



Within-class multimodal classification

Huan Wan¹  · Hui Wang¹ · Bryan Scotney² · Jun Liu¹ · Wing W. Y. Ng³

Received: 10 June 2019 / Revised: 25 May 2020 / Accepted: 24 June 2020 /

Published online: 11 August 2020

© The Author(s) 2020

Abstract

In many real-world classification problems there exist multiple subclasses (or clusters) within a class; in other words, the underlying data distribution is within-class multimodal. One example is face recognition where a face (i.e. a class) may be presented in frontal view or side view, corresponding to different modalities. This issue has been largely ignored in the literature or at least under studied. How to address the within-class multimodality issue is still an unsolved problem. In this paper, we present an extensive study of within-class multimodality classification. This study is guided by a number of research questions, and conducted through experimentation on artificial data and real data. In addition, we establish a case for within-class multimodal classification that is characterised by the concurrent maximisation of between-class separation, between-subclass separation and within-class compactness. Extensive experimental results show that within-class multimodal classification consistently leads to significant performance gains when within-class multimodality is present in data. Furthermore, it has been found that within-class multimodal classification offers a competitive solution to face recognition under different lighting and face pose conditions. It is our opinion that the case for within-class multimodal classification is established, therefore there is a milestone to be achieved in some machine learning algorithms (e.g. Gaussian mixture model) when within-class multimodal classification, or part of it, is pursued.

Keywords Within-class multimodality · Linear discriminant analysis · Subclass discriminant analysis · Separability-oriented subclass discriminant analysis

1 Introduction

Understanding the underlying data distribution before applying a machine learning process is an important step in the analysis of data, as otherwise, wrong choices may be made in the different stages of the machine learning process. Every single algorithm used in machine learning has, either explicitly or implicitly, some assumptions about the data for it to work

✉ Huan Wan
wan-h@ulster.ac.uk

effectively. For linear regression, the typical assumptions include linearity (there is linear relationship between the independent and dependent variables), exogeneity (the errors between observed and predicted values should have conditional mean zero), multicollinearity (the independent variables must all be linearly independent), homoscedasticity (the errors have the same variance in each observation) and normality (the errors have normal distribution) [7, 23]. For random forests [2], one assumption is that changes in the dependent variable are best described by hyper-rectangles in the independent variables (because they are based on trees). Another assumption is that no future value of the dependent variable will be outside of the range of values already in the training data. If the distribution of data can be described as the canonical statistical distributions it is possible to gain much inferential and predictive power [15]. The key to any successful use of data in an analysis or in making a decision is applying the correct machine learning/statistical modelling technique to the data at hand.

In this paper we consider a particular type of data distribution where there are multiple modalities (concentrations/clusters of data) within each class, *within-class multimodality*, and study how to choose the right feature extraction methods to model such data more effectively. Fig. 1(a) illustrates within-class multimodality at a conceptual level, where there are two and three modalities respectively in Class One and Class Two. Within-class multimodality is prevalent in the real world. For example, we can recognise people under different illuminations, and also in different poses. If we represent face images of the same person under different illuminations, it is likely that different images with different illuminations will be in different clusters (see Fig. 1(b) for an illustration). Actually, face recognition under varying illuminations is a challenging problem[25, 31]. The same can be said of face recognition in different head poses (see Fig. 1(c) for an illustration). Another potential application is energy disaggregation of appliances by non-intrusive load monitoring (NILM) [8, 11, 12, 19], namely disaggregating the total consumption readings into the consumption patterns of each individual appliance, where the total consumption reading of a house represents a class and the appliances in a house are the modalities within this class. Therefore, dividing a class into multiple modalities is similar to disaggregating the total consumption of all appliance into the consumption of each appliance.

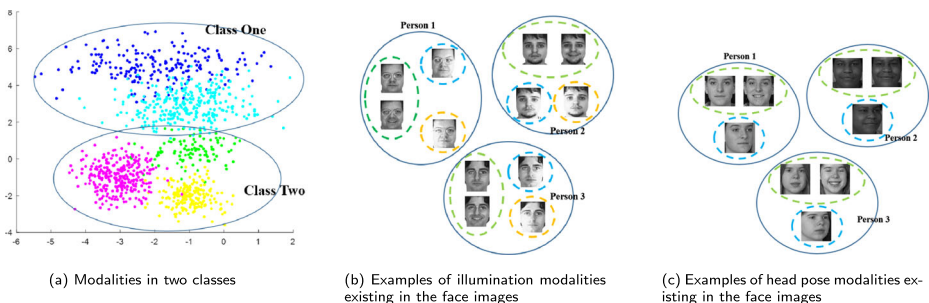


Fig. 1 (a) There are two modalities in Class One, and three modalities in Class Two, where different modalities are marked by different colours. (b) Each person has three different illumination modalities: two face images in the green dotted circle are taken under normal lighting; one face image in the cyan dotted circle is taken under normal lighting and right light on; one face image in the orange dotted circle is taken under normal lighting and left light on. (c) Each person has two different head pose modalities: two face images in the green dotted circle are taken with frontal head pose, and one face image in the cyan dotted circle is taken with rightwards head pose

Within-class multimodality has been largely ignored in the literature, or at least under studied. The closest studies are *linear discriminant analysis* (LDA) [5, 22], *subclass discriminant analysis* (SDA) [32], *Mixture subclass discriminant analysis* (MSDA) [6], and *separability-oriented subclass discriminant analysis* (SSDA) [26]. Unlike LDA which separates different classes under the assumption that each class is unimodal, SDA, MSDA and SSDA recognize that a class may be multimodal and seek to find LDA dimensions based on multimodality descriptors through the notion of subclass. SDA, MSDA and SSDA have better classification performance than LDA, which indicates the importance of within-class multimodality for classification. LDA is a classical approach to discriminant dimensionality reduction. It transforms data from the original data space into a lower dimensional space (LDA space) so that the within-class compactness is maximised whilst the between-class separation is maximised. This is achieved through maximising the well-known Fisher objective, which is composed by the within-class scatter matrix and between-class scatter matrix [5, 22]. In the presence of within-class multimodality, LDA reduces dimensionality by merging multiple modalities in each class into a single modality. SDA extends LDA in order to separate classes at a subclass level rather than at a class level. It transforms data into a lower dimensional LDA space so that the between-subclass separation is maximised, and within-class compactness is maximised. The SDA subclasses are discovered using the *leave-one-out-test* (LOOT) criterion proposed in [32] or the *stability criterion* [18]. MSDA extends SDA by replacing SDA's within-class scatter matrix with a new within-subclass scatter matrix. SSDA further extends SDA to minimise the level of overlap between subclasses within every class; thus the between-class separation is maximised, between-subclass separation is maximised and within-class compactness is maximised. The SSDA subclasses are discovered by the agglomerative hierarchical clustering algorithm using a new criterion called the separability criterion [26], which aims to divide each class into several non-overlapping clusters.

A lot is known about *within-class unimodality classification*,¹ whose aim is to build a model assuming there is one modality per class. It is well-known that simultaneously minimising intra-class variance and maximising inter-class variance will increase learning performance [4, 28, 29]. However, not enough is known about *within-class multimodality classification*, when data distribution is within-class multimodal. Existing studies (e.g. SDA and SSDA) only scratch the surface in multimodality, and many questions remain unanswered. In this paper, we present an extensive study of within-class multimodality classification as guided by the following five key questions about within-class multimodality that are important for the understanding of multimodality, the design of new learning algorithms and the improvement of existing learning algorithms.

- Question 1: Why do we consider multimodality?
- Question 2: How many clusters should we use?
- Question 3: How should we utilise the clusters?
- Question 4: Do we have real benefits?
- Question 5: If we keep increasing modalities, what will happen?

¹Unimodality is when data distribution has one centre of concentration, whereas multimodality is when data distribution has multiple centres of concentration.

The study of these questions is important for a number of reasons. Firstly, it will reveal a relationship between the modality of the data distribution and the comparative performance of the classification, so it is possible to gain an insight into the data through the comparative model performance using different data dimensionality reduction techniques. Secondly, it will establish the fact that different dimensionality reduction techniques are suitable for different data distributions. Thirdly, it will provide a direction for improving other machine learning algorithms such as neural networks by designing new loss functions.

