中国科学 : 信息科学　2023 年　第 53 卷　第 12 期: 2325–2340
SCIENTIA SINICA Informationis
论文

面向多维分类的监督式降维

贾彬彬[1,2], 张敏灵[1,3*]

1. 东南大学计算机科学与工程学院, 南京 210096
2. 兰州理工大学电气工程与信息工程学院, 兰州 730050
3. 计算机网络和信息集成教育部重点实验室 (东南大学), 南京 210096
* 通信作者. E-mail: zhangml@seu.edu.cn

收稿日期: 2022–09–20; 修回日期: 2023–01–04; 接受日期: 2023–04–05; 网络出版日期: 2023–12–12

国家自然科学基金 (批准号: 62225602, 62306131) 资助项目

摘要 与传统多类分类相比, 多维分类中每个对象仍由一个示例 (特征向量) 表示, 但同时与多个类别变量相关联, 各类别变量基于异构类别空间刻画对象的语义. 降维可以有效地缓解维度灾难并加速模型训练, 已有多维分类研究均关注于设计性能更好的学习算法, 尚未出现面向多维分类数据降维方面的工作. 本文基于特征空间和语义空间的相关性, 首次面向多维分类数据设计了一种名为 SDeM 的监督式线性降维方法. 该方法使用 Hilbert-Schmidt 独立判据衡量两个空间的相关性, 通过最大化投影特征空间与语义空间在该度量下的相关性确定投影矩阵. 实验结果表明, 相比于无监督式降维方法, SDeM 所得降维特征更有利于多维分类方法取得更好的泛化性能.

关键词 机器学习, 多维分类, 降维, 空间相关性, Hilbert-Schmidt 独立准则

1 引言

在传统的监督学习框架下, 一种常见的任务是在一个类别变量 (class variable) 的监督下训练分类模型, 例如多类分类 (multi-class classification, MCC). 然而, 在很多实际应用场景中, 对象的语义往往具有多维度特性, 仅用单个类别变量难以刻画对象丰富的语义信息. 例如, 电商网站需要同时向消费者提供智能手机的品牌、操作系统、处理器、屏幕材质、价格区间等多个角度的分类选项, 用以帮助消费者更方便地挑选适合自己的手机. 为了更便于对现实世界中具有多维度语义信息的对象进行建模, 一种很自然的方式是将每个对象与多个类别变量相关联, 从而对应于多维分类 (multi-dimensional classification, MDC) 框架 [1]. 相比于多类分类, 多维分类中每个对象仍然由一个示例 (特征向量) 表示, 但同时与多个类别变量相关联. 其中, 每个类别变量对应一个类别空间 (class space), 用于从不同的角度描述对象的语义信息. 实际上, 多维分类数据中的学习需求广泛存在于文本挖掘 [2,3]、计算机视觉 [4,5]、生物信息学 [6,7]、生态信息学 [8,9] 等很多实际应用场景.

引用格式: 贾彬彬, 张敏灵. 面向多维分类的监督式降维. 中国科学: 信息科学, 2023, 53: 2325–2340, doi: 10.1360/SSI-2022-0363
Jia B-B, Zhang M-L. Supervised dimensionality reduction for multi-dimensional classification (in Chinese). Sci Sin Inform, 2023, 53: 2325–2340, doi: 10.1360/SSI-2022-0363

© 2023《中国科学》杂志社

实际应用中的数据往往具有很高的维度, 例如计算机视觉中常常把图像的每个像素作为一维特征使用, 而自然语言处理中常常把字典中的每个词作为一维特征使用. 高维度特征将导致维度灾难 [10, 11] 以及繁重的计算等问题. 实际上, 虽然高维度特征可以比较充分地描述对象的性质, 但与学习任务相关的特征信息也许仅仅是高维度特征的某个低维 "嵌入" (embedding). 因此, 我们可以通过某种数学变换将原始高维度特征空间转变为一个低维 "子空间", 即降维 (dimensionality reduction) [12~14]. 由于多维分类广泛存在于文本挖掘、计算机视觉等实际应用场景, 因此多维分类的模型学习亦需要面临高维度特征带来的各种问题. 虽然诸如主成分分析 (principal component analysis, PCA)、多维缩放 (multi-dimensional scaling, MDS) 等与标记信息形式无关的降维方法可以直接应用于多维分类数据, 但无监督学习模式使它们所得降维特征并不一定与多维分类数据的监督信息匹配, 进而导致在其降维特征基础上学得多维分类模型的性能欠佳. 不同于传统多类分类问题, 多维分类输出空间包含多个类别空间, 在某个类别空间属于相同类别的样本, 在其他类别空间则可能属于不同的类别. 因此, 有必要专门研究面向多维分类的监督式降维方法, 使其学得降维特征尽可能同时匹配不同类别空间所蕴含的监督信息. 当前已有多维分类研究主要关注如何通过建模类别变量之间的依赖关系构建泛化性能更好的多维分类模型, 尚未出现面向多维分类数据降维方面的研究.

本文首次研究了面向多维分类数据的降维问题, 针对多维分类数据特点设计了一种监督式线性降维方法, 命名为 SDeM (即 supervised dimensionality reduction for MDC). 具体来说, SDeM 使用 Hilbert-Schmidt 独立判据 [15] 衡量两个空间的相关性, 基于降维所得特征空间与语义空间的相关性构建优化目标, 通过最大化目标函数求得投影矩阵, 使基于该投影矩阵所得低维特征空间与语义空间最为相关. 相比于经典的无监督线性降维方法主成分分析, SDeM 在优化目标函数中引入了监督信息. 对比实验使用了 4 种多维分类方法, 分别基于 SDeM、主成分分析、多维缩放所得降维特征在 15 个多维分类数据集上训练多维分类模型, 实验结果表明本文的监督式多维分类降维方法 SDeM 所得降维特征更有利于模型取得更好的泛化性能, 进而验证了 SDeM 的有效性.

本文剩余部分组织如下: 第 2 节简要介绍相关工作, 第 3 节详细地给出 SDeM 的技术细节, 第 4 节展示对比实验结果, 第 5 节对全文进行总结并简要讨论进一步的研究方向.

2 相关工作

与多维分类最相关的学习框架包括传统的多类分类以及多标记分类 (multi-label classification, MLC) [16, 17]. 对于多类分类, 当仅关注多维分类问题的任意单个维度时, 其对应为一个多类分类问题, 因此可将多类分类看作是当多维分类的维度个数等于 1 时的特例 (即单维分类). 对于多标记分类, 当仅关注多标记分类问题的任意单个标记时, 其对应为一个二类分类问题, 也就是说多标记分类的每个标记可看作是一个取值个数等于 2 的类别变量, 因此可将多标记分类看作是当多维分类每个维度的类别标记个数都等于 2 时的特例 [18]1). 但从概念上讲, 多维分类假设语义空间是异构的, 不同类别空间从不同的角度来描述对象的语义; 而多类分类和多标记分类则假设语义空间是同构的, 所有类别标记均来自同一个类别空间, 其中多标记分类不同于多类分类之处在于它不再限制每个示例仅可以关联类别标记集合中的单个标记, 从而更便于对多义性对象进行建模.

多维分类问题可以通过独立地为每个维度训练一个多类分类器进行求解, 学术界将此策略称为 binary relevance (BR) 方法. 但由于 BR 未考虑类别变量之间的依赖关系, 这将会影响训练所得分类器的泛化性能, 因此已有多维分类研究主要关注如何通过建模类别变量之间的依赖关系构建泛化性能

1) 多维分类、多标记分类和多类分类三者之间的关系可以参见文献 [18] 的图 1.

更好的多维分类模型. 根据依赖关系建模的方式, 可将已有工作大致分为显式建模与隐式建模两种类型. 显式建模方法通过某种结构在输出空间直接建模类别变量之间的依赖关系, 例如基于有向无环图的多维贝叶斯网分类器系列方法 [19~21]、基于链式结构的 classifier chains 系列方法 [22~24]、基于堆叠结构的两级式依赖关系建模系列方法 [25~27] 以及基于正则化约束的统计机器学习方法 [28]; 隐式建模方法则通过操作特征空间或基于变换后的标记空间训练预测模型间接建模类别变量之间的依赖关系, 例如基于特征增广的系列方法 [18,29,30] 和基于标记编码的系列方法 [31~33]. 值得一提的是, 很多监督型方法学得的模型为一个映射矩阵 (或向量), 亦可实现特征映射, 但映射特征空间的维度一般由类别个数决定, 并不能根据需要选择, 与降维任务存在本质不同.

降维是一项基本的机器学习任务 [12~14], 可以通过特征选择和特征映射两种途径实现, 本文关注后者[2)]. 根据降维过程是否需要标记信息, 降维包括无监督降维 (例如主成分分析) 和监督降维 (例如线性判别分析) 两种类型. 无监督降维方法由于其降维过程仅利用特征信息, 与标记形式无关, 对不同学习任务具有普适性; 但所得降维特征蕴含的分辨信息并不一定适宜于依据相应的标记信息训练预测模型. 监督降维方法针对学习任务的标记形式进行专门设计, 虽然普适性不如无监督降维, 但它通过协同利用特征信息和标记信息, 可以使所得降维特征与标记信息更加匹配; 其中, 监督降维方法的优势在很多学习任务的降维研究中已得到验证 [34,35], 甚至通过某些辅助方式生成一些标记信息亦有助于提高学得特征的分辨能力 [36,37]. 以多标记降维研究 [38] 为例, MDDM 方法 [39] 通过最大化特征空间和标记空间的依赖关系进行降维, MLDA 方法 [40] 则将面向多类分类问题的监督降维方法线性判别分析推广至多标记场景学得投影矩阵, 实验结果表明这些监督式的降维方法得到的低维特征空间比无监督降维方法更有利于学得泛化性能更好的多标记模型.

