文章编号：2095-2163(2021)03-0001-08 中图分类号：TP391.41 文献标志码：A

# 基于自适应结构图的半监督语音情感特征选择

罗 辉，韩纪庆

（哈尔滨工业大学 计算机科学与技术学院，哈尔滨 150001）

**摘 要**：本文研究了语音情感识别中的半监督特征选择问题，即如何利用未标记语音情感数据来帮助选择具有情感判别性的特征。为了解决这个问题，提出了一种新的基于图的半监督特征选择方法。其可以根据标签适应度和流形平滑度，在图上估计一个预测标签矩阵，从而有效地利用标记数据中的标签信息，以及标记数据和未标记数据中的流形结构信息。与现有的基于图的方法相比，该方法能同时进行特征选择和局部结构学习，从而自适应地确定图相似度矩阵。同时，还对图相似度矩阵进行了约束，使其包含更准确的数据结构信息，从而可以选择更有判别性的特征。此外，提出了一种有效的迭代算法来优化该问题。在典型语音情感数据集上的实验结果表明，本文提出的方法是有效的。

**关键词**：语音情感识别；半监督特征选择；自适应结构图

# Semi-supervised speech emotion feature selection based on adaptive structured graph

LUO Hui, HAN Jiqing

(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

**【Abstract】** This paper considers the problem of semi-supervised feature selection in speech emotion recognition, that is, how to use unlabeled speech emotion data to help select the features with emotion discriminability. To address this problem, the paper proposes a novel graph-based semi-supervised feature selection method. The proposed method can estimate a prediction label matrix on the graph with respect to the label fitness and the manifold smoothness, thus it can effectively utilize label information from labeled data as well as a manifold structure information from both labeled and unlabeled data. In comparison with the existing graph-based algorithms, the proposed approach can perform feature selection and local structure learning simultaneously, so the graph similarity matrix can be determined adaptively. At the same time, the paper constrains the similarity matrix to make it contain more accurate data structure information, therefore the proposed approach can select features that are more discriminative. Moreover, an efficient iterative algorithm is proposed to optimize the problem. Experimental results on typical speech emotion datasets show that the proposed method is effective.

**【Key words】** speech emotion recognition; semi-supervised feature selection; adaptive structured graph

## 0 引 言

随着电子技术和计算机技术的发展，人们需要具有情感识别能力的新型语音对话系统。然而，要实现这一目标，还需要克服许多困难。首先，在特征提取方面，尚不清楚哪些语音特征能有效区分语音情感[1]。其次，不同的句子、说话者、说话风格和语速等因素都会引起不同的声学变化，给语音情感识别增加了新的挑战[1-2]。

特征选择不仅可以突出情感所带来的可变性，还能减少情感之外其它因素的干扰，并能保留原始特征的可解释性[1]。根据标签信息的可用性，特征选择方法可分为有监督方法、无监督方法和半监督方法。其中，由于半监督特征选择能够通过同时使用标记和未标记数据来最大化数据的有效性，因此，可将其作为有监督方法和非监督方法之间一个很好的折衷方案[3-5]。

在目前的研究工作中，有许多不同的半监督特征选择方法，总地来说大致可分为 3 种类型，即：滤波式方法、封装式方法和嵌入式方法[6]。其中，由于嵌入式方法在许多方面都具有优势，因此受到了越来越多的关注[5,7-8]。在各种嵌入式特征选择方法中，基于图的半监督特征选择方法因其非参数性、判别性和直推性而受到了广大研究者的青睐[9]。由于局部流形结构在计算效率和表征能力上优于全局结构，因此大多数嵌入式方法都试图发掘数据内部的局部结构，并用其进行特征选择[10]。经典的基于图的半监督特征选择方法主要包含 2 个独立的步

**基金项目**：国家自然科学基金联合基金项目(U1736210)；国家重点研发计划(2017YFB1002102)。

**作者简介**：罗 辉(1989-)，男，博士研究生，主要研究方向：语音情感识别、语音信号处理；韩纪庆(1964-)，男，博士，教授，博士生导师，主要研究方向：语言信号处理、音频信息处理。

**收稿日期**：2020-09-16

骤。首先,通过挖掘局部内部结构信息,来构造相似图矩阵。然后,利用稀疏约束来选择有价值的特征[11-12]。尽管如此,这些方法依然存在一些缺点。一方面,传统的基于图的特征选择方法将构造相似图矩阵和选择特征分成 2 个独立的步骤,其在原始数据中构造的相似图矩阵并不会随着后续的处理而改变。然而,实际数据中往往包含大量的噪声样本和特征,使得所构造的相似图矩阵不可靠[13],从而破坏数据的局部流形结构,最终导致特征选择的性能下降。另一方面,传统方法得到的相似图矩阵通常不能反映理想的邻域结构。根据局部连通性可知,最优相似图矩阵中的连通分量应与类别数保持一致,使得每个连通分量对应一个情感类别[14-15]。然而,简单地使用 $k$ 最近邻准则进行邻域分配,很难得到理想的相似图。

为了解决上述问题,本文提出一种新的基于自适应结构图的半监督语音情感特征选择(Adaptive Structured Graph based Feature Selection, ASGFS)方法。该方法可以同时进行特征选择和局部结构学习,从而选择出更有判别性的语音情感特征。此外,使用基于图拉普拉斯的半监督学习,来更好地利用标记数据和无标记数据的特征选择和标签同时进行预测,在满足标签数据的标签适应度和整个数据结构的流形平滑度的前提下,同时进行特征选择和标签预测。在 3 个典型的语音情感数据上的实验表明,本文所提出的方法能够改善语音情感识别的性能。

## 1 基于图的半监督学习

关于本次研究中所采用的符号表示,可将其描述为:对于任一矩阵 $\boldsymbol{W} \in R^{p\times q}$,$\boldsymbol{w}^i$ 和 $\boldsymbol{w}_j$ 分别表示该矩阵的第 $i$ 行和第 $j$ 列。对于任一向量 $\boldsymbol{w} \in R^p$,$w_i$ 表示其第 $i$ 个元素。令 $\mathbf{1}_p \in R^p$ 表示元素全为 1 的列向量,$\boldsymbol{I}_p \in R^{p\times p}$ 表示单位阵。定义 $\boldsymbol{X} = [\boldsymbol{x}_1,\cdots,\boldsymbol{x}_l,\boldsymbol{x}_{l+1},\cdots,\boldsymbol{x}_n] \in R^{d\times n}$ 为语音情感特征矩阵,其中,$l$ 表示标注样本数,$n$ 为总的训练样本数,$d$ 为原始特征维数。定义 $\boldsymbol{Y}_l \in R^{l\times c}$ 为标签矩阵,其中,$c$ 表示类别数。如果样本 $\boldsymbol{x}_i$ 被标注为第 $j$ 类,则标签矩阵的元素 $y_{ij} = 1$;否则,$y_{ij} = 0$。