We create artificial data sets having a range of modalities and conduct extensive experiments in order to answer Questions 1-3 (and possibly Question 5). We also select real world data sets that clearly have multiple modalities and conduct extensive experiments to answer Question 4. The contributions of this paper are highlighted as follows:

- We answered the abovementioned five key questions.
- We obtained the following useful findings: 1) when within-class multimodality is present, the concurrent maximisation of between-class separation, within-class compactness and between-subclass separation can lead to significant performance gains; 2) within-class multimodal classification offers a competitive solution to face recognition under different lighting and face pose conditions, where each lighting/pose condition corresponds to a separate modality in the data space; 3) There is correlation between multimodality and performance gain in within-class multimodality classification. Optimal performance can be expected if the number of modalities in the within-class multimodality classification algorithm is the same as the true number of within-class modalities

The rest of the paper is organised as follows. Section 2 presents relevant work including linear discriminant analysis (LDA), subclass discriminant analysis (SDA) and separability-oriented subclass discriminant analysis (SSDA). Section 3 focuses on artificial data sets and their rationale. Section 4 attempts to answer various questions about multimodality using artificial data sets, and Section 5 attempts to answer other questions using real data sets. Section 6 concludes the paper with a summary.

In the rest of the paper we use cluster, subclass and modality in different contexts but these terms are interchangeable in this paper.

2 Related work

In this section, we present an overview of related work, including the LDA, SDA and SSDA to provide the context for this work and introduce the necessary technical notations.

2.1 Linear discriminant analysis

Linear discriminant analysis (LDA) is a classical method for discriminant analysis. It has been widely used in many areas, such as pattern recognition [13, 14] and machine learning [10, 27]. LDA seeks to find a linear combination of features that separates two or more classes of objects. The resulting combination may be used as a linear classifier, or more commonly, for dimensionality reduction before later classification [30]. LDA uses a between-class scatter matrix S_b to measure the separability of classes, and uses a within-class scatter matrix S_w to measure the compactness of each class. Then LDA attempts to find a linear projective matrix W that projects data into a new space, LDA space, that is spanned by LDA features (or LDA dimensions), such that a measure of the between-class

scatter matrix S_b in the new space is maximised and simultaneously the same measure of the within-class scatter matrix S_w in the new space is minimised. S_b and S_w are defined, respectively, as follows:

$$S_b = \frac{1}{N} \sum_{i=1}^C N_i (\mu_i - \mu)(\mu_i - \mu)^T, \tag{1}$$

$$S_w = \frac{1}{N} \sum_{i=1}^C \sum_{j=1}^{N_i} (x_{ij} - \mu_i)(x_{ij} - \mu_i)^T, \tag{2}$$

where N is the number of samples, N_i is the number of samples in class i , C is the number of classes, μ_i is the mean of class i , μ is global mean of all samples, and x_{ij} denotes the j^{th} sample in class i .

LDA is an optimisation process, with the following *Fisher objective*:

$$J^{LDA}(W) = \frac{tr(W^T S_b W)}{tr(W^T S_w W)}, \tag{3}$$

where W is a projective matrix that projects data from the data space to the LDA space. In order to find an LDA space that can separate different classes well, LDA needs to find the matrix $W^* = \arg \max_W J^{LDA}(W)$. It turns out that the sought-after projective matrix W^* is

composed of the eigenvectors corresponding to the largest eigenvalues of $S_w^{-1} S_b$ [26], under the assumption that every class is Gaussian distributed and has the same covariance.

2.2 Subclass discriminant analysis

Subclass discriminant analysis (SDA) [32] is a variant of LDA that separates classes at a subclass level rather than at a class level, based on the observation that the data distribution in a class may be multimodal (i.e., forming clusters). This is achieved by dividing each class into a set of subclasses and then running an LDA-like optimisation process to maximise between-subclass separation and within-class compactness.

The between-class scatter matrix S_b of LDA is replaced by the between-subclass scatter matrix, which is defined below (4):

$$S_b^{SDA} = \sum_{i=1}^{C-1} \sum_{j=1}^{K_i} \sum_{l=i+1}^C \sum_{n=1}^{K_l} p_{ij} p_{ln} (\mu_{ij} - \mu_{ln})(\mu_{ij} - \mu_{ln})^T, \tag{4}$$

where C denotes the number of classes, K_i (K_l) denotes the number of subclasses in class i (l), μ_{ij} (μ_{ln}) denotes the mean of the j^{th} (n^{th}) subclass in class i (l), $p_{ij} = \frac{N_{ij}}{N}$ ($p_{ln} = \frac{N_{ln}}{N}$) denotes the prior of the j^{th} (n^{th}) subclass of class i (l), and N_{ij} (N_{ln}) is the number of samples in j^{th} (n^{th}) subclass of class i (l).

The within-class scatter matrix of SDA is re-defined as the sample covariance matrix as below (5):

$$S_w^{SDA} = \frac{1}{N} \sum_{j=1}^N (x_j - \mu)(x_j - \mu)^T, \tag{5}$$

where N , x_j , and μ are the number of instances, the j^{th} instance and the mean of all instances, respectively.

The Fisher objective is re-defined as follows (6):

$$J^{SDA}(W) = \frac{\text{tr}(W^T S_b^{SDA} W)}{\text{tr}(W^T S_w^{SDA} W)}. \quad (6)$$

In order to divide each class into the same number of subclasses, a *leave-one-out-test* (LOOT) criterion [32] or a faster *stability criterion* [18] is used together with a nearest neighbour based clustering algorithm [32]. Firstly, the clustering algorithm is used to sort the samples of each class so that samples with smaller Euclidean distance stay closer. To achieve this, two samples A and B are found in each class that have the largest Euclidean distance between each other, and are taken as the 1^{st} and n^{th} samples in the sorted data. After that, the samples ranked from 1^{st} to $(n/2)^{th}$ are near A , and the samples ranked from $(n/2 + 1)^{th}$ to n^{th} are near B . Then, based on the number of subclasses set by the user, the sorted samples are divided into the specified number of subclasses for each class. Finally, the LOOT criterion or stability criterion is used to find the optimal number of subclass for each class.

2.3 Separability-oriented subclass discriminant analysis

Separability-oriented subclass discriminant analysis (SSDA) [26] is an extension of SDA, which also separates classes at subclass level. It aims to (1) maximise the between-subclass separation within every class; (2) maximise the within-class compactness; and (3) maximise the overall between-class separation. This is achieved through an LDA-like optimisation process operating at subclass level and with a different Fisher objective.

The way to find optimal subclasses for each class is very different from SDA. SSDA aims to find subclasses with no or little overlap through agglomerative hierarchical clustering guided by a *separability criterion* [26]. The resulting clustering is one that maximises the average euclidean distance (*AED*) between the mean of a class and the means of subclasses in the class.

Three versions of SSDA exist [26], each having different combination of between-class scatter matrix and within-class scatter matrix. One version is reviewed here. The between-class scatter matrix in SSDA, S_b^{SSDA} , is defined in terms of the subclasses:

$$S_b^{SSDA} = \sum_{i=1}^C \frac{N_i}{N} \sum_{j=1}^{K_i} (\mu_{ij} - \mu)(\mu_{ij} - \mu)^T, \quad (7)$$

where N is the number of samples in the data set, N_i is the number of samples in class i ($i = 1, 2, \dots, C$, C is the number of class) such that $\sum_{i=1}^C N_i = N$, K_i is the number of subclasses in class i , μ is the mean of the whole data set and μ_{ij} is the mean of subclass j of class i .

The within-class scatter matrix is the standard LDA within-class matrix, $S_w^{SSDA} = S_w$. Therefore, the Fisher objective of SSDA $J^{SSDA}(W)$ is below, replacing S_b by S_b^{SSDA} . Moreover, we summarise the idea of SSDA in the Algorithm 1 and show the main steps of SSDA algorithm using a flowchart, see Fig. 2. Here, the notations used in the flowchart have same meaning as those used in the Algorithm 1.

$$J^{SSDA}(W) = \frac{\text{tr}(W^T S_b^{SSDA} W)}{\text{tr}(W^T S_w^{SSDA} W)} = \frac{\text{tr}(W^T S_b^{SSDA} W)}{\text{tr}(W^T S_w W)}. \quad (8)$$

Algorithm 1 SSDA: In this algorithm, C is the number of classes, AED_{ik} is the euclidean distance between the mean of class i and the means of subclasses in cla K_i^* is the number of subclasses found by SSDA for class i

Input: A set of training data X with class labels and the maximum number of su K .