Hilbert-Schmidt 独立判据 [15] 用于在再生核希尔伯特空间 (reproducing kernel Hilbert space) 衡量两组随机变量的相关性. 由于它的有效性和简洁性, Hilbert-Schmidt 独立判据已被用于解决从监督学习到无监督学习、从传统机器学习到深度学习领域的各种任务 [41]. 例如, Gangeh 等 [42] 基于 Hilbert-Schmidt 独立判据设计了一种面向大规模基因表达数据的高效特征选择方法, Bao 等 [43] 基于 Hilbert-Schmidt 独立判据设计了一种面向偏标记学习数据的特征降维方法, Song 等 [44] 将聚类视为从数据推断标记的过程, 并使用 Hilbert-Schmidt 独立判据作为潜在的推断规则.

3 SDeM 方法

假设 $\mathcal{X}$ 表示由 d 维特征构成的输入空间, $\mathcal{Y}=C_1\times C_2\times\cdots\times C_q$ 表示由 q 个类别空间的笛卡尔积 (Cartesian product) 构成的输出空间, 其中 C_j 包含 K_j 个类别 ($1\leqslant j\leqslant q$), 即 $C_j=\{c_1^j,c_2^j,\ldots,c_{K_j}^j\}$. 给定多维分类数据集 $\mathcal{D}=\{(\boldsymbol{x}_i,\boldsymbol{y}_i)\mid 1\leqslant i\leqslant m\}$, 其中 $\boldsymbol{x}_i=[x_{i1},x_{i2},\ldots,x_{id}]^{\mathrm{T}}\in\mathcal{X}$ 表示 d 维特征向量 (即示例), $\boldsymbol{y}_i=[y_{i1},y_{i2},\ldots,y_{iq}]^{\mathrm{T}}\in\mathcal{Y}$ 表示与 $\boldsymbol{x}_i$ 相关联的 q 维类别向量 (即标记向量), 且有 $y_{ij}\in C_j$ (即类别向量中每项对应一个不同的类别空间), 则多维分类的学习任务是根据数据集 $\mathcal{D}$ 学得映射函数 $f:\mathcal{X}\mapsto\mathcal{Y}$, 可以预测未见示例 $\boldsymbol{x}_*$ 的类别向量 $f(\boldsymbol{x}_*)\in\mathcal{Y}$.

对于本文关注的降维任务来说, SDeM 方法尝试学得映射矩阵 $\boldsymbol{W}=[\boldsymbol{w}_1,\boldsymbol{w}_2,\ldots,\boldsymbol{w}_{d'}]\in\mathbb{R}^{d\times d'}$ 将 d 维特征向量 $\boldsymbol{x}$ 投影为 d' 维特征向量 $\boldsymbol{x}'$, 即 $\boldsymbol{x}'=\boldsymbol{W}^{\mathrm{T}}\boldsymbol{x}$, 其中 $d'\ll d$. 对于映射矩阵 $\boldsymbol{W}$, 从特征表示的层面来说, 我们希望 $\boldsymbol{w}_1,\boldsymbol{w}_2,\ldots,\boldsymbol{w}_{d'}$ 是一组标准正交基, 即 $\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}=\boldsymbol{I}_{d'}$, 其中 $\boldsymbol{I}_{d'}$ 表示 d' 阶单位矩阵. 从标记信息利用的层面来说, 我们希望基于 $\boldsymbol{W}$ 降维所得特征更有利于模型的学习. 具体来讲, SDeM 方法使用 Hilbert-Schmidt 独立判据作为衡量两个空间相关性的度量, 通过最大化训练样本

2) 文献中的 “降维” 一般指特征映射途径, 而特征选择途径通常直接称为 “特征选择”.

的降维特征空间与标记空间的相关性解得映射矩阵 $\boldsymbol{W}$.

令 $\boldsymbol{X} \in \mathbb{R}^{m\times d}$ 表示数据集 $\mathcal{D}$ 的特征矩阵, 其中 $\boldsymbol{X}$ 的第 i 行 (记为 $\boldsymbol{X}_{i:}$) 对应第 i 个示例 $\boldsymbol{x}_i$ 的转置, 即 $\boldsymbol{X}_{i:} = \boldsymbol{x}_i^{\mathrm{T}}$; 同理, 令大小为 $m \times q$ 的矩阵 $\boldsymbol{Y}$ 表示数据集 $\mathcal{D}$ 的标记矩阵, 其中 $\boldsymbol{Y}$ 的第 i 行 (记为 $\boldsymbol{Y}_{i:}$) 对应第 i 个标记向量 $\boldsymbol{y}_i$ 的转置, 即 $\boldsymbol{Y}_{i:} = \boldsymbol{y}_i^{\mathrm{T}}$. 由于多维分类假设类别变量均为名义型 (nominal) 变量, 因此离散型矩阵 $\boldsymbol{Y}$ 的这种标记表示形式并不便于数值计算. 为了有效表达标记信息, 同时亦便于后续数值计算, 本文使用名义型变量最本真的独热 (one-hot) 形式表达样本的语义信息. 具体来说, 我们将每个类别向量 $\boldsymbol{y}_i$ 表示为它的独热形式 $\boldsymbol{z}_i = [\boldsymbol{z}_i^1; \boldsymbol{z}_i^2; \cdots; \boldsymbol{z}_i^q] \in \{0,1\}^{\sum_{j=1}^{q} K_j}$; 其中, 表达式中的分号表示列向量 $\boldsymbol{z}_i$ 由 q 个列向量 $\boldsymbol{z}_i^j$ $(1 \leqslant j \leqslant q)$ 首尾拼接组成, 而长为 K_j 的列向量 $\boldsymbol{z}_i^j = [z_{i1}^j, z_{i2}^j, \ldots, z_{iK_j}^j]^{\mathrm{T}} \in \{0,1\}^{K_j}$ 对应于 $\boldsymbol{y}_i$ 的第 j 项 y_{ij} 的独热形式, 即

$$z_{ia}^j = \begin{cases} 1, \text{ if } y_{ij} = c_a^j, \\ 0, \text{ otherwise}, \end{cases} \tag{1}$$

其中 z_{ia}^j 表示向量 $\boldsymbol{z}_i^j$ 的第 a 项 $(1 \leqslant a \leqslant K_j)$. 令 $\boldsymbol{Z} \in \{0,1\}^{m\times\sum_{j=1}^{q} K_j}$ 表示数据集 $\mathcal{D}$ 的独热形式标记矩阵, 其中 $\boldsymbol{Z}$ 的第 i 行 (记为 $\boldsymbol{Z}_{i:}$) 对应第 i 个标记向量 $\boldsymbol{z}_i$ 的转置, 即 $\boldsymbol{Z}_{i:} = \boldsymbol{z}_i^{\mathrm{T}}$. 值得一提的是, 矩阵 $\boldsymbol{Z}$ 只是本文采用的标记信息表达形式, 未来还可以探索其他更有效的表达形式.

令 m 阶方阵 $\boldsymbol{K}$ 和 $\boldsymbol{L}$ 分别表示数据集 $\mathcal{D}$ 的降维所得特征向量以及标记向量的线性核矩阵, 它们的第 (i,j) 项 K_{ij} 和 L_{ij} 分别定义如下:

$$K_{ij} = \langle \boldsymbol{W}^{\mathrm{T}}\boldsymbol{x}_i, \boldsymbol{W}^{\mathrm{T}}\boldsymbol{x}_j \rangle, \ L_{ij} = \langle \boldsymbol{z}_i, \boldsymbol{z}_j \rangle, \tag{2}$$

即 $\boldsymbol{K} = \boldsymbol{X}\boldsymbol{W}(\boldsymbol{X}\boldsymbol{W})^{\mathrm{T}} = \boldsymbol{X}\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}$ 和 $\boldsymbol{L} = \boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}$, 其中 $\langle \cdot, \cdot \rangle$ 表示向量的内积.