基于图的半监督学习方法中,先要定义一个相似图。图中所有节点对应着标注数据和未标注数据,各边反映了样本之间的相似性。这些方法通常假设标签在图上满足平滑性,可表示如下:

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \|\boldsymbol{z}_i - \boldsymbol{z}_j\|_2^2 s_{ij}, \tag{1}$$

其中,$\boldsymbol{Z} = [\boldsymbol{z}_1,\boldsymbol{z}_2,\cdots,\boldsymbol{z}_n]^{\mathrm{T}} \in R^{n\times c}$ 为预测标签矩阵,$\boldsymbol{S} = [s_{ij}] \in R^{n\times n}$ 表示一个预先定义的相似图矩阵。采用高斯函数构造相似图矩阵的方法可表示如下:

$$S_{ij} = \begin{cases} e^{-\frac{\|x_i - x_j\|_2^2}{2\sigma^2}}, & \boldsymbol{x}_i \in N_k(\boldsymbol{x}_j) \text{ 或} \boldsymbol{x}_j \in N_k(\boldsymbol{x}_i), \\ 0, & \text{其它}. \end{cases} \tag{2}$$

其中,$N_k(\boldsymbol{x}_i)$ 表示语音情感样本 $\boldsymbol{x}_i$ 在原始高维空间中的 $k$ 近邻集合。

之前的研究工作分别通过约束标签的适应度和流形的平滑度,介绍了利用局部和全局一致性(Local and global consistency, LGC)[16],以及高斯场和谐波函数(Gaussian fields and harmonic functions, GFHF)[17],在图上估计预测标签矩阵的方法。此外,许多方法利用流形正则来进行半监督扩展[18-19],例如岭回归(ridge regression)、支持向量机(Support Vector Machine, SVM)和线性判别分析(Linear Discriminant Analysis, LDA)。灵活流形嵌入(Flexible Manifold Embedding, FME)是一个半监督流形学习的统一框架[20],可表示为以下优化问题:

$$\min_{\boldsymbol{W},\boldsymbol{Z},\boldsymbol{b}} Tr(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{Z}) + \gamma(\|\boldsymbol{W}\|_F^2 + \mu\|\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W} + \mathbf{1}_n\boldsymbol{b}^{\mathrm{T}} - \boldsymbol{Z}\|_F^2).$$
$$\text{s.t. } \boldsymbol{Z}_l = \boldsymbol{Y}_l \tag{3}$$

其中,$Tr(\cdot)$ 为迹运算;$\boldsymbol{Z} = \begin{pmatrix}\boldsymbol{Z}_l \\ \boldsymbol{Z}_u\end{pmatrix}$;$\boldsymbol{b} \in R^c$ 为偏置项;$\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}$ 为拉普拉斯矩阵;对角阵 $\boldsymbol{D}$ 的第 $i$ 个对角元素 $d_{ii} = \sum_{j=1}^{n} s_{ij}$;系数 $\mu$ 和 $\gamma$ 是用来平衡不同项的参数。通过求解优化问题(3),可以同时学习预测标签矩阵 $\boldsymbol{Z}$ 和分类函数 $h(\boldsymbol{W},\boldsymbol{b})$。

考虑到图拉普拉斯是半监督学习的基础,并且由于语音情感数据通常包含多种结构,可由流形正则进行刻画,因此就可将流形正则用于语音情感分析[21-22]。基于此,本文将提出一种基于图的半监督语音情感特征选择方法,其可在特征选择中自适应地学习局部流形结构。

## 2 基于自适应结构图的半监督特征选择

本节将详细介绍文中所提出的 ASGFS 模型,并针对其给出一种有效的优化求解算法。

### 2.1 ASGFS 模型

与预先在原始特征空间中构造二值图(binary-based graph)或基于核的图(kernel-based graph)不

同，本文采用自适应过程来学习相似图矩阵。根据文献[14]可知，2 个数据相邻的概率可以看作是两者之间的相似性，即语音情感样本 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 之间的距离越小，则成为彼此近邻的概率就越大。假设 2 个样本的标签越接近，其相邻的概率越大，则概率 $s_{ij}$ 可通过求解如下优化问题来确定：

$$\min_{s_{ij}} \sum_{i,j} \|\boldsymbol{z}_i - \boldsymbol{z}_j\|_2^2 s_{ij}^{\alpha}$$
$$\text{s.t. } 0 \leqslant s_{ij} \leqslant 1, \boldsymbol{s}^i \boldsymbol{1}_n = 1. \quad (4)$$

其中，参数 $\alpha$ 用于避免模型得到平凡解(trivial solution)。如果设置 $\alpha = 1$，则只有离样本 $\boldsymbol{x}_i$ 最近的数据点可以概率 1 作为其近邻。根据经验，本文设置参数 $\alpha = 2$。

优化问题(4)中的约束条件可以确保 $s_{ij}$ 是对概率的描述，并且矩阵 $\boldsymbol{S}$ 可以看作与传统图方法一致的相似图矩阵。由于 $\boldsymbol{z}_i$ 的维度远低于 $\boldsymbol{x}_i$，因此自适应结构图在一定程度上不仅能减轻维数灾难问题，而且能有效地捕捉数据的内在结构。

在目前的研究中，许多半监督特征选择方法在选择特征时，都尽量保持局部流形结构而不是全局结构。因此，要利用矩阵 $\boldsymbol{S}$ 来描述局部结构，必须假设其大多数元素都是零。为此，在优化问题(4)的约束条件中添加了一个稀疏约束，即 $\|\boldsymbol{s}^i\|_0 = k$。

在 FME 中，正则项 $\|\boldsymbol{W}\|_F^2$ 用来约束模型的复杂度，避免产生过拟合问题。由于本文的任务是进行特征选择，所以矩阵 $\boldsymbol{W}$ 应具有结构稀疏的特性。为此，考虑利用 $L_{2,1}$ 范数正则[23]来使矩阵 $\boldsymbol{W}$ 具备行稀疏性，则 $\boldsymbol{W}$ 的非零行 $\boldsymbol{w}^i$ 对应于第 $i$ 个特征的权重。因此，矩阵 $\boldsymbol{W}$ 可视为最强判别性特征的组合系数，从而能够实现特征选择。

