)	355.02	51.24	12.90	15.44	219.32	256316.28	30886.48	399.88	113.16	31.40	82.21
KMSDA(PolyPlus)	228.58	106.97	6.38	19.31	215.76	236059.91	38842.89	136.90	169.28	3.51	70.41

that LDA is unable to process multimodal data effectively. The new method is designed to capture local structure information from within and between classes. We observe that most existing subclass-based LDA variants select subclasses locally, i.e. based only on data within individual classes, neglecting information between classes at class boundaries. To solve this problem, GSDA finds subclasses, *global subclasses*, by applying the HC-SC clustering algorithm to the whole data rather than one class of data at a time, then applying the local structure refining strategy to the clusters. Finally, GSDA finds a subspace that maximises the average distance between these global subclasses and concurrently minimises the average distance within every global subclass. This is achieved by re-defining Fish-Rao's criterion using new scatter matrices, one for between global subclasses and one for within global subclasses, and then applying the LDA optimisation process.

Extensive experiments using a variety of challenging data sets with multiple modalities and a mixture of subclass-based LDA methods and locality-based LDA methods have produced convincing results to conclude that GSDA is a new state-of-the-art method for discriminant analysis. GSDA has outperformed LDA consistently and also outperformed subclass-based and locality-based DA methods in most of our experiments in terms of both accuracy and runtime. In particular, GSDA has consistently outperformed both subclass-based and locality-based DA methods on imbalanced data sets. This suggests that GSDA is a competitive solution to multimodal data analysis and, in particular, a state-of-the-art solution to the challenging problem of imbalanced classification.

We have argued, using examples and supported by our experiments in Section 4, that this superior performance is due to the fact that GSDA is able to find subclasses at class boundaries so that pushing such subclasses apart can effectively separate different classes. In future work, the idea of using global subclasses to separate different classes will be extended from LDA to partial least square regression/classification. The idea will also be exploited to improve existing machine-learning algorithms and even design new ones.

Table 11
Running time in seconds, of the DA methods on YouTube data set.

Methods \ Datasets	YouTube
LDA	4.70
SDA	12.07
MSDA	31730.14
SSDA	7.01
GSDA	10.78
DMEG	54.32
ALLDA	12.10
LSDA	47.78
KSDA(RBF)	8207.46
KMSDA(Gaussian)	116396.15
KMSDA(Linear)	121923.74
KMSDA(Poly)	169384.55
KMSDA(PolyPlus)	116007.91

CRediT authorship contribution statement

Huan Wan: Conceptualization, Data curation, Methodology, Software, Writing – original draft. **Hui Wang:** Conceptualization, Methodology, Supervision, Writing – review & editing. **Bryan W. Scotney:** Supervision, Writing – review & editing. **Jun Liu:** Supervision, Writing – review & editing. **Xin Wei:** Visualization, Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Huan Wan reports financial support was provided by National Natural Science Foundation of China.

Data availability

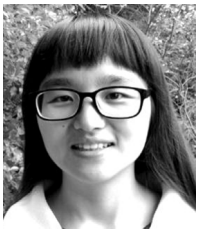
Data will be made available on request.

Acknowledgements

The work is supported by the National Natural Science Foundation of China under Grant No. 62106090 and 62106093, and the Jiangxi Urgent Need for Overseas Talents project under Grant No. 20223BCJ25026 and 20223BCJ25040.