Output: Transformation matrix W^* .

for $i = 1$ to C **do**

for $k = 1$ to K **do**

 Calculate AED_{ik} using the agglomerative hierarchical clustering guided by separability criterion.

end for

$K_i^* = \text{argmax}_k(AED_{ik})$.

 Calculate S_b^{SSDA} with K_i^* subclasses using Eq.(7).

 Calculate S_w using Eq.(2).

 The columns of transformation matrix W^* is given by the eigenvectors corresponding to the largest eigenvalues of $S_w^{-1} S_b^{SSDA}$.

end for

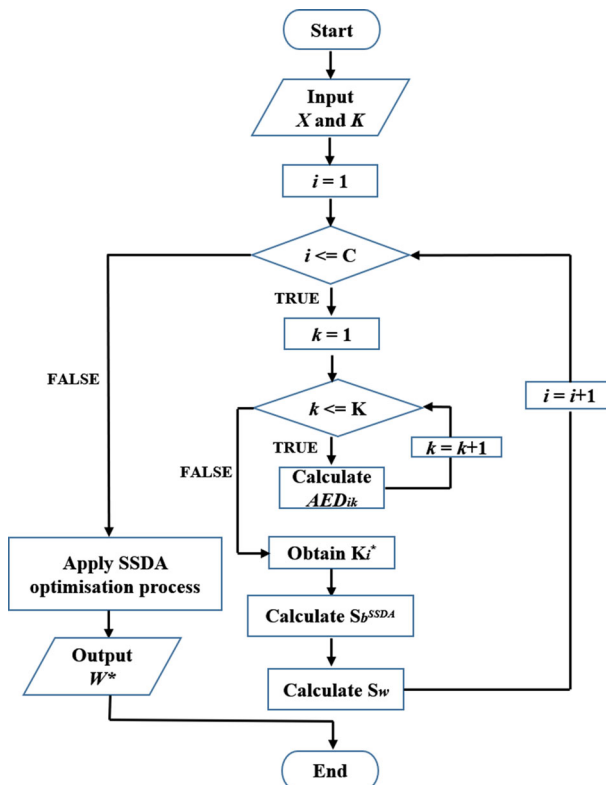


Fig. 2 The flowchart of SSDA algorithm

3 Artificial data

In order to answer the research questions mentioned above, we generate four types of artificial data.

- Type 1, consists of two different classes and samples in each class are from a single multivariate normal distribution. This type is denoted by C2M1.
- Type 2, consists of two different classes and every class has two subclasses of samples generated from two multivariate normal distributions. This type is denoted by C2M2.
- Type 3, consists of two different classes and every class has three subclasses of samples generated from three multivariate normal distributions. This type is denoted by C2M3.
- Type 4, consists of three different classes, and every class has three subclasses of samples generated from three multivariate normal distributions. This type is denoted by C3M3.

The number of variables is one parameter in a multivariate normal distribution, which is set to 30 for all types of artificial data in our studies. Two other important parameters are: the mean μ and covariance σ , which are needed to generate artificial data from a multivariate normal distribution. In our studies, the mean μ is a 1-by-30 vector and the values of the mean vector are integers chosen randomly from the range [1, 10]. Covariance σ is a 30-by-30 diagonal matrix. There are two covariance matrixes for C2M1, one for each class. The values of one covariance matrices for C2M1 are integers chosen randomly from the range [10, 21], and the values of the other covariance matrix are integers chosen randomly from the range [20, 41].

There are four covariance matrices for C2M2, one for each subclasses and two for each class (there are two subclasses in each class). For class one, the values of the covariance matrices for the two subclasses are integers chosen randomly from the range [10, 21], and the values of the covariance matrices for the two subclasses of class two are integers chosen randomly from the range [20, 41].

There are six covariance matrices for C2M3, one for each subclass and three for each class. For class one, the values of the covariance matrices for the three subclasses are integers chosen from the ranges [10, 21] randomly. For class two, the values of the covariance matrices for the three subclasses are integers chosen randomly from the range [20, 41].

There are nine covariance matrices for C3M3, one for each subclass and three for each class. For class one, the values of the covariance matrices for the three subclasses are integers chosen from the ranges [1, 10] randomly. For class two and class three, the values of the covariance matrices for the three subclasses are integers chosen randomly from the ranges [10, 21] and [20, 41], respectively.

In total 10 data sets are generated for each type, and every class of every artificial data set (any type) has 1000 samples. Therefore C2M1, C2M2 and C2M3 each has a total of 2000 samples with 1000 per class. For C2M2 and C2M3, the samples in each class are randomly placed into two and three subclasses respectively according to a probability distribution which varies from data set 1 to 10. C3M3 has a total of 3000 samples with 1000 per class. The samples in each class are randomly placed into three subclasses in the same way as for C2M2 and C2M3. The actual number of samples per subclass are shown in Tables 2, 3 and 4.

4 Multimodality in artificial data

Multiple modalities exist in data. In order to have full insights about the issue of within-class multimodality, various questions can be asked and answered. In the Introduction, some

questions are posed explicitly, and the rest of this paper is to seek answers to these questions. Some questions will be answered using artificial data in this section. Other questions will be answered using real-world data in the next section.

4.1 Q1: Is it necessary to address within-class multimodality?

To answer this question we consider and compare experimentally three approaches in the presence of within-class multimodality:

- separating within-class modalities for every class through the extraction of features by dimensionality reduction methods such as SDA and SSDA;
- merging within-class modalities as a uni-modality for every class in the process of feature extraction using a dimensionality reduction method such as LDA; and
- doing nothing about within-class multimodality and using the original data for classification.

In order to evaluate these three approaches, we conduct experiments using k-nearest neighbour (kNN, $k=1$) as the classifier on all of the artificial data sets. We consider four cases: (1) Original: the original artificial data sets without any processing for dimensionality reduction (2) LDA processed (3) SDA processed (4) SSDA processed. In addition, we use one half of each data set for training and the other half for testing.

Tables 1, 2, 3 and 4 show the experimental results in the four cases on all of the artificial data sets. From these results, we can observe the following:

- It is apparent that SSDA outperforms Original and LDA on all artificial data sets. In particular, SSDA improves classification accuracy over Original by at least 9% on all of the C2M1, C2M2 and C2M3 data sets, and by at least 14% on the C3M3 data sets.
- LDA, SDA and SSDA outperform Original consistently, so dimensionality reduction in the style of LDA can indeed improve classification performance significantly. Whilst this is not new, it indicates that doing nothing about multimodality is suboptimal.
- When there is only one modality per class: it is clear from Table 1 that the differences between LDA, SDA and SSDA do not appear to be significant. This suggests that when

Table 1 Classification accuracy with kNN ($k=1$) of Original, LDA, SDA and SSDA on ten C2M1 data sets

Methods Data sets	Original	LDA	SDA	SSDA
C2M1-1	0.8700	0.9700	0.9750	0.9700
C2M1-2	0.8590	0.9540	0.9640	0.9540
C2M1-3	0.8430	0.9500	0.9580	0.9660
C2M1-4	0.8180	0.9490	0.9610	0.9540
C2M1-5	0.8540	0.9540	0.9540	0.9540
C2M1-6	0.8730	0.9620	0.9650	0.9660
C2M1-7	0.8730	0.9670	0.9750	0.9690
C2M1-8	0.8630	0.9660	0.9660	0.9700
C2M1-9	0.8170	0.9320	0.9380	0.9320
C2M1-10	0.8590	0.9620	0.9620	0.9620

Table 2 Classification accuracy with kNN (k=1) of Original, LDA, SDA and SSDA on ten C2M2 data sets, along with the ratio between the numbers of samples from different subclasses in each class

Data sets	Methods & ratio	Original	LDA	SDA	SSDA	ratio	
						Class One	Class Two
C2M2-1		0.7850	0.8390	0.9190	0.9370	684:316	701:299
C2M2-2		0.8430	0.9070	0.9410	0.9600	676:324	693:307
C2M2-3		0.8630	0.9500	0.9690	0.9750	521:479	508:492
C2M2-4		0.7970	0.8900	0.9430	0.9610	479:521	499:501
C2M2-5		0.8180	0.8770	0.8770	0.9300	491:509	497:503
C2M2-6		0.8530	0.9220	0.9430	0.9520	486:514	512:488
C2M2-7		0.8640	0.9190	0.9560	0.9590	289:711	305:695
C2M2-8		0.8000	0.9020	0.9250	0.9300	274:726	294:706
C2M2-9		0.7600	0.8860	0.9080	0.9100	186:814	208:792
C2M2-10		0.8230	0.9230	0.9270	0.9450	793:207	796:204

there is only one modality per class, doing dimensionality reduction using SDA or SSDA makes little difference from using LDA.