将 $\mathcal{X}$ 包含的特征变量以及 $\mathcal{Y}$ 包含的类别变量视为两组随机变量, Hilbert-Schmidt 独立判据在 $\mathcal{X}$ 和 $\mathcal{Y}$ 的再生核希尔伯特空间计算交叉协方差算子 (cross-covariance operator) 范数的平方. 实际应用中通常基于数据集 $\mathcal{D}$ (即空间 $\mathcal{X}$ 和 $\mathcal{Y}$ 中两组随机变量的 m 个独立观测) 计算 Hilbert-Schmidt 独立判据的经验估计:

$$\mathrm{HSIC}(\mathcal{D}, \mathcal{F}, \mathcal{G}) = \frac{1}{(m-1)^2}\mathrm{tr}(\boldsymbol{H}\boldsymbol{K}\boldsymbol{H}\boldsymbol{L}), \tag{3}$$

其中, $\mathcal{F}$ 和 $\mathcal{G}$ 分别表示由 $\mathcal{X}$ 和 $\mathcal{Y}$ 映射的再生核希尔伯特空间, $\mathrm{tr}(\cdot)$ 表示矩阵的迹, $\boldsymbol{H} = \boldsymbol{I}_m - \frac{1}{m}\boldsymbol{E}_m$, $\boldsymbol{I}_m$ 和 $\boldsymbol{E}_m$ 分别为 m 阶单位矩阵和 m 阶元素全为 1 的方阵. 对于 m 阶方阵 $\boldsymbol{K}$ 来讲 (方阵 $\boldsymbol{L}$ 同理):

$$\boldsymbol{H}\boldsymbol{K} = \left(\boldsymbol{I}_m - \frac{1}{m}\boldsymbol{E}_m\right)\boldsymbol{K} = \boldsymbol{K} - \frac{1}{m}\boldsymbol{E}_m\boldsymbol{K} \triangleq \boldsymbol{K} - \boldsymbol{M}, \tag{4}$$

其中 $\boldsymbol{M} = \frac{1}{m}\boldsymbol{E}_m\boldsymbol{K}$. 根据 $\boldsymbol{E}_m$ 的定义可知, 矩阵 $\boldsymbol{M}$ 的第 (i,j) 项 $M_{ij} = \frac{1}{m}\sum_{a=1}^{m} K_{aj}$, 即 $\boldsymbol{M}$ 第 j 列所有元素均为矩阵 $\boldsymbol{K}$ 第 j 列的均值, 也就是说 $\boldsymbol{H}\boldsymbol{K}$ (或 $\boldsymbol{H}\boldsymbol{L}$) 表示对矩阵 $\boldsymbol{K}$ (或 $\boldsymbol{L}$) 按列进行中心化.

综合考虑 SDeM 方法在特征表示和标记信息利用两个层面的要求, 将式 (2) 定义的特征向量核矩阵 $\boldsymbol{K} = \boldsymbol{X}\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}$ 代入式 (3) 并忽略掉其常系数, 可得如下优化问题:

$$\max_{\boldsymbol{W}} \ \mathrm{tr}(\boldsymbol{H}\boldsymbol{X}\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{L}) \quad \text{s.t. } \boldsymbol{W}^{\mathrm{T}}\boldsymbol{W} = \boldsymbol{I}_{d'}.$$

根据矩阵迹的性质 $\mathrm{tr}(\boldsymbol{A}\boldsymbol{B}\boldsymbol{C}) = \mathrm{tr}(\boldsymbol{C}\boldsymbol{A}\boldsymbol{B})$, 上述优化问题可以等价为

$$\max_{\boldsymbol{W}} \ \mathrm{tr}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{L}\boldsymbol{H}\boldsymbol{X}\boldsymbol{W}) \quad \text{s.t. } \boldsymbol{W}^{\mathrm{T}}\boldsymbol{W} = \boldsymbol{I}_{d'}. \tag{5}$$

对该优化问题使用拉格朗日 (Lagrange) 乘子法可得

$$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{HLHX}\boldsymbol{w}_i = \lambda_i \boldsymbol{w}_i, \quad 1 \leqslant i \leqslant d'. \tag{6}$$

于是, 只需对矩阵 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{HLHX} \in \mathbb{R}^{d\times d}$ 进行特征值分解, 将求得的 d 个特征值进行排序: $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_d$, 使用前 d' 个特征值对应的特征向量构成投影矩阵 $\boldsymbol{W}$ 即可.

值得注意的是, 式 (5) 与局部保持投影 (locality preserving projections, LPP) [45] 的优化目标有些类似, 但二者亦存在很大的差异. 从本质上来讲, 局部保持投影的优化目标是通过最小化相似样本降维后的欧氏距离得到的, 而式 (5) 是通过最大化训练样本的降维特征空间与标记空间的相关性得到的. 从公式层面来讲, 若令 $\hat{\boldsymbol{L}} = \boldsymbol{HLH}$ 对应于局部保持投影优化目标中的拉普拉斯 (Laplace) 矩阵, 根据 $\boldsymbol{H}$ 的定义可知, $\hat{\boldsymbol{L}}$ 的第 (i,j) 项 $\hat{L}_{ij} = L_{ij} - \mu_i - \mu_j + \mu$, 其中 μ_i 和 μ_j 表示 $\boldsymbol{L}$ 的第 i 行和第 j 行的平均值, μ 表示 $\boldsymbol{L}$ 的整体平均值; 此时, 虽然可将 $\hat{\boldsymbol{L}}$ 进一步表示为 $\hat{\boldsymbol{D}} - \hat{\boldsymbol{S}}$ 的拉普拉斯矩阵形式, 其中 $\hat{\boldsymbol{D}}$ 为对角阵, 第 (i,i) 项 $\hat{D}_{ii}$ 等于相似度矩阵 $\hat{\boldsymbol{S}}$ 第 i 行之和, 但 $\hat{\boldsymbol{S}}$ 与标记向量的线性核矩阵 $\boldsymbol{L}$ 差异很大, 已无法简单从类似于局部保持投影的角度去解释式 (5) 的物理意义了.

另外, 本文使用的 Hilbert-Schmidt 独立判据在多视图降维方法 ISRL 中亦有应用 [46], 二者在技术上存在一定的相似性和差异性. 在相似性层面, 它们均借助于 Hilbert-Schmidt 独立判据度量两个空间的相关性, 其中 SDeM 关注投影特征空间与语义空间之间的相关性, 而 ISRL 关注不同视图之间的独立性. 在差异性层面, SDeM 通过最大化 Hilbert-Schmidt 独立判据学习面向线性降维的投影矩阵, 而 ISRL 通过最小化 Hilbert-Schmidt 独立判据学习面向衡量不同视图与各视图共有信息交互重要性的权重.

算法 1 给出了 SDeM 方法的工作流程.

算法 1 SDeM 方法

输入: 多维分类数据集 $\mathcal{D}$, 低维空间维数 d' 或域值 th;
输出: 投影矩阵 $\boldsymbol{W}$;

1: 根据式 (1) 构建标记矩阵 $\boldsymbol{Z}$;
2: 计算矩阵 $\boldsymbol{L} = \boldsymbol{ZZ}^{\mathrm{T}}$ 和 $\boldsymbol{H} = \boldsymbol{I}_m - \frac{1}{m}\boldsymbol{E}_m$;
3: 计算矩阵 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{HLHX}$;
4: 对 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{HLHX}$ 做特征值分解, 得到特征值 $\lambda_1, \lambda_2, \ldots, \lambda_d$ 及其对应的特征向量 $\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_d$, 其中 $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_d$;
5: **if** 给定 d' **then**
6: 取最大的 d' 个特征值对应的特征向量 $\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_{d'}$;
7: **else** (即给定 th)
8: 找出满足 $\sum_{i=1}^{d'} \lambda_i \geqslant \mathrm{th} \cdot \sum_{i=1}^{d} \lambda_i$ 的前 d' 个最大的特征值对应的特征向量 $\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_{d'}$;
9: **end if**
10: 返回 $\boldsymbol{W} = [\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_{d'}]$.

SDeM 基于独热编码构建标记信息矩阵 $\boldsymbol{Z}$, 致使 $\boldsymbol{Z}$ 的秩等于 $\sum_{j=1}^{q} K_j - q$, 即类别标记个数之和减去维度个数[3)], 进而其对应的核矩阵 $\boldsymbol{L} = \boldsymbol{ZZ}^{\mathrm{T}}$ 的秩亦等于 $\sum_{j=1}^{q} K_j - q$. 因此, $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{HLHX}$ 的秩不大于 $\sum_{j=1}^{q} K_j - q$. 也就是说, 降维后特征空间的维数 d' 亦不大于 $\sum_{j=1}^{q} K_j - q$.

SDeM 方法与主成分分析的关系. 作为最经典的无监督降维方法, 主成分分析旨在构建 d' 维的低维特征空间, 希望所有样本点在低维空间的投影尽可能分开, 亦即使得投影后样本点的方差最大化.

3) 严格上说, $\boldsymbol{Z}$ 的秩等于 m 和 $\sum_{j=1}^{q} K_j - q$ 之中的较小者, 但一般情况下有 $\sum_{j=1}^{q} K_j - q \ll m$.