References

[1] Svante Wold, Kim Esbensen, Paul Geladi, Principal component analysis, Chemometr. Intell. Lab. Syst. 2 (1–3) (1987) 37–52.
[2] Aijia Ouyang, Yanmin Liu, Shengyu Pei, Xuyu Peng, Minghua He, Qian Wang, A hybrid improved kernel LDA and PNN algorithm for efficient face recognition, Neurocomputing (2019).
[3] Souheila Benkhaira, Abdesslem Layeb, Face recognition using RLDA method based on mutated cuckoo search algorithm to extract optimal features, Int. J. Appl. Metaheuristic Comput. (IJAMC) 11 (2) (2020) 118–133.
[4] Mohcene Bessaoudi, Abdelmalik Ouamane, Mebarka Belahcene, Ammar Chouchane, Elhocine Boutellaa, Salah Bourennane, Multilinear side-information based discriminant analysis for face and kinship verification in the wild, Neurocomputing 329 (2019) 267–278.
[5] Na Han, Jigang Wu, Xiaozhao Fang, Jie Wen, Shanhua Zhan, Shengli Xie, Xuelong Li, Transferable linear discriminant analysis, IEEE Trans. Neural Netw. Learn. Syst. (2020).
[6] Michal Oravec, Anel Beganović, Lukáš Gál, Michal Čeppan, Christian W. Huck, Forensic classification of black inkjet prints using Fourier transform near-infrared spectroscopy and linear discriminant analysis, Forensic Sci. Int. 299 (2019) 128–134.
[7] Saritha Unnikrishnan, John Donovan, Russell Macpherson, David Tormey, An integrated histogram-based vision and machine learning classification model for industrial emulsion processing, IEEE Trans. Ind. Inform. (2020).
[8] Ronald A. Fisher, The use of multiple measurements in taxonomic problems, Ann. Eugen. 7 (2) (1936) 179–188.
[9] Ronald A. Fisher, The statistical utilization of multiple measurements, Ann. Eugen. 8 (4) (1938) 376–386.
[10] C. Radhakrishna Rao, The utilization of multiple measurements in problems of biological classification, J. R. Stat. Soc. Ser. B Stat. Methodol. 10 (2) (1948) 159–203.
[11] Masashi Sugiyama, Local fisher discriminant analysis for supervised dimensionality reduction, in: Proceedings of the 23rd International Conference on Machine Learning, 2006, pp. 905–912.
[12] Deng Cai, Xiaofei He, Kun Zhou, Jiawei Han, Hujun Bao, Locality sensitive discriminant analysis, in: IJCAI, Vol. 2007, 2007, pp. 1713–1726.
[13] Xuelong Li, Mulin Chen, Feiping Nie, Qi Wang, Locality adaptive discriminant analysis, in: IJCAI, 2017, pp. 2201–2207.
[14] Feiping Nie, Zheng Wang, Rong Wang, Zhen Wang, Xuelong Li, Adaptive local linear discriminant analysis, ACM Trans. Knowl. Discov. Data (TKDD) 14 (1) (2020) 1–19.
[15] Shuicheng Yan, Dong Xu, Benyu Zhang, Hong-Jiang Zhang, Qiang Yang, Stephen Lin, Graph embedding and extensions: A general framework for dimensionality reduction, IEEE Trans. Pattern Anal. Mach. Intell. 29 (1) (2006) 40–51.
[16] Jianhui Chen, Jieping Ye, Qi Li, Integrating global and local structures: A least squares framework for dimensionality reduction, in: 2007 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2007, pp. 1–8.
[17] Zizhu Fan, Yong Xu, David Zhang, Local linear discriminant analysis framework using sample neighbors, IEEE Trans. Neural Netw. 22 (7) (2011) 1119–1132.
[18] Manli Zhu, Aleix M. Martinez, Subclass discriminant analysis, IEEE Trans. Pattern Anal. Mach. Intell. 28 (8) (2006) 1274–1286.
[19] Nikolaos Gkalelis, Vasileios Mezaris, Ioannis Kompatsiaris, Mixture subclass discriminant analysis, IEEE Signal Process. Lett. 18 (5) (2011) 319–322.
[20] Huan Wan, Hui Wang, Gongde Guo, Xin Wei, Separability-oriented subclass discriminant analysis, IEEE Trans. Pattern Anal. Mach. Intell. 40 (2) (2018) 409–422.
[21] Calyampudi Radhakrishna Rao, Calyampudi Radhakrishna Rao, Mathematischer Statistiker, Calyampudi Radhakrishna Rao, Calyampudi Radhakrishna Rao, Linear statistical inference and its applications, Vol. 2, Wiley New York, 1973.
[22] Aleix M. Martinez, Manli Zhu, Where are linear feature extraction methods applicable? IEEE Trans. Pattern Anal. Mach. Intell. 27 (12) (2005) 1934–1944.
[23] David J. Hand, Veronica Vinciotti, Local versus global models for classification problems: Fitting models where it matters, Amer. Statist. 57 (2) (2003).
[24] Laurens van der Maaten, Geoffrey Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (Nov) (2008) 2579–2605.
[25] Joseph C. Dunn, A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters, J. Cybern. 3 (1973) 32–57.
[26] Di You, Onur C. Hamsici, Aleix M. Martinez, Kernel optimization in discriminant analysis, IEEE Trans. Pattern Anal. Mach. Intell. 33 (3) (2010) 631–638.
[27] Nikolaos Gkalelis, Vasileios Mezaris, Ioannis Kompatsiaris, Tania Stathaki, Mixture subclass discriminant analysis link to restricted Gaussian model and other generalizations, IEEE Trans Neural Netw. Learn. Syst. 24 (1) (2012) 8–21.
[28] Jesús Alcalá-Fdez, Alberto Fernández, Julián Luengo, Joaquín Derrac, Salvador García, Luciano Sánchez, Francisco Herrera, Keel data-mining software tool: data set repository, integration of algorithms and experimental analysis framework, J. Mult.-Valued Logic Soft Comput. 17 (2011).
[29] Dheeru Dua, Casey Graff, UCI Machine Learning Repository, University of California, Irvine, School of Information and Computer Sciences, 2017.
[30] Lior Wolf, Tal Hassner, Itay Maoz, Face recognition in unconstrained videos with matched background similarity, in: CVPR 2011, IEEE, 2011, pp. 529–534.
[31] Marko Heikkilä, Matti Pietikäinen, Cordelia Schmid, Description of interest regions with center-symmetric local binary patterns, in: Computer Vision, Graphics and Image Processing, Springer, 2006, pp. 58–69.