- As for LDA and its variants, we can rank order them in terms of their performance: $LDA \leq SDA \leq SSDA$ on the artificial data sets with within-class multimodality, namely C2M2, C2M3 and C3M3. This suggests that dealing with within-class multimodality the SSDA way is better.
- When there are multiple modalities per class: from Tables 2, 3 and 4, it is clear that doing dimensionality reduction at the subclass level as in SDA or SSDA is better than at the class level as in LDA. Furthermore, SSDA clearly outperforms SDA in these experiments. This suggests that separating subclasses (in other words, reducing the overlap of different subclasses) within every class and at the same time separating all classes is

Table 3 Classification accuracy with kNN (k=1) of Original, LDA, SDA and SSDA on ten C2M3 data sets, along with the ratio between the numbers of samples from different subclasses in each class

Data sets	Methods & ratio	Original	LDA	SDA	SSDA	ratio	
						Class One	Class Two
C2M3-1		0.7720	0.8370	0.8740	0.9250	208:531:261	189:535:276
C2M3-2		0.8380	0.8750	0.8810	0.9490	359:187:454	362:216:422
C2M3-3		0.7450	0.7990	0.8940	0.9220	358:360:282	327:380:293
C2M3-4		0.8090	0.8930	0.9180	0.9490	141:354:505	130:398:472
C2M3-5		0.7830	0.8680	0.9430	0.9490	11:347:642	7:351:642
C2M3-6		0.7850	0.8400	0.8400	0.9190	8:347:645	2:343:655
C2M3-7		0.7960	0.8470	0.8470	0.9290	188:652:160	194:612:194
C2M3-8		0.7830	0.8420	0.8420	0.9330	437:394:169	438:403:159
C2M3-9		0.7840	0.8390	0.8390	0.9200	431:142:427	413:150:437
C2M3-10		0.7710	0.8460	0.8460	0.9060	426:161:413	452:147:401

Table 4 Classification accuracy with kNN (k=1) of Original, LDA, SDA and SSDA on ten C3M3 data sets, along with the ratio between the numbers of samples from different subclasses in each class

Data sets	Methods & ratio				ratio		
	Original	LDA	SDA	SSDA	Class One	Class Two	Class Three
C3M3-1	0.7593	0.8480	0.8947	0.9433	659:127:214	654:155:191	666:131:203
C3M3-2	0.7740	0.8287	0.9053	0.9413	559:114:327	569:115:316	562:107:331
C3M3-3	0.7773	0.8587	0.9080	0.9400	776:147:77	785:152:63	736:174:90
C3M3-4	0.7120	0.7267	0.7267	0.9220	313:278:409	326:237:437	321:246:433
C3M3-5	0.7767	0.8067	0.8067	0.9393	330:262:408	265:287:448	317:284:399
C3M3-6	0.7273	0.7647	0.8800	0.9253	425:243:332	449:242:309	400:260:340
C3M3-7	0.7847	0.8180	0.8920	0.9280	168:435:397	195:425:380	176:441:383
C3M3-8	0.7720	0.8560	0.8560	0.9413	165:405:430	183:412:405	163:450:387
C3M3-9	0.7840	0.8680	0.8893	0.9433	36:613:351	46:622:332	47:601:352
C3M3-10	0.7987	0.8787	0.8787	0.9520	16:500:484	23:462:515	20:487:493

a better approach than simply pulling subclasses in a class from the subclasses of other classes.

- When the number of modalities per class increases: according to Tables 1, 2 and 3, in general the classification accuracy drops in all methods, suggesting that the complexity of the problem increases. This can be seen more clearly in Fig. 3. Interestingly, the margin of performance drop is the smallest with SSDA, suggesting that SSDA is more robust than Original, LDA and SDA when the number of modalities per class changes.

From these observations we can draw the conclusion that it is indeed necessary to deal with the issue of within-class multimodality. Furthermore, this conclusion will be confirmed by using the real data sets in Section 5.

4.2 Q2: How many within-class modalities should we use?

There is a clear difference between SDA and SSDA in terms of classification accuracy as shown in Tables 1, 2, 3 and 4. SDA and SSDA are both trying to separate classes at subclass level but they are different in two ways: (1) how to find the within-class multimodalities; (2) once found, how to make use of these modalities. We examine the first issue in this subsection and discuss the second issue in Subsection 4.3.

SDA uses a stability criterion to find class modalities, whereas SSDA uses a separability criterion. Tables 5, 6, 7 and 8 show the numbers of class modalities found by SDA and

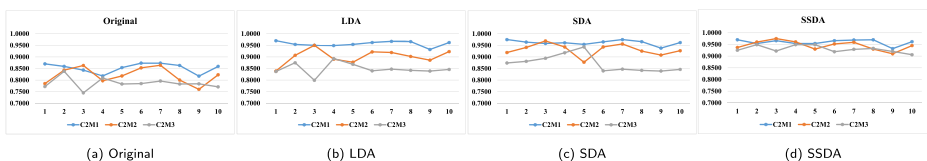


Fig. 3 The classification performance of Original, LDA, SDA and SSDA on ten C2M1 data sets, ten C2M2 data sets and ten C2M3 data sets: In the line charts, the horizontal axis shows the ten data sets from C2M1, C2M2 and C2M3, and the vertical axis shows the classification accuracy

Table 5 The number of subclasses found by SDA and SSDA for each class in the C2M1 data sets

Methods Data sets	SDA		SSDA	
	Class One	Class Two	Class One	Class Two
C2M1-1	3	3	1	1
C2M1-2	3	3	1	1
C2M1-3	3	3	2	2
C2M1-4	4	4	2	2
C2M1-5	1	1	1	1
C2M1-6	3	3	2	2
C2M1-7	2	2	4	4
C2M1-8	1	1	2	2
C2M1-9	2	2	1	1
C2M1-10	1	1	1	1

SSDA for the 10 data sets, of type C2M1, C2M2, C2M3 and C3M3, respectively. It is clear that the numbers are quite different for SDA and SSDA. The numbers found by SSDA in general are quite close to the true numbers of within-class modalities, and identical in most of the data sets. Apart from in a few cases, the numbers found by SDA are quite different to the true numbers.

Furthermore, SSDA can even find true within-class modalities for classes with imbalanced proportions of data between subclasses. For instance, SSDA separates each of Class One, Class Two and Class Three of C3M3-10 into three modalities, when their subclass ratios are 16 : 500 : 484, 23 : 462 : 515 and 20 : 487 : 493 respectively.

All of these observations suggest that (1) the steadily good classification performance is guaranteed by the correct the number of modalities found; and (2) SSDA can find the number of within-class modalities more correctly than SDA, which will be verified on the two face databases in Section 5.2.

Table 6 The number of subclasses found by SDA and SSDA for each class in the C2M2 data sets

Methods Data sets	SDA		SSDA	
	Class One	Class Two	Class One	Class Two
C2M2-1	5	5	4	2
C2M2-2	6	6	2	2
C2M2-3	4	4	2	2
C2M2-4	6	6	2	2
C2M2-5	1	1	3	3
C2M2-6	4	4	2	2
C2M2-7	4	4	3	2
C2M2-8	6	6	2	2
C2M2-9	3	3	2	2
C2M2-10	2	2	4	2

Table 7 The number of subclasses found by SDA and SSDA for each class in the C2M3 data sets

Methods Data sets	SDA		SSDA	
	Class One	Class Two	Class One	Class Two
C2M3-1	8	8	3	3
C2M3-2	10	10	3	3
C2M3-3	3	3	3	3
C2M3-4	6	6	2	3
C2M3-5	3	3	3	3
C2M3-6	15	15	3	3
C2M3-7	1	1	3	3
C2M3-8	1	1	3	3
C2M3-9	1	1	3	3
C2M3-10	1	1	4	4

4.3 Q3: How should we utilise the modalities?

After the multiple within-class modalities are found, we need to utilise them in order to reduce dimensionality for the purpose of effective classification. SDA and SSDA provide different solutions, all based on the LDA optimisation process but with different Fisher objectives. To compare these two solutions, we apply the SDA and SSDA optimisation processes on the artificial data sets consist of within-class modalities (i.e., C2M2, C2M3 and C3M3). In addition, the true number of within-class modalities (True-MN) is used in the both SDA and SSDA. The experimental results are presented in Tables 9, 10 and 11.