根据以上分析，本文所提出的 ASGFS 模型可表示成如下优化问题：

$$\min_{\boldsymbol{W},\boldsymbol{Z},\boldsymbol{b},\boldsymbol{S}} \sum_{i,j} \|\boldsymbol{z}_i - \boldsymbol{z}_j\|_2^2 s_{ij}^{\alpha} + \gamma \|\boldsymbol{W}\|_{2,1} +$$
$$\mu \|\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W} + \boldsymbol{1}_n \boldsymbol{b}^{\mathrm{T}} - \boldsymbol{Z}\|_F^2$$
$$\text{s.t. } 0 \leqslant \boldsymbol{S} \leqslant 1, \boldsymbol{s}^i \boldsymbol{1}_n = 1, \|\boldsymbol{s}^i\|_0 = k, \boldsymbol{Z}_l = \boldsymbol{Y}_l. \quad (5)$$

在得到优化问题(5)的最优解之后，可使用 $\boldsymbol{w}^i$ 的 $L_2$ 范数 $\|\boldsymbol{w}\|_2$ 来评估所对应语音情感特征的重要性。在实际应用中，可根据重要程度对所有特征进行排序，从中选择特定维数的特征子集来进行后续的识别任务。

## 2.2　模型的优化算法

由于优化问题(5)的目标函数中包含非平滑的 $L_{2,1}$ 范数正则，因此本节采用交替优化的方式来对其进行求解。

首先，固定变量 $\boldsymbol{W}$、$\boldsymbol{Z}$ 和 $\boldsymbol{S}$ 而只更新变量 $\boldsymbol{b}$，关于变量 $\boldsymbol{b}$ 的优化问题可以表示如下：

$$\min_{\boldsymbol{b}} \|\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W} + \boldsymbol{1}_n \boldsymbol{b}^{\mathrm{T}} - \boldsymbol{Z}\|_F^2, \quad (6)$$

通过对式(6)中的目标函数求偏导数，并令该导数为零，可得如下更新公式：

$$\boldsymbol{b} = \frac{1}{n}(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{1}_n - \boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{1}_n), \quad (7)$$

将式(7)带入到优化问题(5)中，并固定变量 $\boldsymbol{W}$ 和 $\boldsymbol{S}$，可得关于变量 $\boldsymbol{Z}$ 的优化问题：

$$\min_{\boldsymbol{Z}} Tr(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{Z}) + Tr((\boldsymbol{Z} - \boldsymbol{Y})^{\mathrm{T}}\boldsymbol{U}(\boldsymbol{Z} - \boldsymbol{Y})) +$$
$$\mu Tr((\boldsymbol{H}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W} - \boldsymbol{H}\boldsymbol{Z})^{\mathrm{T}}(\boldsymbol{H}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W} - \boldsymbol{H}\boldsymbol{Z})), \quad (8)$$

其中，$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{Y}_l \\ \boldsymbol{0} \end{pmatrix} \in R^{n\times c}$；$\boldsymbol{U} \in R^{n\times n}$ 是对角阵，如果语音情感样本 $\boldsymbol{x}_i$ 为标注数据，则对角元素 $u_{ii} = \infty$；否则，$u_{ii} = 1$；$\boldsymbol{L}$ 为对应的拉普拉斯矩阵，矩阵 $\boldsymbol{H} = \boldsymbol{I}_n - \frac{1}{n}\boldsymbol{1}_n\boldsymbol{1}_n^{\mathrm{T}}$ 满足 $\boldsymbol{H} = \boldsymbol{H}^{\mathrm{T}} = \boldsymbol{H}^2$。

通过对式(8)中的目标函数求偏导数，并令该导数为零，可得如下更新公式：

$$\boldsymbol{Z} = (\boldsymbol{L} + \boldsymbol{U} + \mu\boldsymbol{H})^{-1}(\boldsymbol{U}\boldsymbol{Y} + \mu\boldsymbol{H}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}), \quad (9)$$

将式(7)和式(9)带入到优化问题(5)中，并固定变量 $\boldsymbol{S}$，可得关于变量 $\boldsymbol{W}$ 的优化问题：

$$\min_{\boldsymbol{W}} Tr(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{W}) - 2Tr(\boldsymbol{G}^{\mathrm{T}}\boldsymbol{W}) + \gamma \|\boldsymbol{W}\|_{2,1}, \quad (10)$$

其中，$\boldsymbol{M} = \boldsymbol{X}\boldsymbol{H}(\mu\boldsymbol{I}_n - \mu^2\boldsymbol{P})\boldsymbol{H}\boldsymbol{X}^{\mathrm{T}}$，$\boldsymbol{G} = \mu\boldsymbol{X}\boldsymbol{H}\boldsymbol{P}\boldsymbol{U}\boldsymbol{Y}$，$\boldsymbol{P} = (\boldsymbol{L} + \boldsymbol{U} + \mu\boldsymbol{H})^{-1}$，$\boldsymbol{Q} = \boldsymbol{U}\boldsymbol{Y} + \mu\boldsymbol{H}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}$。

通过对式(10)中的目标函数求偏导数，并令该导数为零，可得如下更新公式：

$$\boldsymbol{W} = (\boldsymbol{B}\boldsymbol{M} + \gamma\boldsymbol{I}_n)^{-1}\boldsymbol{B}\boldsymbol{G}, \quad (11)$$

其中，$\boldsymbol{B}$ 为对角阵，其对角元素 $b_{ii} = 2\|\boldsymbol{w}^i\|_2$。由于矩阵 $\boldsymbol{B}$ 与优化变量 $\boldsymbol{W}$ 有关，因此需要利用迭代算法进行求解[11]。

固定变量 $\boldsymbol{W}$、$\boldsymbol{b}$ 和 $\boldsymbol{Z}$ 而只更新变量 $\boldsymbol{S}$。由于优化问题(5)中关于变量 $\boldsymbol{S}$ 的约束条件是作用在其每一行上的，所以可以逐行更新变量 $\boldsymbol{S}$。下面以第 $i$ 行为例，阐述其更新过程。关于变量 $\boldsymbol{s}^i$ 的优化问题可表示如下：

$$\min_{\boldsymbol{s}^i} \sum_j \|\boldsymbol{z}_i - \boldsymbol{z}_j\|_2^2 s_{ij}^{\alpha}$$
$$\text{s.t. } \boldsymbol{0} \leqslant \boldsymbol{s}^i \leqslant \boldsymbol{1}_n, \boldsymbol{s}^i \boldsymbol{1}_n = 1, \|\boldsymbol{s}^i\|_0 = k, \quad (12)$$