表 1　实验数据集的特点

Table 1　Characteristics of the experimental data sets

Data set	#Example (m)	#Dim. (q)	#Label/Dim. ($K_1, \ldots, K_q$)	#Feature (d)	Ref.
Oes10	403	16	3	298	[47]
Song	785	3	3	98	[18]
BeLaE	1930	5	5	45	[48]
Voice	3136	2	4, 2	19	[49]
Scm20d	8966	16	4	61	[47]
Rf1	8987	8	4, 4, 3, 4, 4, 3, 4, 3	64	[47]
Thyroid	9172	7	5, 5, 3, 2, 4, 4, 3	34	[50]
Scm1d	9803	16	4	280	[47]
CoIL	9822	5	6, 10, 10, 4, 2	620	[50]
TIC	9822	3	6, 4, 2	640	[50]
Flickr	12198	5	3, 4, 3, 4, 4	1536	[51]
Disfa	13095	12	5, 5, 6, 3, 4, 4, 5, 4, 4, 4, 6, 4	136	[52]
Fera	14052	5	6	136	[5]
Adult	18419	4	7, 7, 5, 2	82	[50]
Default	28779	4	2, 7, 4, 2	39	[53]

依据本文所定义的符号, 主成分分析的优化目标通常写为

$$\max_{\boldsymbol{W}} \ \mathrm{tr}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{W}) \quad \text{s.t. } \boldsymbol{W}^{\mathrm{T}}\boldsymbol{W} = \boldsymbol{I}_{d'}.$$

但以上优化目标需要所有样本必须是中心化的, 即 $\boldsymbol{x}_i \leftarrow \boldsymbol{x}_i - \frac{1}{m}\sum_{j=1}^{m}\boldsymbol{x}_j$, 写成矩阵形式即 $\boldsymbol{X} \leftarrow \boldsymbol{H}\boldsymbol{X}$. 因此, 主成分分析的优化目标可以直接统一表达为

$$\max_{\boldsymbol{W}} \ \mathrm{tr}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{H}\boldsymbol{X}\boldsymbol{W}) \quad \text{s.t. } \boldsymbol{W}^{\mathrm{T}}\boldsymbol{W} = \boldsymbol{I}_{d'}. \tag{7}$$

注意, 上式目标函数原本为 $\mathrm{tr}(\boldsymbol{W}^{\mathrm{T}}(\boldsymbol{H}\boldsymbol{X})^{\mathrm{T}}\boldsymbol{H}\boldsymbol{X}\boldsymbol{W})$, 由于 $\boldsymbol{H}$ 为对称阵, 故有 $(\boldsymbol{H}\boldsymbol{X})^{\mathrm{T}}\boldsymbol{H}\boldsymbol{X} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{H}\boldsymbol{X}$.

对比式 (5) 和 (7) 可以发现[4], 本文 SDeM 方法和主成分分析的区别在于目标函数中是否存在蕴含监督信息的标记核矩阵 $\boldsymbol{L}$. 通过接下来的对比实验可以看到, SDeM 方法通过引入标记信息可以提高所得降维特征的分辨能力, 有助于后续训练的预测模型取得更好的泛化性能.

4　实验

4.1　实验设置

4.1.1　数据集

本文的对比实验共使用了 15 个多维分类数据集, 表 1 总结了这些数据集的详细特点, 包括样本个数 (#Example)、维度个数 (#Dim.)、每个维度包含的类别标记个数 (#Label/Dim.) 以及特征维度 (#Feature); 为便于理解, 括号中还给出了其对应的数学符号. 其中, 对于 "#Label/Dim.", 若数据集所有维度包含的类别标记个数相同, 则仅给出该值; 否则, 每个维度包含的标记个数依序给出.

4) 根据 $\boldsymbol{H}$ 的定义易知 $\boldsymbol{H}\boldsymbol{H} = \boldsymbol{H}$, 即式 (7) 的目标函数可进一步化简为 $\mathrm{tr}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{X}\boldsymbol{W})$, 文中保留式 (7) 的形式主要是为了方便与式 (5) 进行对比, 用以直观说明本文 SDeM 方法和主成分分析的区别.

4.1.2 评价指标

令 $\mathcal{S}=\{(\boldsymbol{x}_i,\boldsymbol{y}_i)\mid 1\leqslant i\leqslant p\}$ 表示多维分类测试集, $f:\mathcal{X}\mapsto\mathcal{Y}$ 表示待评估的多维分类模型; 对于测试样本 $\boldsymbol{x}_i$ 来说, 令 $\boldsymbol{y}_i=[y_{i1},y_{i2},\ldots,y_{iq}]^{\mathrm{T}}\in\mathcal{Y}$ 表示其真实的类别向量, $\hat{\boldsymbol{y}}_i=f(\boldsymbol{x}_i)=[\hat{y}_{i1},\hat{y}_{i2},\ldots,\hat{y}_{iq}]^{\mathrm{T}}$ 表示由 f 预测的类别向量; 则样本 $\boldsymbol{x}_i$ 被预测正确的类别标记个数可定义为 $r^{(i)}=\sum_{j=1}^{q}\mathbf{1}_{y_{ij}=\hat{y}_{ij}}$, 其中对于谓词 $\mathbf{1}_{\pi}$, 当 π 成立时 $\mathbf{1}_{\pi}$ 返回值为 1, 否则返回值为 0. 基于这些符号, 本文用于评价多维分类模型泛化性能的 3 种评价指标的定义如下 [1,18,20,31].

- 汉明分值 (Hamming score, HS):

$$\mathrm{HS}_{\mathcal{S}}(f)=\frac{1}{p}\sum_{i=1}^{p}\frac{1}{q}\cdot r^{(i)}. \tag{8}$$

该指标测量样本的类别空间被预测正确个数的平均值.

- 精确匹配 (exact match, EM):

$$\mathrm{EM}_{\mathcal{S}}(f)=\frac{1}{p}\sum_{i=1}^{p}\mathbf{1}_{r^{(i)}=q}. \tag{9}$$

该指标测量所有类别空间均被预测正确的样本所占的比例.

- 亚精确匹配 (sub-exact match, SEM):

$$\mathrm{SEM}_{\mathcal{S}}(f)=\frac{1}{p}\sum_{i=1}^{p}\mathbf{1}_{r^{(i)}\geqslant q-1}. \tag{10}$$

该指标测量至少有 $q-1$ 个类别空间被预测正确的样本所占的比例.

容易看出, 以上 3 个评价指标的指标数值越大代表模型泛化性能越好. 在本文的对比实验中, 采用十折交叉验证 (ten-fold cross validation) 的方法来对模型进行评估, 并计算出各评价指标的均值和方差用来比较各对比方法的优劣.

4.1.3 对比方法

作为面向多维分类数据降维研究的初步尝试, 目前尚不存在其他监督式多维分类降维方法用于对比. 因此, 实验部分主要通过与经典的无监督降维方法进行对比, 来查验专门设计监督式降维方法是否可以提高降维特征的分辨能力. 具体来说, 本文将 SDeM 与主成分分析、多维缩放两个经典的无监督降维方法进行对比. 通过第 3 节的分析可以知道, 本文 SDeM 方法和主成分分析的区别在于目标函数中是否存在蕴含监督信息的标记核矩阵 $\boldsymbol{L}$; 多维缩放则要求样本在降维所得低维空间中的距离与在原始特征空间中的距离尽可能接近.

由于 SDeM 方法、主成分分析、多维缩放的降维过程均独立于多维分类模型训练, 因此可以使用任意多维分类方法基于它们的降维特征进行学习, 通过对比同一种多维分类方法分别基于它们降维特征学得预测模型的泛化性能来判断 SDeM 方法和两种对比降维方法结果的优劣. 具体来说, 本文共使用如下 4 种多维分类方法:

• BR 方法将多维分类问题按维度分解为 q 个独立的多类分类问题, 针对每个维度分别训练一个多类分类器. BR 方法完全忽略了维度之间的依赖关系, 是多维分类研究中最基本的基准方法.

• gMML 方法 [31] 首先基于 “一对其余” (one-vs-rest) 方式对每个类别空间进行变换, 然后在变换后的输出空间结合度量学习技术训练预测模型.

• SEEM 方法 [26] 首先针对每对类别空间进行第一级训练建模二阶依赖关系, 然后基于第一级的预测结果针对每个类别空间完成第二级训练从而建模高阶依赖关系.

• EDCC 方法 [24] 首先基于 "一对一" (one-vs-one) 方式将多维分类问题逐维分解为多个二类分类问题, 然后基于链式结构训练一串二类分类器完成多维分类模型的学习.

以上 4 种方法除 gMML 之外均需要一个多类分类器, 本文使用 LIBLINEAR [54] 实现的支持向量机担任[5]. 对于 gMML 方法, 其参数配置为 $\lambda = 10$, $t = 0.7$, $\gamma = 0.1$ 和 $k = 20$; 对于 SEEM 方法, 唯一的参数 k 设置为 10; 对于 EDCC 方法, 集成的基分类器个数设置为 10.

对于每种多维分类方法 $\mathcal{A} \in \{$BR, gMML, SEEM, EDCC$\}$, 我们使用 $\mathcal{A}$-SDeM, $\mathcal{A}$-PCA 和 $\mathcal{A}$-MDS 分别表示使用 SDeM 方法、主成分分析和多维缩放降维所得特征训练的多维分类模型, 本文通过对比 $\mathcal{A}$-SDeM, $\mathcal{A}$-PCA 和 $\mathcal{A}$-MDS 关于每个评价指标的泛化性能来评价 SDeM 方法的有效性. 为了公平对比, SDeM 方法、主成分分析和多维缩放保留相同的降维特征维度, 设置为 $d' = \sum_{j=1}^{q} K_j - q$.

4.2 实验结果

表 2∼4 分别给出了关于汉明分值、精确匹配和亚精确匹配的实验结果. 其中, 若 $\mathcal{A}$-SDeM 的实验结果在某数据集上优于 $\mathcal{A}$-PCA 和 $\mathcal{A}$-MDS 的实验结果, 则将 $\mathcal{A}$-PCA 和 $\mathcal{A}$-MDS 相应的实验结果增加下划线进行强调显示. 为了分析 $\mathcal{A}$-SDeM 的性能在 15 个数据集上是否显著优于 $\mathcal{A}$-PCA 和 $\mathcal{A}$-MDS, 我们使用了 Wilcoxon 符号秩检验 (Wilcoxon signed-ranks test) [55] 作为统计检验工具. 表 5 总结了统计检验结果, 其中 win/tie/loss 分别表示 $\mathcal{A}$-SDeM 与对比方法相比, 性能更好/相同/更差, 方括号中还给出了检验相应的 p 值, 其中显著度 α 设置为 0.05.