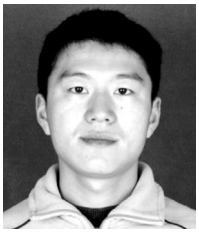


Huan Wan received the master's degree in computer application and technology from the School of Mathematics and Computer Science, Fujian Normal University, China, and the PhD degree in computing from the University of Ulster, UK. She is currently a lecturer at the School of Computer and Information Engineering, Jiangxi Normal University, China. Her current research interests are feature extraction, face verification and pattern recognition.

Hui Wang is Professor of Computer Science at Queen's University Belfast. His research interests are machine learning, knowledge representation and reasoning, combinatorial data analytics, and their applications in image, video, spectra and text data analyses. He has over 300 publications in these areas. He is the principal investigator of a number of regional, national and international projects in the areas of image/video analytics (EPSRC funded MVSE 2021–2024, Horizon 2020 funded DESIREE and ASGARD, FP7 funded SAVASA, Royal Society funded VIAD), spectral data analytics (EPSRC funded VIPIRS on virus detection 2020–2022), text analytics (INI funded DEEPFLOW, Royal Society funded BEACON), and intelligent content management (FP5 funded ICONS); and is co-investigator of several other EU funded projects. He was a Professor of Computer Science at Ulster University, on various roles in the School of Computing including Head of the AI Research Centre, Research Director. He is an associate editor of IEEE Transactions on Cybernetics, founding Chair of IEEE SMCS Northern Ireland Chapter (2009–2018), and a member of IEEE SMCS Board of Governors (2010–2013).

Bryan W. Scotney received the B.Sc. degree in mathematics from Durham University, UK in 1980 and the Ph.D. degree in mathematics from the University of Reading, UK in 1985. He is Professor of Informatics at Ulster University, UK, and was Director of Ulster University's Computer Science Research Institute since its formation in 2005 until May 2015. He has over 300 publications, spanning a range of research interests in mathematical computation, especially in digital image processing and computer vision, pattern recognition and classification, statistical databases, reasoning under uncertainty, and applications to healthcare informatics, official statistics, biomedical and vision sciences, and telecommunications network management. He has collaborated widely with academic, government and commercial partners, and much of his work has been supported by funding from the European Union Framework Programmes and the UK Research Councils. Prof. Scotney was President of the Irish Pattern Recognition and Classification Society 2007–2014, and a member of the Governing Board of the International Association for Pattern Recognition (IAPR), 2007–2014. He is currently Guest Professor at Keio University, Tokyo.

Jun Liu received the BSc and MSc degrees in applied mathematics, and the PhD degree in information engineering from Southwest Jiaotong University, Chengdu, China, in 1993, 1996, and 1999, respectively. He is currently a Reader in Computer Science at Ulster University, Northern Ireland, United Kingdom. He has been working in the field of AI for many years. His current research interests include logic and reasoning methods for intelligent systems and formal verification; intelligent DSSs and information management, with applications in health care, engineering, and industry field, etc. (e.g., safety and risk analysis; situation awareness and emergency systems; scenario/activity recognition); information fusion and data combinations; data mining and KBS; applied computational intelligence for uncertainty analysis and optimisation. He is a senior member of the IEEE. He has over 160 publications in these areas. He is an Associate Editor of IEEE Transaction on Fuzzy Systems, and an Associate Editor of Knowledge-Based Systems.



Xin Wei received a bachelor of engineering degree in computer science and technology from Shangrao Normal University, China; the master's degree in computer application and technology from the School of Mathematics and Computer Science, Fujian Normal University, China, and the PhD degree in artificial intelligence from the University of Ulster, UK. He is currently a lecturer at the School of Software, Nanchang University, China. His research interests include face recognition, image representation, deep learning and medical image processing.