From Tables 9, 10 and 11, it is clear that the performance of SSDA with True-MN is consistently higher than SDA with True-MN. Therefore, it suggests that the SSDA optimisation process can better utilise the modalities than the SDA optimisation process. Furthermore, it

Table 8 The number of subclasses found by SDA and SSDA for each class in the C3M3 data sets

Methods Data sets	SDA			SSDA		
	Class One	Class Two	Class Three	Class One	Class Two	Class Three
C3M3-1	4	4	4	3	3	3
C3M3-2	5	5	5	3	3	3
C3M3-3	3	3	3	3	3	3
C3M3-4	1	1	1	3	4	3
C3M3-5	1	1	1	3	3	3
C3M3-6	2	2	2	3	3	3
C3M3-7	5	5	5	3	3	3
C3M3-8	1	1	1	3	3	3
C3M3-9	6	6	6	3	3	3
C3M3-10	1	1	1	3	3	3

Table 9 The classification accuracy with kNN ($k=1$) of SDA and SSDA using the true number of within-class modalities on the C2M2 data sets

Methods Data sets	SDA with True-MN	SSDA with True-MN
C2M2-1	0.9080	0.9260
C2M2-2	0.9350	0.9600
C2M2-3	0.9670	0.9750
C2M2-4	0.9460	0.9610
C2M2-5	0.9230	0.9230
C2M2-6	0.9480	0.9520
C2M2-7	0.9380	0.9570
C2M2-8	0.9300	0.9300
C2M2-9	0.8970	0.9100
C2M2-10	0.9270	0.9380

shows that maximising inter-subclass and inter-class separation at the same time is a worthwhile goal of LDA-like dimensionality reduction when the true modalities are found in data.

5 Multimodality in real data

Separating within-class multimodalities results in good performance on artificial data, when the modality of the data is known. For real-world data, the modality of the data is unknown even if we believe that there should be multimodality, e.g., as in the problem of face recognition discussed in Section 1. Can we obtain real benefits by addressing within-class multimodality in real-world data in the same way as for artificial data? This is the question we want to answer in this section. We consider two types of data. One is general data from

Table 10 The classification accuracy with kNN ($k=1$) of SDA and SSDA using the true number of within-class modalities on the C2M3 data sets

Methods Data sets	SDA with True-MN	SSDA with True-MN
C2M3-1	0.9120	0.9250
C2M3-2	0.9120	0.9490
C2M3-3	0.8940	0.9220
C2M3-4	0.9320	0.9490
C2M3-5	0.9430	0.9490
C2M3-6	0.8920	0.9190
C2M3-7	0.8960	0.9290
C2M3-8	0.8970	0.9330
C2M3-9	0.9000	0.9200
C2M3-10	0.8730	0.8980

Table 11 The classification accuracy with kNN (k=1) of SDA and SSDA using the true number of within-class modalities on the C3M3 data sets

Methods Data sets	SDA with True-MN	SSDA with True-MN
C3M3-1	0.9093	0.9433
C3M3-2	0.9153	0.9413
C3M3-3	0.9080	0.9400
C3M3-4	0.8580	0.9167
C3M3-5	0.9213	0.9393
C3M3-6	0.8933	0.9253
C3M3-7	0.9047	0.9280
C3M3-8	0.9253	0.9413
C3M3-9	0.9220	0.9433
C3M3-10	0.9480	0.9520

the UCI data repository [3]; the other is face image data, as it is intuitively plausible that there is within-class multimodality associated with lighting conditions and head pose.

In our experiments, we consider k-nearest neighbor (kNN, k=1) as the classifier. We conduct a study on the within-class classification problem by focusing on extracting discriminant features in this paper. Some commonly used classifiers have built-in feature selection/extraction functions. For example, Support Vector Machine (SVM) and Decision Tree (DT) select features as part of the learning process. KNN classifier does not have any built-in feature selection/extraction function, so it is selected and used in our experiments. Additionally, we use ten-fold cross-validation as the evaluation framework, and *Estimated Mean Accuracy (EMA)* and *Standard Error of the Mean (SEM)* [9] as the evaluation metrics: $EMA = \frac{\sum_{i=1}^{10} p_i}{10}$, where p_i denotes the percentage of correct classification in the i th fold validation; $SEM = \frac{\delta}{\sqrt{10}}$, where $\delta = \sqrt{\frac{\sum_{i=1}^{10} (p_i - EMA)^2}{9}}$. So, the higher EMA and lower SEM are, the better classification performance is. Moreover, to make the evaluation results more reliable, we ran each experiment 10 times using ten-fold cross-validation, and reported the average EMA (AEMA) and average SEM (ASEM).

5.1 General data

We select eleven UCI data sets using two criteria: (1) all attributes must be numerical; (2) there must be many attributes so that dimensionality reduction is meaningful. General information about the eleven UCI data sets is shown in Table 12.

Furthermore, we compare SSDA and SDA against adaptive local linear discriminant analysis (ALLDA) [20]. To compare ALLDA as fairly as possible, we follow the experimental settings used in the [20] since we do not have the source code of ALLDA. In [20], four UCI data sets are used to test the performance of ALLDA. They are Australian, Heart, Pima and Diabetes, respectively. We can not find the Diabetes data set corresponding to the description in the [20], so we compare SDA and SSDA with ALLDA on the remaining three data sets. The experimental settings used in the [20] are: (1) several samples are randomly selected from every class with same proportion as training data and the rest of samples as testing data. The splits of Australian, Heart and Pima data sets are described in the Table 13;

Table 12 General information about the ten UCI data sets used, where #I denotes the number of instances, #C denotes the number of classes and #A denotes the number of attributes

Name of data set (Acronym)	#I	#C	#A
QSAR Biodegradation (QSAR-B)	1055	2	41
Climate Model Simulation Crashes (CMSC)	540	2	18
Diabetic Retinopathy (DR)	1151	2	19
Multiple Feature-fou (MF-fou)	2000	10	76
Musk(Version 1)-Clean1 (M1-C1)	476	2	166
Parkinsons	195	2	22
Statlog Project (SP)	846	4	18
White Wine Quality (WWQ)	4898	7	11
Yeast	1484	10	8
Isolet	7797	26	617
Vertebral	310	2	6

(2)1-nearest neighbor is used as the classifier and each experiment is conducted using 20 random splits; (3)the mean accuracy (Macc) and standard deviation (Std) are the evaluation of classification performance.

Experimental results are presented in Tables 14 and 15. The experimental results of ALLDA in the Table 15 are cited from [20]. From these results we note the following observations:

- LDA, SDA and SSDA achieve better performance than Original on the majority of the UCI data sets. This furtherly verifies the conclusion drawn by using artificial data sets that it is necessary to deal with the issue of within-class multimodality.
- SSDA achieves better performance than the other three methods on the majority of data sets. In particular, SSDA outperforms LDA on all UCI data sets.
- Compared with Original and LDA, both SDA and SSDA have superior performance on CMSC, DR, MF-fou, Parkinsons, Yeast and Isolet. This suggests that these data sets are likely to have salient within-class multimodalities. Fig. 4 is a visualisation of these data sets in a two-dimensional space by t-SNE [16], where different colours represent different classes. t-SNE is a technique for visualising high-dimensional data sets by giving each sample a location in a two- or three-dimensional space. It can be observed that these data sets comprise different class and some class clusters consist of several clusters, which correspond to within-class modalities. In particular, the presence of multimodality is clear in Parkinsons, where class one consists of several red clusters and class two consists of several cyan clusters.