根据报道的实验结果, 我们可以得到如下观察:

• 如表 2∼4 所示, 对于 $\mathcal{A}$-SDeM 和 $\mathcal{A}$-PCA, 在共 180 种配置 (4 种方法 × 15 个数据集 × 3 个评价指标) 中, $\mathcal{A}$-SDeM 共在 138 种配置上优于 $\mathcal{A}$-PCA、在 20 种配置上与 $\mathcal{A}$-PCA 相同、在 22 种配置上次于 $\mathcal{A}$-PCA; 对于 $\mathcal{A}$-SDeM 和 $\mathcal{A}$-MDS, $\mathcal{A}$-SDeM 共在 137 种配置上优于 $\mathcal{A}$-MDS、在 12 种配置上与 $\mathcal{A}$-MDS 相同、在 31 种配置上次于 $\mathcal{A}$-MDS.

• 如表 5 所示, 关于所有 3 种评价指标及对于所有 4 种多维分类方法, $\mathcal{A}$-SDeM 可以取得统计上优于 $\mathcal{A}$-PCA 的泛化性能; 结合 SDeM 方法与主成分分析的密切联系, 该实验结果清晰地说明了降维过程引入监督信息确实有利于提升降维所得特征的分辨能力.

• 相比于 $\mathcal{A}$-MDS, 统计检验结果显示 $\mathcal{A}$-SDeM 可以取得优于或至少与 $\mathcal{A}$-MDS 持平的实验结果, 该结果亦有效说明了监督式降维相比于无监督降维的优势. 值得一提的是, 多维缩放在降维时需要同时使用训练样本和测试样本, 而 SDeM 和主成分分析仅基于训练样本进行降维, 然后再使用学得的映射矩阵对测试样本完成降维.

• 尽管基于相同的降维特征训练预测模型, 但对于不同的多维分类方法来说, $\mathcal{A}$-SDeM, $\mathcal{A}$-PCA 和 $\mathcal{A}$-MDS 之间的优劣关系并不相同. 例如, BR-SDeM 仅在 3 种配置上次于 BR-PCA, 但 EDCC-SDeM 在 8 种配置上次于 EDCC-PCA, 可能的原因在于 BR 方法训练预测模型时使用的基分类器采用 "一对其余" 方式解决多类分类问题, 这与 SDeM 方法降维时使用独热变换对应的标记空间类似; EDCC 方法训练预测模型时首先使用了 "一对一" 方式分解, 与 SDeM 方法降维时使用独热变换对应的标记空间差异较大. 因此, 未来可以开展针对某种特定多维分类方法的降维研究.

• 类似地, 还可以观察到 $\mathcal{A}$-SDeM, $\mathcal{A}$-PCA 和 $\mathcal{A}$-MDS 之间的优劣关系亦与数据集有关. 因此, 未来还可以开展针对特定多维分类应用 (数据集) 的降维研究.

5) LIBLINEAR 配置为 "L2-regularized L1-loss support vector classification (dual)", 即 "-s 3".

表 2 关于汉明分值的实验结果 (均值 ± 方差)

Table 2 Experimental results (mean±std.) in terms of Hamming score

Data set (d'/d)	BR-SDeM	BR-PCA	BR-MDS	gMML-SDeM	gMML-PCA	gMML-MDS
Oes10 (32/298)	0.754±0.018	0.748±0.024	0.795±0.012	0.775±0.020	0.774±0.019	0.772±0.020
Song (6/98)	0.777±0.028	0.773±0.024	0.562±0.049	0.778±0.024	0.776±0.024	0.776±0.024
BeLaE (20/45)	0.402±0.019	0.401±0.017	0.331±0.011	0.414±0.020	0.412±0.017	0.416±0.017
Voice (4/19)	0.794±0.014	0.753±0.013	0.692±0.013	0.810±0.009	0.779±0.014	0.783±0.010
Scm20d (48/61)	0.624±0.007	0.623±0.008	0.552±0.007	0.600±0.007	0.600±0.007	0.600±0.007
Rf1 (21/64)	0.796±0.005	0.794±0.006	0.697±0.007	0.730±0.006	0.730±0.007	0.733±0.007
Thyroid (19/34)	0.962±0.002	0.959±0.002	0.714±0.047	0.960±0.002	0.960±0.003	0.960±0.003
Scm1d (48/280)	0.698±0.004	0.694±0.005	0.624±0.005	0.696±0.006	0.695±0.007	0.699±0.008
CoIL (27/620)	0.828±0.006	0.701±0.007	0.561±0.005	0.824±0.005	0.725±0.006	0.727±0.006
TIC (9/640)	0.839±0.006	0.696±0.007	0.570±0.009	0.849±0.006	0.789±0.007	0.789±0.007
Flickr (13/1536)	0.747±0.005	0.725±0.004	0.645±0.004	0.743±0.004	0.714±0.005	0.715±0.006
Disfa (42/136)	0.894±0.002	0.894±0.002	0.553±0.004	0.884±0.003	0.884±0.003	0.884±0.003
Fera (25/136)	0.608±0.013	0.606±0.009	0.503±0.007	0.589±0.007	0.589±0.007	0.589±0.007
Adult (17/82)	0.689±0.003	0.662±0.004	0.578±0.006	0.692±0.004	0.672±0.004	0.672±0.004
Default (11/39)	0.644±0.004	0.592±0.005	0.492±0.019	0.662±0.004	0.619±0.004	0.619±0.004
Data set (d'/d)	SEEM-SDeM	SEEM-PCA	SEEM-MDS	EDCC-SDeM	EDCC-PCA	EDCC-MDS
Oes10 (32/298)	0.735±0.037	0.715±0.043	0.729±0.028	0.800±0.017	0.798±0.014	0.802±0.012
Song (6/98)	0.776±0.027	0.775±0.028	0.769±0.030	0.778±0.022	0.777±0.025	0.778±0.022
BeLaE (20/45)	0.397±0.018	0.390±0.015	0.391±0.011	0.448±0.018	0.440±0.018	0.445±0.019
Voice (4/19)	0.803±0.011	0.782±0.013	0.762±0.013	0.840±0.015	0.793±0.010	0.842±0.011
Scm20d (48/61)	0.691±0.017	0.699±0.017	0.699±0.013	0.688±0.006	0.688±0.006	0.687±0.005
Rf1 (21/64)	0.942±0.006	0.939±0.014	0.920±0.007	0.911±0.004	0.912±0.003	0.910±0.004
Thyroid (19/34)	0.935±0.059	0.931±0.059	0.900±0.056	0.964±0.002	0.960±0.003	0.960±0.003
Scm1d (48/280)	0.756±0.010	0.751±0.025	0.734±0.019	0.829±0.002	0.827±0.003	0.835±0.003
CoIL (27/620)	0.848±0.024	0.788±0.006	0.768±0.006	0.870±0.006	0.806±0.005	0.807±0.007
TIC (9/640)	0.854±0.006	0.804±0.006	0.774±0.010	0.857±0.007	0.811±0.007	0.813±0.008
Flickr (13/1536)	0.757±0.005	0.734±0.006	0.708±0.005	0.759±0.005	0.729±0.005	0.731±0.005
Disfa (42/136)	0.861±0.043	0.875±0.041	0.899±0.016	0.904±0.002	0.904±0.002	0.904±0.002
Fera (25/136)	0.682±0.007	0.679±0.008	0.641±0.007	0.637±0.008	0.635±0.007	0.638±0.008
Adult (17/82)	0.696±0.005	0.676±0.005	0.660±0.008	0.707±0.004	0.689±0.006	0.689±0.005
Default (11/39)	0.652±0.003	0.628±0.006	0.600±0.006	0.666±0.003	0.620±0.004	0.620±0.005

4.3 敏感性分析

图 1 给出了 $\mathcal{A}$-SDeM (其中 $\mathcal{A} \in \{$BR, gMML, SEEM, EDCC$\}$) 在数据集 Rf1, CoIL, Adult, Default 上关于 3 个评价指标的泛化性能随降维特征维度 d' 增加时的变化曲线. 图 1 中, 横轴对应降维特征维度 d', 其变化范围为 $\{0.1, 0.2, \ldots, 1\} \times (\sum_{j=1}^{q} K_j - q)$ (四舍五入), 纵轴为相应评价指标的值. 从图 1 中可以看出, 除少数例外情况之外, 整体上来讲 $\mathcal{A}$-SDeM 的泛化性能随着 d' 的增加会变好; 但同时也可以看到, 当 $d' \geqslant 0.5 \times (\sum_{j=1}^{q} K_j - q)$ 时, 各多维分类方法关于各指标的性能提升随着 d' 的增加将变

表 3　关于精确匹配的实验结果 (均值 ± 方差)