如果忽略约束条件 $\|\boldsymbol{s}^i\|_0 = k$ 和 $\boldsymbol{0} \leqslant \boldsymbol{s}^i \leqslant \boldsymbol{1}_n$，则拉格朗日函数可表示为：

$$\mathcal{L}(s_{ij}) = \sum_j (\|\boldsymbol{z}_i - \boldsymbol{z}_j\|_2^2 s_{ij}^{\alpha}) - \eta\left(\sum_j s_{ij} - 1\right), \quad (13)$$

令 $v_{ij} = \|\boldsymbol{z}_i - \boldsymbol{z}_j\|_2^2$，根据 Karush-Kuhn-Tucker 条件[24]可知，$s_{ij}$ 的最优解可表示如下：

$$s_{ij} = \frac{(v_{ij})^{\frac{1}{1-\alpha}}}{\sum_{t=1}^{n}(v_{it})^{\frac{1}{1-\alpha}}}, \tag{14}$$

需要说明的是，上面的推导过程忽略了约束条件 $\boldsymbol{0} \leqslant \boldsymbol{s}^i \leqslant \boldsymbol{1}_n$。然而，由于 $v_{ij} \geqslant 0$，容易验证，式(14)中的结果满足该约束。总之，当变量 $\boldsymbol{Z}$ 固定时，可以通过式(14)来计算变量 $\boldsymbol{S}$ 的闭式解。

现在考虑另外一个约束条件 $\|\boldsymbol{s}^i\|_0 = k$，其目的是使得向量 $\boldsymbol{s}^i$ 中只有 $k$ 个非零元素。由于需要使式(12)中的目标函数最小，所有只需要优化其中 $k$ 个非零元素。同时，$v_{ij}$ 越大，目标函数的值就越大。因此，可以将 $\boldsymbol{s}^i$ 中与最大的 $n-k$ 个 $v_{ij}$ 相对应的元素设为零，而只更新其余的 $k$ 个元素。具体来说，假设最小的 $k$ 个 $v_{ij}$ 为 $\{v_{ij_1}, v_{ij_2}, \cdots, v_{ij_k}\}$（不包含 $v_{ii}$），那么优化问题(12)的最优解 $\boldsymbol{s}^i$ 为：

$$s_{ij} = \begin{cases} \dfrac{(v_{ij_m})^{\frac{1}{1-\alpha}}}{\sum_{t=1}^{k}(v_{ij_t})^{\frac{1}{1-\alpha}}}, & m = 1,2,\cdots,k, \\ 0, & \text{其它.} \end{cases} \tag{15}$$

最终，ASGFS 模型的求解算法总结在算法 1 中。

**算法 1　自适应结构图特征选择算法**

**输入**　语音情感数据 $\boldsymbol{X} \in R^{d\times n}$，标签 $\boldsymbol{Y} \in R^{n\times c}$，参数 $\alpha$、$\gamma$、$\mu$

**输出**　权值 $\boldsymbol{W} \in R^{d\times c}$

1 初始化 $\boldsymbol{B} = \boldsymbol{I}_d$ 以及相似图矩阵 $\boldsymbol{S}$[11]；

2 计算对角阵 $\boldsymbol{U} \in R^{n\times n}$；

3 计算矩阵 $\boldsymbol{H} = \boldsymbol{I}_n - \frac{1}{n}\boldsymbol{1}_n\boldsymbol{1}_n^{\mathrm{T}}$；

4 while 不收敛 do

5　计算图拉普拉斯矩阵 $\boldsymbol{L} \in R^{n\times n}$；

6　计算 $\boldsymbol{P} = (\boldsymbol{L} + \boldsymbol{U} + \mu\boldsymbol{H})^{-1}$；

7　计算 $\boldsymbol{M} = \boldsymbol{X}\boldsymbol{H}(\mu\boldsymbol{I}_n - \mu^2\boldsymbol{P})\boldsymbol{H}\boldsymbol{X}^{\mathrm{T}}$；

8　计算 $\boldsymbol{G} = \mu\boldsymbol{X}\boldsymbol{H}\boldsymbol{P}\boldsymbol{U}\boldsymbol{Y}$；

9　求解优化问题(10)来更新 $\boldsymbol{W}$；

10　利用式(15)更新 $\boldsymbol{S}$；

11 end

## 3　实验与分析

### 3.1　数据集

本节将在 3 个典型的语音情感数据集上验证 ASGFS 方法的性能，包括 Berlin[25]、eNTERFACE[26]和 CASIA[27]。这些数据集记录了各种离散的情绪状态，例如愤怒、快乐、悲伤等。在语音情感特征提取方面，采用 2010 副语言挑战赛的配置，并利用开源工具 openSMILE 进行特征提取[28]。首先，为每个情感音频文件提取 34 个低阶特征（Low-level Descriptors，LLDs），例如音高、梅尔倒谱系数和响度等，并计算其一阶差分，得到 68 个低阶特征表示。然后，将 19 个统计函数部分或全部作用于每一个低阶特征上，得到超音段特征。此外，还为每个情感音频文件提取音高的起始时间以及会话的持续时间。最终，得到 1 582 维的语音情感特征表示。在提取特征之后，采用说话人依赖的归一化策略，对数据进行独立的预处理，并将每个特征值标准化，使其均值为 0，标准差为 1。

### 3.2　实验设置

在数据库的划分方面，首先采用说话人依赖的策略，将各数据集中每个类的样本随机分为 2 部分。其中，一半作为训练数据，另一半作为测试数据。然后，分别将训练集中每个类 5%、10% 和 15% 的样本作为半监督学习中的标注数据，其余的作为未标注数据。

为了验证特征选择方法的有效性，利用基于径向基（Radial Basis Function，RBF）核函数的支持向量机（Support Vector Machine，SVM）和随机森林（Random-Forest，RF）作为分类器来评价所选特征的分类性能，并采用是非加权平均召回率（Unweighted Average Recall，UAR）作为性能评价指标。其计算公式如下：

$$UAR = \frac{1}{c}\sum_{k=1}^{c}\frac{|\boldsymbol{x}: \boldsymbol{x} \in D_k \cap \hat{y} = y|}{\boldsymbol{x}: \boldsymbol{x} \in D_k}. \tag{16}$$

其中，$y$ 表示样本 $\boldsymbol{x}$ 的真实标签；$\hat{y}$ 表示分类器对该样本的预测标签；$D_k$ 表示属于第 $k$ 类的样本集合。