Table 13 General information and the split about Australian, Heart and Pima data set, where #C denotes the number of classes, #Training denotes the number of training data, #Testing denotes the number of testing data and #A denotes the number of attributes

Name of data set	#C	#Training	#Testing	#A
Australian	2	207	483	14
Heart	2	54	216	13
Pima	2	149	619	8

Table 14 AEMA±ASEM values with kNN (k=1) of Original, LDA, SDA and SSDA on Eleven UCI data sets

Data sets \ Methods	Original	LDA	SDA	SSDA
	AEMA ± ASEM	AEMA ± ASEM	AEMA ± ASEM	AEMA ± ASEM
QSAR-B	0.7928 ± 0.0139	0.7954 ± 0.0111	0.7580 ± 0.0132	0.8381 ± 0.0102
CMSC	0.8895 ± 0.0106	0.9384 ± 0.0087	0.9389 ± 0.0074	0.9454 ± 0.0093
DR	0.6172 ± 0.0129	0.6448 ± 0.0138	0.6451 ± 0.0148	0.6796 ± 0.0122
MF-fou	0.8269 ± 0.0068	0.8152 ± 0.0071	0.8374 ± 0.0067	0.8343 ± 0.0063
M1-C1	0.8578 ± 0.0142	0.7881 ± 0.0193	0.7462 ± 0.0219	0.8814 ± 0.0149
Parkinsons	0.8454 ± 0.0254	0.8389 ± 0.0219	0.8424 ± 0.0255	0.8616 ± 0.0213
SP	0.7020 ± 0.0136	0.7879 ± 0.0116	0.7744 ± 0.0122	0.8313 ± 0.0105
WWQ	0.5980 ± 0.0058	0.6254 ± 0.0062	0.6096 ± 0.0066	0.6339 ± 0.0080
Yeast	0.5238 ± 0.0152	0.5217 ± 0.0137	0.5295 ± 0.0133	0.5328 ± 0.0148
Isolet	0.8967 ± 0.0030	0.9469 ± 0.0260	0.9488 ± 0.0025	0.9594 ± 0.0022
Verbebral	0.8390 ± 0.0188	0.7742 ± 0.0203	0.8265 ± 0.0198	0.8119 ± 0.0229

5.2 Face image data

We conduct face recognition experiments on two widely used face databases: AR face database [17] and FERET face database [21]. Face Recognition is a multi-class classification problem, where each person is regarded as a class. Face recognition attempts to determine whether a face image is from someone in the database when we have a collection of images for each person in the database. A person's set of face images may contain multiple modalities when they are captured in different illumination conditions or head poses. So, the purpose of this study is to test whether the within-class multimodality methods discussed in this paper can bring benefit to this problem.

In our experiments, the images are represented using their pixel values, resulting in large numbers of features. Therefore, our face recognition task becomes a *small sample size* (SSS) problem [24]. To deal with this problem, a two stage PCA + LDA method [1] is used. We use PCA to reduce data dimensionality, retaining principal components that can explain 90% of the variance, before LDA, SDA and SSDA are used. Details of the two face databases used in our experiments are the given below:

- *AR face database*: The AR face database contains frontal-view face images of 126 different persons (70 males and 56 females). Each person was photographed under different lighting conditions (normal lighting, normal lighting and left light on, normal

Table 15 Macc±Std values with kNN (k=1) of Original, LDA, SDA, SSDA and ALLDA on Australian, Heart and Pima data set, where the results of ALLDA are cited from the paper [20]

Data sets \ Methods	Original	LDA	SDA	SSDA	ALLDA
	Macc ± Std	Macc ± Std	Macc ± Std	Macc ± Std	Macc ± Std
Australian	0.6340 ± 0.0163	0.8000 ± 0.0247	0.6262 ± 0.0184	0.8148 ± 0.0164	0.7775 ± 0.0198
Heart	0.6116 ± 0.0279	0.7750 ± 0.0328	0.7303 ± 0.0254	0.7887 ± 0.0207	0.7431 ± 0.0064
Pima	0.6670 ± 0.0167	0.6845 ± 0.0181	0.6945 ± 0.0187	0.6933 ± 0.0162	0.6763 ± 0.0249

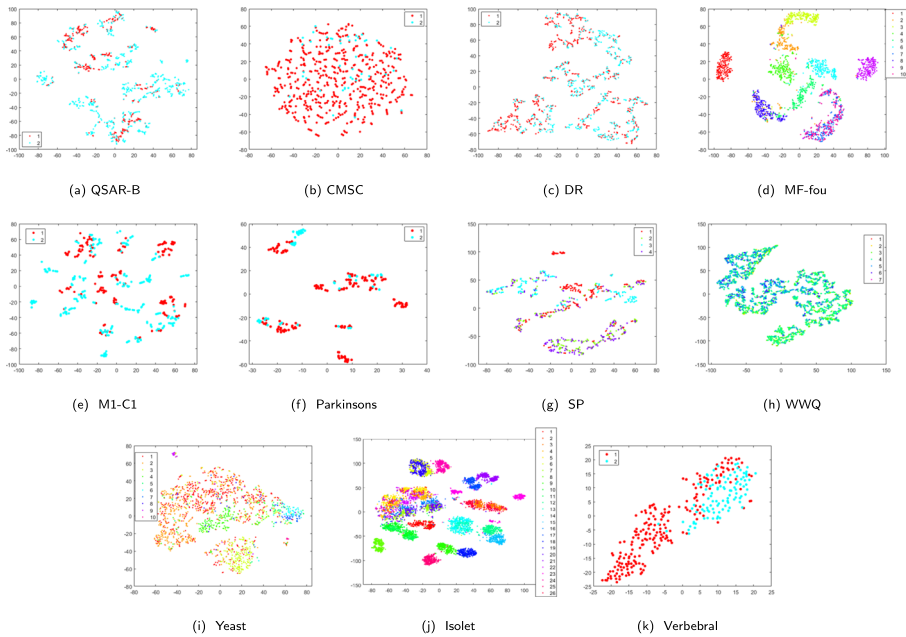


Fig. 4 The data visualisation of QSAR-B, CMSC, DR, MF-fou, M1-C1, Parkinsons, WWQ, SP, Yeast, Isolet and Verbebral in a two-dimensional space

lighting and right light on, normal lighting and both left and right lights on) and distinct facial expressions (neutral expression, smile, anger, and scream), and some images have partial occlusions (sunglasses or scarf). For each person, a total of 13 images were taken in each session for a total of two sessions, which were separated by an interval of two weeks. Hence, there are 26 frontal face images per person. In our experiments, we use a subset of the AR face data set, which comprises 700 face images from 100 persons. We use 7 non-occluded face images of each person taken under different lighting conditions



(a) Examples of images in the AR face database



(b) Examples of images in the FERET face database

Fig. 5 Sample images from the face databases

Table 16 EMA±SEM values with kNN (k=1) of Original, LDA, SDA and SSDA on the AR face database

AR	Methods	Original	LDA	SDA	SSDA
		EMA ± SEM	EMA ± SEM	EMA ± SEM	EMA ± SEM
1		0.5099 ± 0.0158	0.5978 ± 0.0195	0.7806 ± 0.0249	0.8397 ± 0.0076
2		0.5107 ± 0.0117	0.5706 ± 0.0339	0.7822 ± 0.0276	0.8511 ± 0.0182
3		0.5092 ± 0.0173	0.5866 ± 0.0187	0.7188 ± 0.0352	0.8431 ± 0.0159
4		0.5081 ± 0.0195	0.5647 ± 0.0250	0.8052 ± 0.0220	0.8356 ± 0.0195
5		0.5068 ± 0.0203	0.5877 ± 0.0225	0.7682 ± 0.0393	0.8517 ± 0.0156
6		0.5129 ± 0.0186	0.5912 ± 0.0237	0.7814 ± 0.0317	0.8432 ± 0.0129
7		0.5128 ± 0.0168	0.5761 ± 0.0218	0.7366 ± 0.0240	0.8309 ± 0.0103
8		0.5136 ± 0.0132	0.5716 ± 0.0221	0.7402 ± 0.0307	0.8326 ± 0.0208
9		0.5115 ± 0.0162	0.5770 ± 0.0270	0.7830 ± 0.0085	0.8539 ± 0.0127
10		0.5088 ± 0.0183	0.5636 ± 0.0285	0.7939 ± 0.0183	0.8459 ± 0.0125
Average		0.5104 ± 0.0168	0.5787 ± 0.0243	0.7690 ± 0.0262	0.8428 ± 0.0146

and different facial expressions from the first session. Also, Besides, we crop the face part of the image and then resize all images to a standard image size of 80 x 100 pixels (see Fig. 5(a) for some examples). Thus, every face image in the AR database has 8000 features.

- *FERET face database*: The FERET face database includes over 10,000 face images, which have different head poses, lighting conditions and expressions. In our experiments, we use a subset of the FERET face database that consists of 700 images from 100 people, with 7 images per person. Again the face portion of each image is cropped out and normalised to a standard image size of 100 x 100 pixels (see Fig. 5(b) for some examples). We have 10000 features for each image of FERET.