Table 3　Experimental results (mean±std.) in terms of exact match

Data set (d'/d)	BR-SDeM	BR-PCA	BR-MDS	gMML-SDeM	gMML-PCA	gMML-MDS
Oes10 (32/298)	0.062±0.034	0.062±0.024	0.097±0.038	0.074±0.033	0.084±0.048	0.079±0.045
Song (6/98)	0.461±0.070	0.458±0.053	0.114±0.044	0.465±0.059	0.460±0.048	0.462±0.050
BeLaE (20/45)	0.013±0.011	0.017±0.007	0.004±0.003	0.018±0.010	0.025±0.009	0.023±0.010
Voice (4/19)	0.617±0.028	0.568±0.023	0.450±0.023	0.663±0.013	0.615±0.020	0.629±0.016
Scm20d (48/61)	0.042±0.004	0.044±0.006	0.029±0.005	0.052±0.007	0.051±0.007	0.055±0.009
Rf1 (21/64)	0.218±0.013	0.214±0.013	0.106±0.007	0.137±0.011	0.137±0.011	0.141±0.010
Thyroid (19/34)	0.756±0.013	0.736±0.011	0.137±0.097	0.741±0.014	0.738±0.016	0.738±0.016
Scm1d (48/280)	0.080±0.007	0.080±0.007	0.066±0.007	0.099±0.009	0.099±0.008	0.106±0.009
CoIL (27/620)	0.394±0.014	0.187±0.016	0.060±0.007	0.376±0.010	0.192±0.012	0.193±0.011
TIC (9/640)	0.579±0.016	0.354±0.017	0.149±0.008	0.589±0.012	0.461±0.016	0.462±0.014
Flickr (13/1536)	0.233±0.010	0.195±0.010	0.131±0.008	0.222±0.007	0.183±0.011	0.184±0.012
Disfa (42/136)	0.393±0.011	0.393±0.010	0.021±0.004	0.379±0.011	0.379±0.011	0.379±0.011
Fera (25/136)	0.195±0.014	0.195±0.011	0.058±0.008	0.195±0.013	0.195±0.013	0.195±0.013
Adult (17/82)	0.201±0.008	0.156±0.006	0.131±0.010	0.207±0.007	0.175±0.009	0.176±0.009
Default (11/39)	0.157±0.007	0.111±0.007	0.051±0.011	0.172±0.009	0.125±0.008	0.125±0.008
Data set (d'/d)	SEEM-SDeM	SEEM-PCA	SEEM-MDS	EDCC-SDeM	EDCC-PCA	EDCC-MDS
Oes10 (32/298)	0.037±0.036	0.035±0.031	0.022±0.030	0.097±0.045	0.102±0.044	0.092±0.047
Song (6/98)	0.457±0.072	0.458±0.071	0.446±0.059	0.460±0.060	0.466±0.056	0.465±0.052
BeLaE (20/45)	0.018±0.010	0.016±0.011	0.013±0.007	0.030±0.015	0.031±0.012	0.029±0.012
Voice (4/19)	0.647±0.021	0.600±0.018	0.611±0.022	0.695±0.026	0.619±0.019	0.709±0.023
Scm20d (48/61)	0.030±0.022	0.024±0.017	0.031±0.013	0.095±0.008	0.095±0.009	0.093±0.010
Rf1 (21/64)	0.631±0.036	0.610±0.084	0.565±0.016	0.495±0.016	0.499±0.015	0.497±0.014
Thyroid (19/34)	0.608±0.319	0.589±0.311	0.449±0.222	0.769±0.013	0.738±0.017	0.738±0.017
Scm1d (48/280)	0.053±0.031	0.052±0.040	0.049±0.031	0.195±0.010	0.194±0.012	0.196±0.010
CoIL (27/620)	0.464±0.078	0.321±0.015	0.288±0.010	0.513±0.017	0.354±0.006	0.359±0.012
TIC (9/640)	0.606±0.018	0.497±0.014	0.455±0.009	0.612±0.015	0.494±0.016	0.500±0.018
Flickr (13/1536)	0.255±0.015	0.213±0.008	0.181±0.010	0.253±0.010	0.207±0.009	0.209±0.011
Disfa (42/136)	0.154±0.203	0.237±0.208	0.367±0.099	0.414±0.010	0.410±0.009	0.411±0.009
Fera (25/136)	0.260±0.011	0.257±0.011	0.213±0.012	0.211±0.013	0.210±0.013	0.210±0.015
Adult (17/82)	0.250±0.009	0.207±0.011	0.187±0.015	0.242±0.009	0.219±0.010	0.220±0.011
Default (11/39)	0.169±0.006	0.143±0.007	0.110±0.006	0.177±0.006	0.131±0.010	0.129±0.010

得十分平缓. 因此, 在实际应用中需要保留多少维特征要根据实际情况具体分析决定. 在 4.2 小节的实验中, 降维特征维度都设置为 $d' = \sum_{j=1}^{q} K_j - q$.

另外, 图 1 还以相应的虚线给出了方法 $\mathcal{A}$ 的性能 (即使用原始特征空间基于方法 $\mathcal{A}$ 训练模型), 以方便比较降维对泛化性能带来的影响. 从图 1 中可以看出, 降维带来的影响与数据集及多维分类方法均有关, 因此实际应用中应该针对具体问题进行具体分析.

表 4 关于亚精确匹配的实验结果 (均值 ± 方差)

Table 4 Experimental results (mean±std.) in terms of sub-exact match

Data set (d'/d)	BR-SDeM	BR-PCA	BR-MDS	gMML-SDeM	gMML-PCA	gMML-MDS
Oes10 (32/298)	0.141±0.057	0.139±0.057	0.191±0.056	0.179±0.050	0.171±0.041	0.171±0.049
Song (6/98)	0.875±0.036	0.865±0.036	0.652±0.086	0.871±0.030	0.870±0.036	0.868±0.037
BeLaE (20/45)	0.118±0.025	0.112±0.012	0.050±0.014	0.128±0.024	0.127±0.024	0.125±0.025
Voice (4/19)	0.971±0.008	0.937±0.009	0.933±0.016	0.958±0.010	0.942±0.012	0.936±0.009
Scm20d (48/61)	0.090±0.006	0.092±0.008	0.058±0.004	0.100±0.009	0.100±0.009	0.103±0.008
Rf1 (21/64)	0.525±0.013	0.522±0.016	0.313±0.014	0.374±0.014	0.375±0.015	0.372±0.014
Thyroid (19/34)	0.981±0.005	0.980±0.004	0.491±0.108	0.982±0.005	0.982±0.005	0.982±0.005
Scm1d (48/280)	0.168±0.012	0.163±0.011	0.122±0.008	0.195±0.011	0.196±0.013	0.203±0.014
CoIL (27/620)	0.793±0.015	0.545±0.013	0.272±0.012	0.786±0.014	0.583±0.015	0.587±0.015
TIC (9/640)	0.941±0.006	0.778±0.011	0.610±0.018	0.958±0.007	0.911±0.009	0.910±0.009
Flickr (13/1536)	0.621±0.014	0.565±0.016	0.426±0.013	0.612±0.015	0.542±0.017	0.543±0.019
Disfa (42/136)	0.627±0.013	0.623±0.011	0.071±0.007	0.590±0.009	0.590±0.009	0.590±0.009
Fera (25/136)	0.398±0.020	0.392±0.014	0.241±0.012	0.378±0.012	0.377±0.012	0.378±0.012
Adult (17/82)	0.640±0.008	0.600±0.010	0.446±0.016	0.649±0.008	0.615±0.009	0.615±0.009
Default (11/39)	0.556±0.007	0.474±0.009	0.285±0.034	0.587±0.007	0.512±0.007	0.512±0.008
Data set (d'/d)	SEEM-SDeM	SEEM-PCA	SEEM-MDS	EDCC-SDeM	EDCC-PCA	EDCC-MDS
Oes10 (32/298)	0.107±0.069	0.094±0.044	0.084±0.047	0.208±0.070	0.196±0.064	0.198±0.061
Song (6/98)	0.874±0.029	0.869±0.036	0.866±0.045	0.876±0.025	0.869±0.039	0.873±0.035
BeLaE (20/45)	0.105±0.014	0.095±0.021	0.096±0.017	0.147±0.025	0.148±0.028	0.153±0.029
Voice (4/19)	0.959±0.007	0.963±0.012	0.913±0.015	0.986±0.006	0.966±0.005	0.974±0.009
Scm20d (48/61)	0.095±0.031	0.097±0.037	0.101±0.016	0.171±0.008	0.169±0.009	0.170±0.006
Rf1 (21/64)	0.918±0.014	0.917±0.026	0.856±0.023	0.828±0.011	0.831±0.012	0.828±0.013
Thyroid (19/34)	0.938±0.088	0.933±0.097	0.862±0.148	0.981±0.004	0.982±0.005	0.982±0.005
Scm1d (48/280)	0.171±0.040	0.162±0.059	0.142±0.052	0.371±0.017	0.363±0.013	0.375±0.009
CoIL (27/620)	0.817±0.036	0.710±0.012	0.667±0.018	0.862±0.017	0.746±0.015	0.747±0.021
TIC (9/640)	0.958±0.008	0.919±0.011	0.884±0.019	0.960±0.008	0.941±0.008	0.942±0.007
Flickr (13/1536)	0.641±0.013	0.590±0.018	0.537±0.014	0.645±0.015	0.574±0.016	0.578±0.015
Disfa (42/136)	0.523±0.187	0.564±0.189	0.657±0.051	0.666±0.007	0.662±0.009	0.663±0.010
Fera (25/136)	0.509±0.013	0.504±0.014	0.449±0.012	0.437±0.009	0.435±0.009	0.439±0.009
Adult (17/82)	0.641±0.008	0.616±0.011	0.588±0.018	0.664±0.012	0.627±0.013	0.626±0.010
Default (11/39)	0.567±0.007	0.523±0.011	0.480±0.008	0.593±0.008	0.510±0.007	0.510±0.008

4.4 监督降维对比实验

监督式降维方法的关键在于如何充分利用标记信息来对数据进行降维. 作为面向多维分类数据降维研究的初步尝试, 为了表明 SDeM 能够更好地利用标记信息, 本小节进一步将 SDeM 与经典的监督降维方法线性判别分析 (linear discriminant analysis, LDA) 进行了比较. 具体来说, 我们使用线性判别分析独立地根据每个维度的标记信息进行降维并训练分类器, 即使用 BR 方法进行学习.