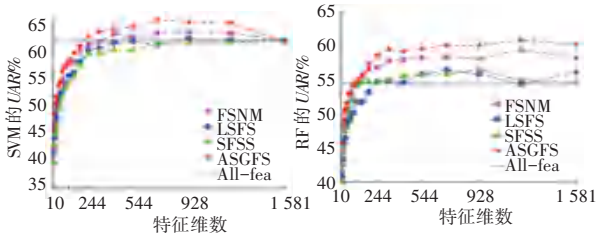
本节使用全部原始特征的分类结果作为基线（记作 All-fea）。除此之外，用于对比的特征选择方法主要包括：基于 $L_{2,1}$ 范数最小化的有监督特征选择（Feature Selection via $L_{2,1}$ - Norms Minimization，FSNM）方法[23]、局部敏感特征选择（Locality Sensitive Feature Selection，LSFS）方法[29]、以及结构化稀疏特征选择（Structural Feature Selection with Sparsity，SFSS）方法[11]。

在参数设置方面，对于所有采用正则技术的方法，各正则化参数的取值范围为{0.001，0.01，0.1，1，10，100，1 000}。对于所有需要构建邻接 Laplacian 图矩阵的方法，最近邻个数 $k$ 固定取值为 5。由于

如何确定所选特征的最优数量仍然是特征选择研究中一个亟待解决的问题，因此本文采用在对数域的[10，1 582]区间内取 20 个数值作为所选特征的维数，并评估每个特征维数的性能。此外，为了更好地反映各方法的性能，每个数据集均进行 10 次独立的采样，以得到不同的训练集和测试集，并在其上验证各方法的性能，将 10 次结果的均值作为最终的性能。