Table 17 EMA±SEM values with kNN (k=1) of Original, LDA, SDA and SSDA on the FERET face database

FERET	Methods	Original	LDA	SDA	SSDA
		EMA ± SEM	EMA ± SEM	EMA ± SEM	EMA ± SEM
1		0.5381 ± 0.0109	0.6065 ± 0.0158	0.6131 ± 0.0179	0.6844 ± 0.0110
2		0.5465 ± 0.0167	0.5646 ± 0.0187	0.6059 ± 0.0196	0.6912 ± 0.0192
3		0.5375 ± 0.0147	0.5866 ± 0.0161	0.5961 ± 0.0218	0.6836 ± 0.0199
4		0.5328 ± 0.0155	0.5898 ± 0.0172	0.6172 ± 0.0153	0.6798 ± 0.0147
5		0.5394 ± 0.0190	0.5979 ± 0.0147	0.6118 ± 0.0234	0.6902 ± 0.0102
6		0.5425 ± 0.0163	0.6038 ± 0.0192	0.6130 ± 0.0221	0.7095 ± 0.0179
7		0.5349 ± 0.0250	0.5867 ± 0.0217	0.6065 ± 0.0160	0.6884 ± 0.0229
8		0.5340 ± 0.0240	0.5913 ± 0.0179	0.6061 ± 0.0166	0.6747 ± 0.0187
9		0.5400 ± 0.0255	0.5961 ± 0.0191	0.6071 ± 0.0185	0.6979 ± 0.0197
10		0.5311 ± 0.0174	0.5710 ± 0.0212	0.6025 ± 0.0223	0.6868 ± 0.0187
Average		0.5377 ± 0.0185	0.5894 ± 0.0182	0.6079 ± 0.0193	0.6887 ± 0.0173

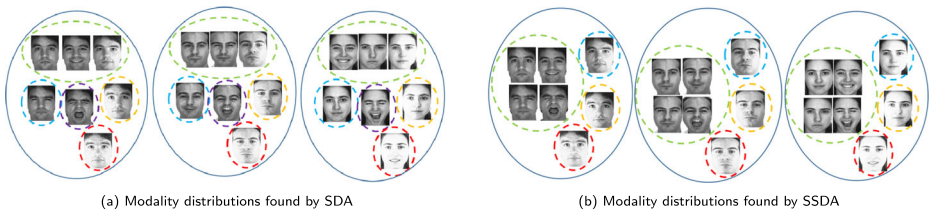


Fig. 6 Examples of modality distributions found by SDA and SSDA on the AR face database, where dotted circles with different colours represent different modalities found by SDA and SSDA. In the (b), the green dotted circle represents the illumination modality with normal lighting; the cyan dotted circle represents the illumination modality with normal lighting and right light on; the orange dotted circle represents the illumination modality with normal lighting and left light on; the red dotted circle represents the illumination modality with normal lighting and both left and right light on

We run experiments with Original, LDA, SDA and SSDA on the AR and FERET face databases 10 times using ten-fold cross-validation. Experimental results are shown in Tables 16 and 17. It is clear that SSDA achieves higher face recognition accuracy than the other three methods on both face image databases; SDA also outperform Original and LDA on both face databases. These results suggest that within-class multimodality does exist in these image databases, and tackling within-class multimodality in the manner of SDA and SSDA does bring benefits.

Furthermore, we want to see what within-class modalities SDA and SSDA can find for AR and FERET, and if the modalities found are consistent with reality. To achieve this, we apply SDA and SSDA on all images of AR and FERET, respectively. Therefore, the maximum number of modalities for each class is set as 7 for both methods since every person only has 7 images in AR and FERET databases. According to the within-modalities found by SDA and SSDA shown in the Figs. 6 and 7, we obtain the following observations:

- From Fig. 6, it readily see that the four modalities found by SSDA correspond to four different illumination conditions existing in the AR database: normal lighting, normal lighting and left light on, normal lighting and right light on, normal lighting and both left and right light on. Although SDA successfully finds two types of illumination modalities: normal lighting and left light on, normal lighting and both left and right light on, it mixes up the images with normal lighting and left light on.
- For FERET database, both SDA and SSDA find different types of within-class modalities for different classes as shown in the Fig. 7. Again, SSDA identifies two types of illumination modalities for each class: normal lighting and low lighting. But SDA fails to find the modality with low lighting for some classes, such as Fig. 7(a)(2).

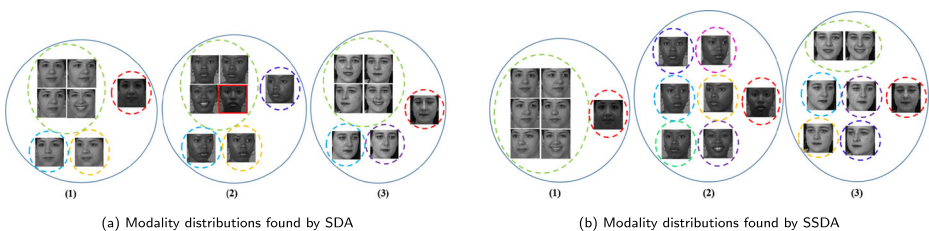


Fig. 7 Examples of modality distributions found by SDA and SSDA on the FERET face database, where dotted circles with different colours represent different modalities found by SDA and SSDA

Table 18 Running time, in seconds, of Original, LDA, SDA and SSDA on eleven UCI data sets and two face databases 10 times using ten-fold cross-validation

Methods Data sets	Original	LDA	SDA	SSDA
QSAR-B	1.7328	2.6739	20.6197	9.7620
CMSC	0.6416	1.0664	6.7935	2.4880
DR	0.7180	1.2667	17.6121	12.0517
MF-fou	1.7049	3.4528	42.9585	10.1369
M1-C1	0.8752	4.4181	17.4592	12.0827
Parkinsons	0.6188	0.9393	8.4054	1.8269
SP	0.6151	1.2622	17.6392	4.5284
WWQ	1.9772	2.8096	87.9767	117.6188
Yeast	0.9834	1.6027	16.4443	9.5505
Isolet	134.1505	1057.5846	1951.6677	1140.4812
Verbebral	0.6817	1.0112	3.8639	3.7453
AR	34.3839	19.2355	63.6227	42.9422
FERET	43.3403	21.9655	306.4279	65.8793

- Apart from identifying the illumination modalities in the FERET database, SSDA can find all correct head pose modalities for some classes (see Fig. 7(b)(3)): frontal modality, leftwards modalities with two different angles and rightwards modalities with two different angles. In addition, SDA also can find some correct head pose modalities for some classes, for example, the modalities represented by the cyan and purple dotted circles shown in the Fig. 7(a)(3).

Therefore, all results from these experiments on two real face databases are consistent with the results on the artificial data sets. When there is within-class multimodality in the data, dealing with the multimodality problem in the manner of either SDA or SSDA is beneficial and, furthermore, the SSDA approach is better than the SDA approach. Interestingly, we have shown that SDA and SSDA offer potential solutions to a challenging problem – face recognition under different lighting and head pose conditions.

5.3 The results: runtime performance

Running results of Original, LDA, SDA and SSDA are shown in Table 18. It is clear that SSDA are slower than Original and LDA but fast than SDA in the most of data sets.

6 Conclusion

Within-class multimodality exists in real-world data and is first studied by [32] and more recently by [26], but many questions are unanswered about within-class multimodality, and its true value is not uncovered fully. This paper presents an extensive study of the within-class multimodality problem through experiments on both artificial data and real data in order to establish a strong case for *within-class multi-modal classification*.

It has been shown using both artificial data and real data that when within-class multimodality is present, maximising between-subclass separation, between-class separation and within-class compactness at the same time in the manner of SDA or SSDA will increase classification performance, with SSDA being the better approach. It is also shown that addressing within-class multimodality this way is optimal if the true number of modalities is known. Interestingly, the experiment on face image databases suggests that SDA and SSDA offer an alternative approach to addressing face recognition under different lighting and head pose conditions.

We believe that a strong case for within-class multimodal classification can be established. We also believe that this classification approach offers a new perspective on improving existing classification algorithms such as Gaussian mixture model and convolutional neural networks, and even devising new classification algorithms. These will be our future research directions.