表 6 给出了 BR-SDeM 与 BR-LDA 的对比实验结果, 其中 BR-SDeM 与 BR-LDA 二者之中的较

表 5　$\mathcal{A}$-SDeM 与 $\mathcal{A}$-PCA 和 $\mathcal{A}$-MDS 之间的 Wilcoxon 符号秩检验 (显著度 α 为 0.05)

Table 5　Wilcoxon signed-ranks test for $\mathcal{A}$-SDeM against $\mathcal{A}$-PCA and $\mathcal{A}$-MDS (significance level $\alpha = 0.05$)

Evaluation	BR-SDeM vs.		gMML-SDeM vs.		SEEM-SDeM vs.		EDCC-SDeM vs.	
metrics	BR-PCA	BR-MDS	gMML-PCA	gMML-MDS	SEEM-PCA	SEEM-MDS	EDCC-PCA	EDCC-MDS
HS	win[6.10E−05]	win[1.22E−04]	win[6.10E−04]	tie[7.30E−02]	win[8.36E−03]	win[5.37E−03]	win[4.27E−04]	tie[8.33E−02]
EM	win[2.03E−02]	win[4.27E−04]	win[4.79E−02]	tie[2.29E−01]	win[8.36E−03]	win[1.03E−02]	win[4.13E−02]	win[4.13E−02]
SEM	win[1.83E−04]	win[4.27E−04]	win[1.25E−02]	win[1.51E−02]	win[1.03E−02]	win[8.36E−03]	win[1.53E−03]	win[3.02E−02]

表 6　BR-SDeM 与 BR-LDA 的对比实验结果 (均值 ± 方差)

Table 6　Comparative experimental results (mean±std.) of BR-SDeM and BR-LDA

	Hamming score		Exact match		Sub-exact match	
Data set (d'/d)	BR-SDeM	BR-LDA	BR-SDeM	BR-LDA	BR-SDeM	BR-LDA
Oes10 (32/298)	**0.754±0.018**	0.458±0.036	**0.062±0.034**	0.000±0.000	**0.141±0.057**	.000±.000
Song (6/98)	0.777±.028	0.777±.023	**0.461±0.070**	0.458±0.059	**0.875±0.036**	0.874±0.023
BeLaE (20/45)	**0.402±0.019**	0.377±0.017	0.013±0.011	**0.014±0.009**	**0.118±0.025**	0.086±0.017
Voice (4/19)	**0.794±0.014**	0.518±0.016	**0.617±0.028**	0.282±0.031	**0.971±0.008**	0.753±0.041
Scm20d (48/61)	**0.624±0.007**	0.495±0.006	**0.042±0.004**	0.021±0.006	**0.090±0.006**	0.048±0.006
Rf1 (21/64)	**0.796±0.005**	0.661±0.010	**0.218±0.013**	0.072±0.014	**0.525±0.013**	0.246±0.019
Thyroid (19/34)	**0.962±0.002**	0.950±0.013	**0.756±0.013**	0.700±0.059	**0.981±0.005**	0.954±0.031
Scm1d (48/280)	**0.698±0.004**	0.498±0.006	**0.080±0.007**	0.002±0.001	**0.168±0.012**	0.010±0.003
CoIL2000 (27/620)	**0.828±0.006**	0.372±0.091	**0.394±0.014**	0.002±0.002	**0.793±0.015**	0.057±0.062
TIC2000 (9/640)	**0.839±0.006**	0.357±0.051	**0.579±0.016**	0.011±0.004	**0.941±0.006**	0.278±0.102
Flickr (13/1536)	**0.747±0.005**	0.662±0.007	**0.233±0.010**	0.144±0.008	**0.621±0.014**	0.462±0.015
Disfa (42/136)	**0.894±0.002**	0.891±0.002	**0.393±0.011**	0.389±0.012	**0.627±0.013**	0.614±0.009
Fera (25/136)	**0.608±0.013**	0.455±0.038	**0.195±0.014**	0.007±0.008	**0.398±0.020**	0.160±0.096
Adult (17/82)	**0.689±0.003**	0.668±0.022	0.201±0.008	0.201±0.024	**0.640±0.008**	0.601±0.044
Default (11/39)	**0.644±0.004**	0.643±0.016	**0.157±0.007**	0.156±0.016	**0.556±0.007**	0.552±0.030

好者进行了加粗显示. 如表 6 所示, 在共 45 种配置 (15 个数据集 × 3 个评价指标) 中, BR-SDeM 共在 42 种配置上优于 BR-LDA、在 2 种配置上与 BR-LDA 相同、在 1 种配置上次于 BR-LDA, 并且在每个数据集上, BR-SDeM 至少关于两个评价指标的性能优于 BR-LDA. 该实验结果表明, 通过研究专门面向多维分类数据的降维方法有助于更好地利用多维分类的标记信息.

5　结论

本文的主要贡献包含 3 层: (1) 首次研究了面向多维分类数据的降维问题; (2) 提出了一种监督式的多维分类降维方法 SDeM; (3) 实验表明, 相比于无监督降维方法, 在降维过程中引入了监督信息的 SDeM 可以有效提升降维特征的分辨能力, 验证了 SDeM 方法的有效性.

本文仅对多维分类的降维问题做了初步探索, 未来还可以在该问题上做更为深入的研究. 例如, SDeM 方法采用了独热形式的标记信息表达形式以及基于 Hilbert-Schmidt 独立判据的相关性评价准则, 致使降维特征的最大维度受限于标记矩阵 $\boldsymbol{Z}$ 的秩, 因此未来可以在 SDeM 方法基础上探索其他形式的标记信息表达以及评价准则, 在提升降维特征分辨能力的同时扩大降维特征最大维度的可选范

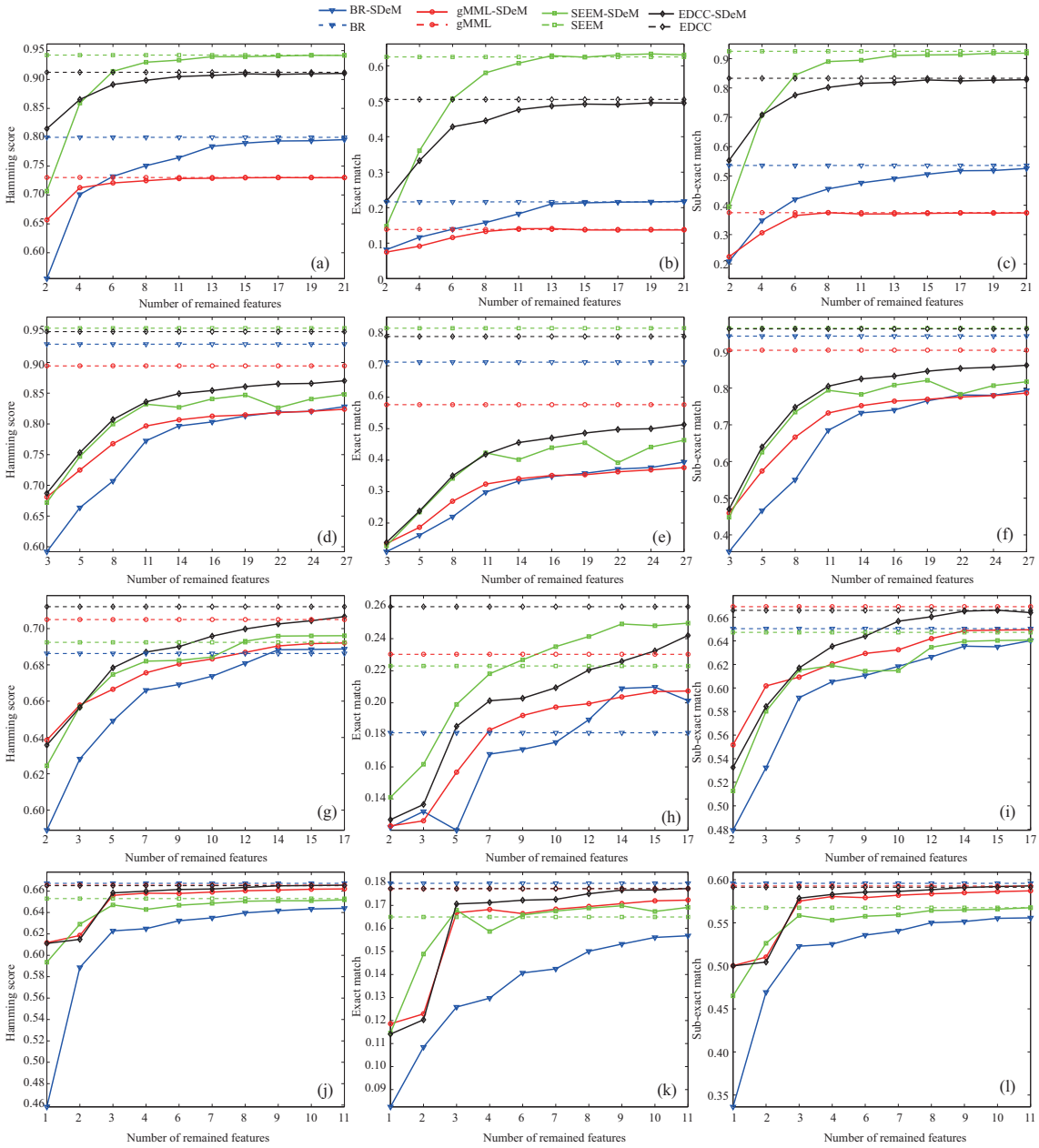

图 1 (网络版彩图) $\mathcal{A}$-SDeM 的性能随降维特征维度 d' 的变化曲线

Figure 1 (Color online) Performance of $\mathcal{A}$-SDeM changes with different reduced feature dimensions d'. (a) Hamming score (Rf1); (b) exact match (Rf1); (c) sub-exact match (Rf1); (d) Hamming score (CoIL); (e) exact match (CoIL); (f) sub-exact match (CoIL); (g) Hamming score (Adult); (h) exact match (Adult); (i) sub-exact match (Adult); (j) Hamming score (Default); (k) exact match (Default); (l) sub-exact match (Default)