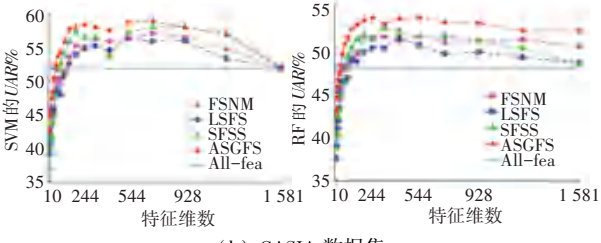
### 3.3　性能对比

图 1～图 3 分别展示了不同特征维数时，SVM 和RF 在 5%、10% 和 15%标注数据上的分类结果。



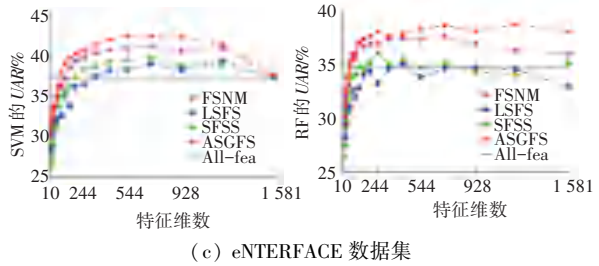

（a）Berlin 数据集

（a）Berlin dataset




（b）CASIA 数据集

（b）CASIA dataset

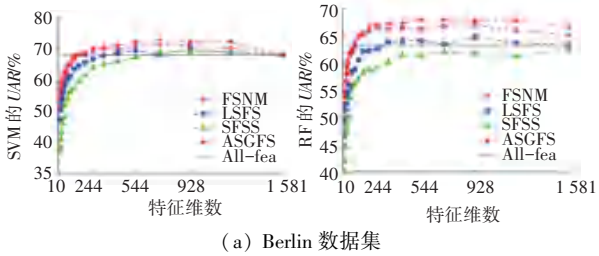



（c）eNTERFACE 数据集

（c）eNTERFACE dataset

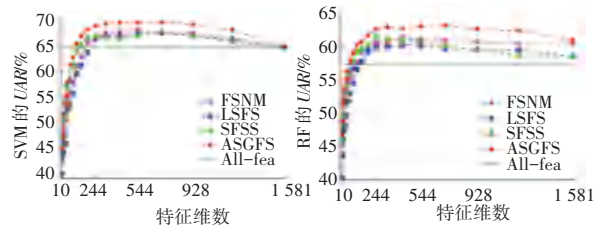
**图 1　SVM 和 RF 分类器在 5%标注数据上的性能曲线**

**Fig. 1　Performance curves of SVM and RF classifier on 5% labeled data**



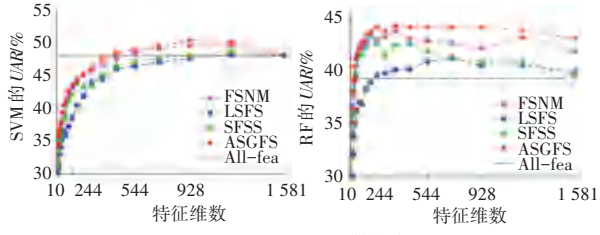

（a）Berlin 数据集

（a）Berlin dataset




（b）CASIA 数据集

（b）CASIA dataset

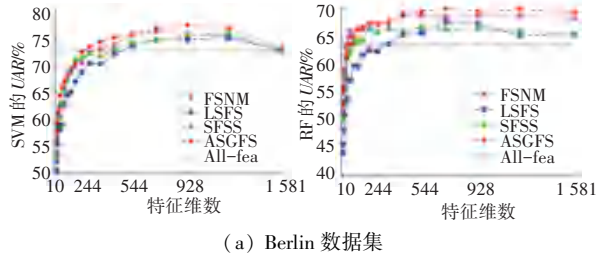



（c）eNTERFACE 数据集

（c）eNTERFACE dataset

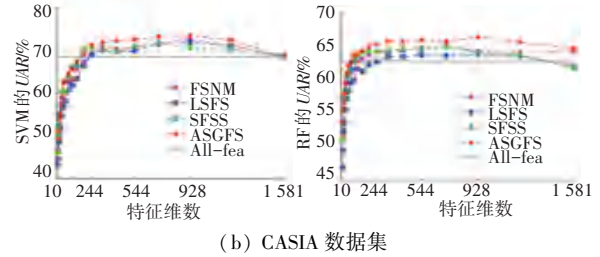
**图 2　SVM 和 RF 分类器在 10%标注数据上的性能曲线**

**Fig. 2　Performance curves of SVM and RF classifier on 10% labeled data**



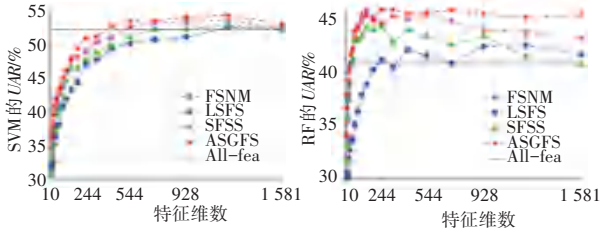

（a）Berlin 数据集

（a）Berlin dataset




（b）CASIA 数据集

（b）CASIA dataset




（c）eNTERFACE 数据集

（c）eNTERFACE dataset

**图 3　SVM 和 RF 分类器在 15%标注数据上的性能曲线**

**Fig. 3　Performance curves of SVM and RF classifier on 15% labeled data**

从图 1~图 3 中的结果可看出,当选择的特征数量较少时,所有特征选择方法的识别性能低于 All-fea的性能。主要原因在于,这些特征丢失了大量对情感识别有用的信息。随着特征维数的增加,所有特征选择方法的性能整体呈现上升趋势。并且,在特征维数的较大变化范围内,都能取得明显优于 All-fea 的性能。这说明原始特征中包含不相关和冗余特征,导致对语音识别系统的性能产生负面影响。此外,在所有的特征选择方法中,ASGFS 方法的整体性能最优。在特征维数相同时,其识别性能优于其它对比方法。而且,其能以最少的特征维数来获得与其它方法相近的性能。因此,本文所提出的方法可以选择更具判别性的语音情感特征。

根据图 1~图 3 的结果,总结了各方法的最高精度参见表 1、表 2。

**表 1 不同方法在 SVM 分类器上的性能对比**

**Tab. 1 Performance comparison of different methods on SVM classifier**

| 数据集 | 标注数据量/% | All-fea | FSNM | LSFS | SFSS | ASGFS |
|---|---|---|---|---|---|---|
| Berlin | 5 | 61.24 | 64.84 | 64.01 | 62.58 | **65.94** |
| | 10 | 68.02 | 72.95 | 71.47 | 70.09 | **73.69** |
| | 15 | 72.67 | 76.68 | 76.64 | 77.19 | **78.19** |
| CASIA | 5 | 51.90 | 58.29 | 58.19 | 59.88 | **60.50** |
| | 10 | 64.50 | 68.79 | 68.21 | 69.04 | **70.10** |
| | 15 | 68.50 | 73.48 | 73.40 | 73.04 | **74.23** |
| eNTERFACE | 5 | 37.20 | 41.79 | 40.70 | 40.66 | **43.04** |
| | 10 | 48.09 | 50.00 | 49.03 | 49.15 | **50.56** |
| | 15 | 52.29 | 54.02 | 53.42 | 53.64 | **54.79** |

**表 2 不同方法在 RF 分类器上的性能对比**

**Tab. 2 Performance comparison of different methods on RF classifier**

| 数据集 | 标注数据量/% | All-fea | FSNM | LSFS | SFSS | ASGFS |
|---|---|---|---|---|---|---|
| Berlin | 5 | 54.47 | 60.41 | 58.99 | 58.80 | **61.61** |
| | 10 | 63.32 | 68.85 | 66.73 | 64.61 | **69.77** |
| | 15 | 64.42 | 70.51 | 69.40 | 69.35 | **71.06** |
| CASIA | 5 | 47.85 | 52.94 | 52.29 | 53.60 | **54.65** |
| | 10 | 57.27 | 62.98 | 61.81 | 62.73 | **63.81** |
| | 15 | 61.37 | 65.17 | 64.69 | 65.71 | **66.17** |
| eNTERFACE | 5 | 34.69 | 38.59 | 36.40 | 37.46 | **39.50** |
| | 10 | 38.81 | 44.57 | 41.99 | 44.29 | **45.23** |
| | 15 | 40.83 | 47.12 | 44.04 | 45.86 | **47.61** |

表 1、表 2 中,粗体数字表示在所有方法中表现最优。从结果可以看出:

(1) 在 2 个分类器中,随着标记数据的增加,所有对比方法在各数据集上的识别性能都会提高。

(2) 所有特征选择方法在 SVM 和 RF 上的识别性能都优于基线系统,说明特征选择可以提高语音情感分类的性能。

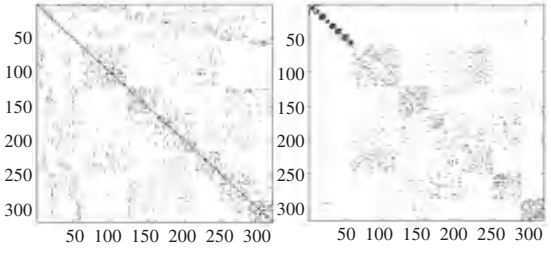
(3) 对于 Berlin 数据集和 eNTERFACE 数据集,有监督特征选择方法 FSNM 优于半监督特征选择方法 LSFS 和 SFSS,这说明在原始特征空间中所构造的相似图可能会对特征选择的性能产生负面影响。

(4) 对于 CASIA 数据集,在大多数情况下,LSFS 和 SFSS 方法的性能都优于 FSNM 方法,这意味着在原始特征空间中所构造的相似图可以在一定程度上刻画该数据的内在结构信息。

(5) 在 3 种不同的标注数据量中, ASGFS 方法在 2 种分类器上的性能都是最优的。相比于基线系统,该方法有着大约 10%的性能提升;相比于 FSNM 方法,有着 2%的性能提升;相比于 LSFS 方法,有着 4%的性能提升;相比于 SFSS,有着 3%的性能提升。主要因为,ASGFS 方法能同时进行特征选择和局部结构学习,从而选择更具判别性的语音情感特征。

### 3.4 图相似度矩阵分析

本节将对 ASGFS 方法所得到的自适应结构图进行分析,并与传统的根据高斯函数构建的图[11]进行对比。图 4~图 6 分别展示了 Berlin、CASIA 和 eNTERFACE 数据集的 2 种不同的图相似度矩阵。从结果可以看到,与传统的方法相比,ASGFS 方法所得到的自适应结构图能够更清晰、更准确地刻画出数据内部的结构信息,从而可以利用其来帮助选择更具判别性的语音情感特征。这也进一步解释了 ASGFS 方法的性能优于其它对比方法的原因。