Acknowledgments The work is partially funded by EU Horizon 2020 project “Analysis System for Gathered Raw Data” (Project Acronym: ASGARD, Project ID: 700381, Project Call: H2020-EU.3.7. - Secure societies - Protecting freedom and security of Europe and its citizens)

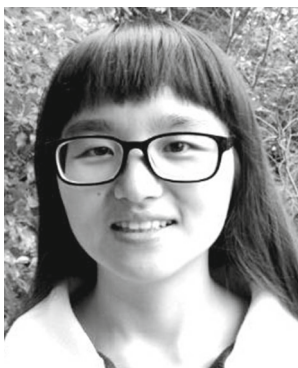
Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Belhumeur PN, Hespanha JP, Kriegman DJ (1997) Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on pattern analysis and machine intelligence* 19(7):711–720
2. Breiman L (2001) Random forests. *Machine learning* 45(1):5–32
3. Dheeru D, Karra Taniskidou E (2017) UCI machine learning repository <http://archive.ics.uci.edu/ml>
4. Em Y, Gag F, Lou Y, Wang S, Huang T, Duan LY (2017) Incorporating intra-class variance to fine-grained visual recognition. In: 2017 IEEE International Conference on Multimedia and Expo (ICME), pp 1452–1457 <https://doi.org/10.1109/ICME.2017.8019371>
5. Fisher RA (1936) The use of multiple measurements in taxonomic problems. *Annals of eugenics* 7(2):179–188
6. Gkalelis N, Mezaris V, Kompatsiaris I (2011) Mixture subclass discriminant analysis. *IEEE Signal Processing Letters* 18(5):319–322
7. Hayashi F (2000) *Econometrics*. Princeton University Press
8. He K, Stankovic L, Liao J, Stankovic V (2016) Non-intrusive load disaggregation using graph signal processing. *IEEE Transactions on Smart Grid* 9(3):1739–1747
9. Huang GB, Mattar M, Berg T, Learned-Miller E (2008) Labeled faces in the wild: A database for studying face recognition in unconstrained environments
10. Jia W, Deng Y, Xin C, Liu X, Pedrycz W (2019) A classification algorithm with linear discriminant analysis and axiomatic fuzzy sets. *Mathematical Foundations of Computing* 2(1):73–81
11. Kaselimi M, Doulamis N, Doulamis A, Voulodimos A, Protopapadakis E (2019) Bayesian-optimized bidirectional lstm regression model for non-intrusive load monitoring. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) IEEE, pp 2747–2751
12. Kaselimi M, Doulamis N, Voulodimos A (2020) Context aware energy disaggregation using adaptive bidirectional lstm models. *IEEE Transactions on Smart Grid*
13. Li CN, Shao YH (2019) Robust and sparse linear discriminant analysis via an alternating direction method of multipliers. *IEEE transactions on neural networks and learning systems*
14. Li H, Zhang L, Huang B, Zhou X (2020) Cost-sensitive dual-bidirectional linear discriminant analysis. *Information Sciences* 510:283–303

15. Louppe G (2014) Understanding random forests: From theory to practice. arXiv preprint arXiv:14077502
16. Lvd M, Hinton G (2008) Visualizing data using t-SNE. *Journal of machine learning research* 9:2579–2605
17. Martínez AM, Kak AC (2001) Pca versus lda. *IEEE transactions on pattern analysis and machine intelligence* 23(2):228–233
18. Martínez AM, Zhu M (2005) Where are linear feature extraction methods applicable? *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27(12):1934–1944
19. Murray D, Stankovic L, Stankovic V, Lulic S, Sladojevic S (2019) Transferability of neural network approaches for low-rate energy disaggregation. In: *ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp 8330–8334
20. Nie F, Wang Z, Wang R, Wang Z, Li X (2020) Adaptive local linear discriminant analysis. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 14(1):1–19
21. Phillips PJ, Moon H, Rizvi SA, Rauss PJ (2000) The feret evaluation methodology for face-recognition algorithms. *IEEE Transactions on pattern analysis and machine intelligence* 22(10):1090–1104
22. Rao CR (1948) The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society Series B (Methodological)* 10(2):159–203
23. Seber GA, Lee AJ (2012) *Linear regression analysis vol 329*. John Wiley & Sons
24. Sharma A, Paliwal KK (2015) Linear discriminant analysis for the small sample size problem: an overview. *International Journal of Machine Learning and Cybernetics* 6(3):443–454
25. Vishwakarma VP, Dalal S (2020) A novel non-linear modifier for adaptive illumination normalization for robust face recognition *Multimedia. Tools and Applications* pp 1–27
26. Wan H, Wang H, Guo G, Wei X (2018) Separability-oriented subclass discriminant analysis. *IEEE transactions on pattern analysis and machine intelligence* 40(2):409–422
27. Wang F, Wang Q, Nie F, Li Z, Yu w, Wang R (2019) Unsupervised linear discriminant analysis for jointly clustering and subspace learning. *IEEE Transactions on Knowledge and Data Engineering*
28. Wen Y, Zhang K, Li Z, Qiao Y (2016) A discriminative feature learning approach for deep face recognition. In: *European conference on computer vision*, Springer, pp 499–515
29. Wen Y, Zhang K, Li Z, Qiao Y (2019) A comprehensive study on center loss for deep face recognition. *Int J Comput Vis* 127(6-7):668–683
30. Ye J, Ji S (2010) *Discriminant analysis for dimensionality reduction: An overview of recent developments* *biometrics: Theory Methods, and Applications* Wiley-IEEE Press New York
31. Zhang W, Zhao X, Morvan JM, Chen L (2018) Improving shadow suppression for illumination robust face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 99:1–1. <https://doi.org/10.1109/TPAMI.2018.2803179>
32. Zhu M, Martínez AM (2006) Subclass discriminant analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28(8):1274–1286

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Huan Wan received the master degree in computer application and technology from the School of Mathematics and Computer Science, Fujian Normal University, China. She is currently pursuing the PhD degree in the School of Computing, Ulster University, UK. Her current research interests are feature extraction, face verification and pattern recognition.



Hui Wang is Professor of Computer Science at Ulster University. His research interests are machine learning, logics and reasoning, combinatorial data analytics, and their applications in image, video, spectra and text analysis. He has over 240 publications in these areas. He played an important role in the development of an algebraic framework for machine learning, Lattice Machine; the original concept of contextual probability, which can be used for uncertainty reasoning/quantification, probability estimation and machine learning; a generic similarity measure, neighbourhood counting, and its specialisations on multivariate data, sequences, tree and graph structures. He is an associate editor of IEEE Transactions on Cybernetics, and an associate editor of International Journal of Machine Learning and Cybernetics. He is the Chair of IEEE SMCS Northern Ireland Chapter, and a member of IEEE SMCS Board of Governors (2010-2013). He is principal investigator of a number of regional, national and international projects in the areas of image/video analytics (Horizon 2020 funded DESIREE and ASGARD, FP7 funded SAVASA, Royal Society funded VIAD), 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.



Bryan 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 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 Senior Lecturer 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 member of the IEEE.



Wing W. Y. Ng received the B.Sc. and Ph.D. degrees from Hong Kong Polytechnic University, Hong Kong, in 2001 and 2006, respectively. He is currently a Professor with the School of Computer Science and Engineering, South China University of Technology, Guangzhou, China. Prof. Ng is the Deputy Director of Guangdong Provincial Key Lab of Computational Intelligence and Cyberspace Information. His current research interests include neural networks, deep learning, generalization error model, and their applications in image and video analysis, smart grid, and smart manufacturing. Prof. Ng is currently an Associate Editor of the International Journal of Machine Learning and Cybernetics. He is the Principle Investigator of three China National Nature Science Foundation Projects and a Program for New Century Excellent Talents in University from the China Ministry of Education. He served as the Board of Governor of the IEEE Systems, Man and Cybernetics Society from 2011 to 2013.

Affiliations

Huan Wan¹  · Hui Wang¹ · Bryan Scotney² · Jun Liu¹ · Wing W. Y. Ng³

Hui Wang
h.wang@ulster.ac.uk

Bryan Scotney
bw.scotney@ulster.ac.uk

Jun Liu
j.liu@ulster.ac.uk

Wing W. Y. Ng
wingng@ieee.org

¹ The School of Computing, Ulster University, Jordanstown, UK

² The School of Computing, Ulster University, Coleraine, UK

³ The School of Computer Science and Engineering, South China University of Technology, Guangzhou, China