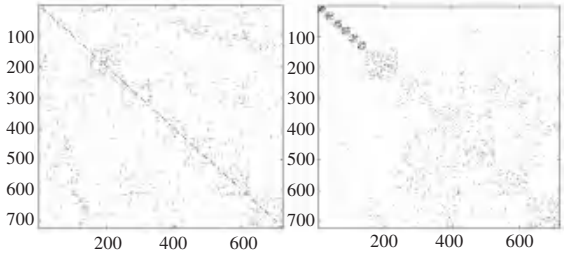



(a) 高斯函数 (b) 自适应结构图

(a) Gaussian function (b) Adaptive structured graph

**图 4 Berlin 数据集上的相似度矩阵对比**

**Fig. 4 Comparison of similarity matrix on Berlin dataset**

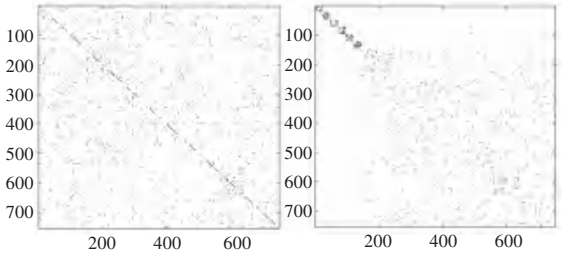



(a) 高斯函数 (b) 自适应结构图

(a) Gaussian function (b) Adaptive structured graph

**图 5 CASIA 数据集上的相似度矩阵对比**

**Fig. 5 Comparison of similarity matrix on CASIA dataset**




（a）高斯函数　（b）自适应结构图

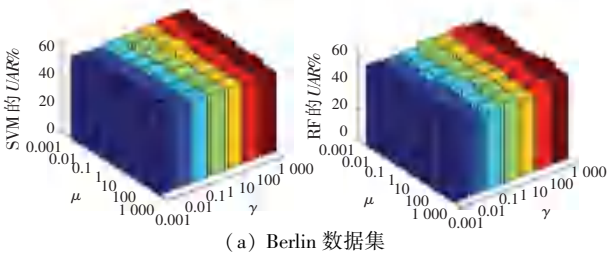
（a）Gaussian function　（b）Adaptive structured graph

**图 6　eNTERFACE 数据集上的相似度矩阵对比**

**Fig. 6　Comparison of similarity matrix on eNTERFACE dataset**

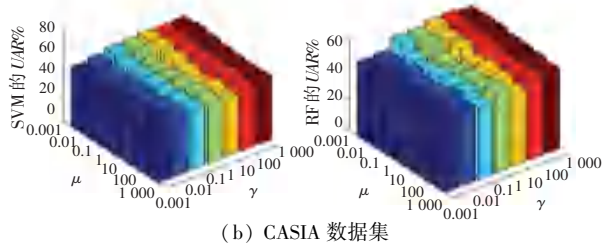
### 3.5　参数敏感性分析

本节将分析 ASGFS 方法对各参数的敏感性。该方法共包含 2 个正则参数：$\gamma$ 和 $\mu$，分别控制着组稀疏约束和分类损失函数对模型的影响程度。图 7 展示了 ASGFS 方法在各参数取不同值时，使用 5%标记数据进行训练的语音情感识别模型的性能。从图 7 中的结果可以看出，不同的参数取值有着不同的识别性能。在 Berlin 数据集上，ASGFS 方法对于参数 $\gamma$ 和 $\mu$ 的不同取值有着较强的鲁棒性。在 CASIA 数据集上，当参数 $\gamma$ 的取值大于 $\mu$ 时，ASGFS 方法的识别性能更优。与之相反，在 eNTERFACE 数据集上，当参数 $\gamma$ 的取值小于 $\mu$ 时，ASGFS 方法能取得更好的性能。



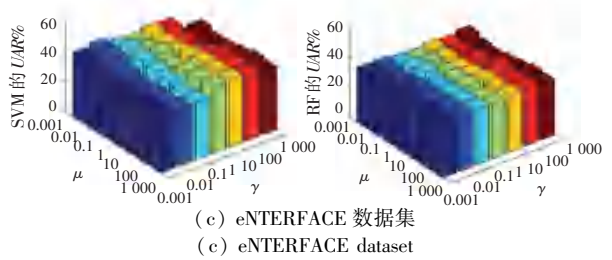

（a）Berlin 数据集

（a）Berlin dataset




（b）CASIA 数据集

（b）CASIA dataset




（c）eNTERFACE 数据集

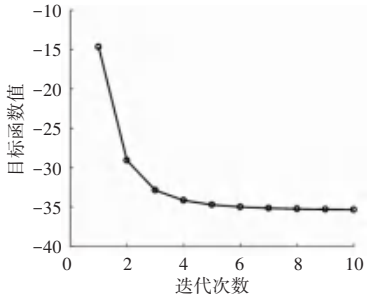
（c）eNTERFACE dataset

**图 7　正则参数 $\gamma$ 和 $\mu$ 的敏感性分析**

**Fig. 7　Sensitivity analysis on regularization parameters $\gamma$ and $\mu$**

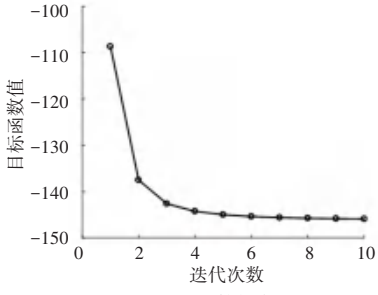
### 3.6　收敛性分析

本节通过实验的方式来研究优化算法 1 的收敛性。在求解 ASGFS 的过程中，通过记录每次迭代后的目标函数值，得到算法的收敛曲线，如图 8 所示。由于算法 1 在不同数量的标注数据上的收敛性是一致的，为简洁起见，图 8 中只展示了 5%标记数据的结果。从图 8 中可以看到，算法 1 是收敛的，且收敛速度非常快。一般来说，算法在 10 次迭代之内就能收敛到一个稳定点。



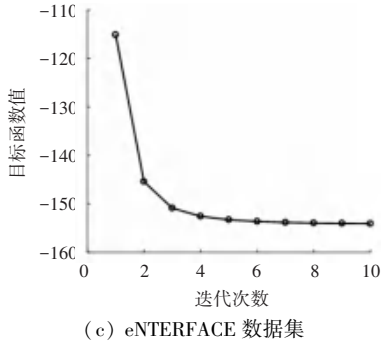

（a）Berlin 数据集

（a）Berlin dataset




（b）CASIA 数据集

（b）CASIA dataset




（c）eNTERFACE 数据集

（c）eNTERFACE dataset

**图 8　ASGFS 方法在 3 个数据集上的收敛曲线**

**Fig. 8　Convergence curves of ASGFS on the three datasets**

## 4　结束语

本文提出了一种新的半监督语音情感特征选择方法。该方法将组稀疏约束、流形正则和直推式分类整合到一个联合特征选择模型中，并且能够同时

进行特征选择和局部结构学习,从而得到自适应结构的图。在 3 个离散语音情感数据集上的实验表明,本文所提出的方法能够选择更具判别性的语音情感特征,从而改善语音情感识别系统的性能。

## 参考文献


[1] AYADI M E, KAMEL M S, KARRAY F. Survey on speech emotion recognition: Features, classification schemes, and databases[J]. Pattern Recognition, 2011, 44(3): 572-587.

[2] PARK J S, KIM J H, OH Y H. Feature vector classification based speech emotion recognition for service robots [J]. IEEE Transactions on Consumer Electronics, 2009, 55(3): 1590-1596.

[3] HAN Y, PARK K, LEE Y K. Confident wrapper-type semi-supervised feature selection using an ensemble classifier[C] // Proceedings of Artificial Intelligence, Management Science and Electronic Commerce. Deng Feng, China: IEEE, 2011: 4581-4586.

[4] LV S, JIANG H, ZHAO L. Manifold based fisher method for semi-supervised feature selection[C] // Proceedings of International Conference on Fuzzy Systems and Knowledge Discovery. Shenyang, China:IEEE, 2013: 664-668.

[5] WANG Jinyan, YAO Jin, SUN Yijun. Semi-supervised local-learning-based feature selection [C] // Proceedings of International Joint Conference on Neural Networks. Beijing, China: IEEE, 2014: 1942-1948.

[6] GUYON I, ELISSEEFF A. An introduction to variable and feature selection[J]. Journal of Machine Learning Research, 2003, 3(6): 1157-1182.

[7] XU Z, KING I, LYU M R, et al. Discriminative semi-supervised feature selection via manifold regularization [J]. IEEE Transactions on Neural Networks, 2010, 21(7): 1033-1047.

[8] ZENG Z, WANG X, ZHANG J, et al. Semi-supervised feature selection based on local discriminative information [J]. Neurocomputing, 2016,172(JANA15PTA1): 102-109.

[9] ZHU X. Semi-supervised learning literature survey[J]. Computer Science, 2008, 37(1): 63-77.

[10] SILVA V D, TENENBAUM J B. Global versus local methods in nonlinear dimensionality reduction [C] // Proceedings of Advances in Neural Information Processing Systems 15. Vancouver, British Columbia, Canada: Nips, 2002: 1959-1966.

[11] MA Z, NIE F, YANG Y, et al. Discriminating joint feature analysis for multimedia data understanding[J]. IEEE Transactions on Multimedia, 2012, 14(6): 1662-1672.

[12] SHI C, RUAN Q, AN G. Sparse feature selection based on graph Laplacian for web image annotation [J]. Image and Vision Computing, 2014, 32(3): 189-201.

[13] WANG D, NIE F, HUANG H. Feature selection via global redundancy minimization[J]. IEEE Transactions on Knowledge and Data Engineering, 2015, 27(10): 2743-2755.

[14] NIE F, WANG X, HUANG H. Clustering and projected clustering with adaptive neighbors[C] // Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2014: 977-986.

[15] NIE F, ZHU W, LI X. Unsupervised feature selection with structured graph optimization [C]// Proceedings of Thirtieth AAAI Conference on Artificial Intelligence. Phoenix, Arizona, USA: AAAI, 2016: 1302-1308.

[16] ZHOU D Y, BOUSQUET O, LAL T N, et al. Learning with local and global consistency[M]// THRUN S, SAUL L, SCHLOPF Proceedings of Advances in Neural Information Processing Systems16. Cambridge: MIT Press, 2004: 321-328.

[17] ZHU X, GHAHRAMANI Z B, LAFFERTY J D. Semi-supervised learning using gaussian fields and harmonic functions [C] // Proceedings of the Twentieth International Conference on Machine Learning. Washington DC: AAAI, 2003: 912-919.

[18] SINDHWANI V, NIYOGI P, BELKIN M, et al. Linear manifold regularization for large scale semi-supervised learning[C] // Proceedings of ICML Workshop on Learning with Partially Classified Training Data. Bonn, Germany: ICML, 2005: 80-83.

[19] CAI D, HE X, HAN J. Semi-supervised discriminant analysis [C] // Proceedings of IEEE International Conference on Computer Vision. Rio de Janeiro, Brazil: IEEE, 2007: 1-7.

[20] NIE F, XU D, TSANG I W H, et al. Flexible manifold embedding: A framework for semi-supervised and unsupervised dimension reduction[J]. IEEE Transactions on Image Processing, 2010, 19(7): 1921-1932.

[21] YOU M, CHEN C, BU J, et al. Emotional speech analysis on nonlinear manifold[C] // Proceedings of International Conference on Pattern Recognition. Las Vegas, Nevada, USA: dblp, 2006: 91-94.

[22] KIM J, LEE S, NARAYANAN S S. An exploratory study of manifolds of emotional speech [C] // Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing. Dallas, Texas, USA: IEEE, 2010: 5142-5145.

[23] NIE F, HUANG H, CAI X, et al. Efficient and robust feature selection via joint L_2,1-norms minimization[C] // Proceedings of Advances in Neural Information Processing Systems 23. Vancouver, British Columbia, Canada: NIPS, 2010: 1813-1821.

[24] BOYD S, VANDENBERGHE L. Convex optimization[M]. New York, USA: Cambridge University Press, 2004.

[25] BURKHARDT F, PAESCHKE A, ROLFES M, et al. A database of German emotional speech[C] // Proceedings of Interspeech. Lisbon, Portugal: ISCA, 2005: 1517-1520.

[26] MARTIN O, KOTSIA I, MACQ B, et al. The enterface'05 audio-visual emotion database [C] // Proceedings of the 22nd International Conference on Data Engineering Workshops. ATLANTA, GA, USA: IEEE, 2006: 8.

[27] Chinese LDC. CASIA-Chinese emotional speech corpus [EB/OL]. [2015-02-24]. http://www.Chineseldc.Org/en/doc/cldc-spc-2005-010/intro.htm.

[28] SCHULLER B, STEIDL S, BATLINER A, et al. The Interspeech 2010 paralinguistic challenge[C] // Proceedings of Interspeech. Makuhari, Chiba, Japan: ISCA, 2010: 2794-2797.

[29] ZHAO Jidong, LU Ke, HE Xiaofei. Locality sensitive semi-supervised feature selection[J]. Neurocomputing, 2008, 71(10): 1842-1849.