围. 再例如, SDeM 方法属于特征映射范畴的线性降维方式, 降维特征仅是原始特征的线性组合致使特征变换能力有限, 因此未来可以研究特征映射途径的非线性降维方法, 甚至从特征选择角度研究面向多维分类数据的降维问题. 另外, 当前越来越多的实际应用需要在开放环境场景 [56,57] 或少量标注样本场景下 [58] 进行学习, 此时会遇到传统机器学习中没有的新问题, 例如物联网应用中常会涉及流数

据和特征增减等难题 [59], 因此专门研究如何在这类特殊应用场景下对多维分类数据进行降维也是值得探索的方向.

参考文献

1 Read J, Bielza C, Larrañaga P. Multi-dimensional classification with super-classes. IEEE Trans Knowl Data Eng, 2014, 26: 1720–1733

2 Shatkay H, Pan F, Rzhetsky A, et al. Multi-dimensional classification of biomedical text: toward automated, practical provision of high-utility text to diverse users. Bioinformatics, 2008, 24: 2086–2093

3 Serafino F, Pio G, Ceci M, et al. Hierarchical multidimensional classification of web documents with MultiWebClass. In: Proceedings of the 18th International Conference on Discovery Science, Banff, 2015. 236–250

4 Lucey P, Cohn J F, Prkachin, K M, et al. Painful data: the UNBCMcMaster shoulder pain expression archive database. In: Proceedings of the 9th IEEE International Conference on Automatic Face and Gesture Recognition, Santa Barbara, 2011. 57–64

5 Valstar M F, Almaev T R, Girard J M, et al. FERA 2015 — second facial expression recognition and analysis challenge. In: Proceedings of the 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, Ljubljana, 2015. 1–8

6 Borchani H, Bielza C, Toro C, et al. Predicting human immunodeficiency virus inhibitors using multi-dimensional Bayesian network classifiers. Artif Intell Med, 2013, 57: 219–229

7 Mihaljevic B, Bielza C, Benavides-Piccione R, et al. Multi-dimensional classification of GABAergic interneurons with Bayesian network-modeled label uncertainty. Front in Computat Neurosci, 2014, 8: 150

8 Džeroski S, Demšar D, Grbović J. Predicting chemical parameters of river water quality from bioindicator data. Appl Intelligence, 2000, 13: 7–17

9 Verma S P, Uscanga-Junco O A, Díaz-González L. A statistically coherent robust multidimensional classification scheme for water. Sci Total Environ, 2021, 750: 141704

10 Keogh E, Mueen A. Curse of dimensionality. In: Encyclopedia of Machine Learning and Data Mining. Boston: Springer, 2017. 314–315

11 Bach F. Breaking the curse of dimensionality with convex neural networks. J Mach Learn Res, 2017, 18: 629–681

12 Cunningham J P, Ghahramani Z. Linear dimensionality reduction: survey, insights, and generalizations. J Mach Learn Res, 2016, 16: 2859–2900

13 Vlachos M. Dimensionality reduction. In: Encyclopedia of Machine Learning and Data Mining. Boston: Springer, 2017. 354–361

14 Ayesha S, Hanif M K, Talib R. Overview and comparative study of dimensionality reduction techniques for high dimensional data. Inf Fusion, 2020, 59: 44–58

15 Gretton A, Bousquet O, Smola A, et al. Measuring statistical dependence with Hilbert-Schmidt norms. In: Proceedings of the 16th International Conference on Algorithmic Learning Theory, Berlin, 2005. 63–77

16 Zhang M L, Zhou Z H. A review on multi-label learning algorithms. IEEE Trans Knowl Data Eng, 2014, 26: 1819–1837

17 Gibaja E, Ventura S. A tutorial on multilabel learning. ACM Comput Surv, 2015, 47: 1–38

18 Jia B-B, Zhang M-L. Multi-dimensional classification via kNN feature augmentation. Pattern Recognition, 2020, 106: 107423

19 van der Gaag L C, de Waal P R. Multi-dimensional Bayesian network classifiers. In: Proceedings of the 3rd European Workshop on Probabilistic Graphical Models, Prague, 2006. 107–114

20 Bielza C, Li G, Larrañaga P. Multi-dimensional classification with Bayesian networks. Int J Approximate Reasoning, 2011, 52: 705–727

21 Gil-Begue S, Bielza C, Larrañaga P. Multi-dimensional Bayesian network classifiers: a survey. Artif Intell Rev, 2021, 54: 519–559

22 Zaragoza J H, Sucar L E, Morales E F, et al. Bayesian chain classifiers for multidimensional classification. In: Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, 2011. 2192–2197

23 Read J, Martino L, Luengo D. Efficient monte carlo methods for multi-dimensional learning with classifier chains. Pattern Recognition, 2014, 47: 1535–1546

24 Jia B-B, Zhang M-L. Decomposition-based classifier chains for multi-dimensional classification. IEEE Trans Artif Intell, 2022, 3: 176–191

25 Arias J, Gamez J A, Nielsen T D, et al. A scalable pairwise class interaction framework for multidimensional classification. Int J Approx Reason, 2016, 68: 194–210

26 Jia B-B, Zhang M-L. Multi-dimensional classification via stacked dependency exploitation. Sci China Inf Sci, 2020, 63: 222102

27 Jia B-B, Zhang M-L. MD-KNN: an instance-based approach for multi-dimensional classification. In: Proceedings of the 25th International Conference on Pattern Recognition, Milan, 2020. 126–133

28 Jia B-B, Zhang M-L. Maximum margin multi-dimensional classification. IEEE Trans Neural Netw Learn Syst, 2022, 33: 7185–7198

29 Wang H, Chen C, Liu W, et al. Incorporating label embedding and feature augmentation for multi-dimensional classification. In: Proceedings of the 34th AAAI Conference on Artificial Intelligence, New York City, 2020. 6178–6185

30 Jia B-B, Zhang M-L. Multi-dimensional classification via selective feature augmentation. Mach Intell Res, 2022, 19: 38–51

31 Ma Z, Chen S. Multi-dimensional classification via a metric approach. Neurocomputing, 2018, 275: 1121–1131

32 Jia B-B, Zhang M-L. Multi-dimensional classification via sparse label encoding. In: Proceedings of the 38th International Conference on Machine Learning, Virtual Conference, 2021. 4917–4926

33 Jia B-B, Zhang M-L. Multi-dimensional classification via decomposed label encoding. IEEE Trans Knowl Data Eng, 2023, 35: 1844–1856

34 Li Z C, Tang J H, Mei T. Deep collaborative embedding for social image understanding. IEEE Trans Pattern Anal Mach Intell, 2019, 41: 2070–2083

35 Li Z C, Tang J H. Semi-supervised local feature selection for data classification. Sci China Inf Sci, 2021, 64: 192108

36 Li Z C, Tang J J. Unsupervised feature selection via nonnegative spectral analysis and redundancy control. IEEE Trans Image Process, 2015, 24: 5343–5355

37 Yan Z, Xiang X G, Li Z C. Item correlation modeling in interaction sequence for graph convolutional session recommendation. Sci Sin Inform, 2022, 52: 1069–1082 [闫昭, 项欣光, 李泽超. 基于交互序列商品相关性建模的图卷积会话推荐. 中国科学: 信息科学, 2022, 52: 1069–1082]

38 Siblini W, Kuntz P, Meyer F. A review on dimensionality reduction for multi-label classification. IEEE Trans Knowl Data Eng, 2021, 33: 839–857

39 Zhang Y, Zhou Z H. Multilabel dimensionality reduction via dependence maximization. ACM Trans Knowl Discov Data, 2010, 4: 1–21

40 Wang H, Ding C, Huang H. Multi-label linear discriminant analysis. In: Proceedings of the 11th European Conference on Computer Vision, Heraklion, 2010. 126–139

41 Wang T, Dai X, Liu Y. Learning with Hilbert-Schmidt independence criterion: a review and new perspectives. Knowledge-Based Syst, 2021, 234: 107567

42 Gangeh M J, Zarkoob H, Ghodsi A. Fast and scalable feature selection for gene expression data using Hilbert-Schmidt independence criterion. IEEE ACM Trans Comput Biol Bioinf, 2017, 14: 167–181

43 Bao W X, Hang J Y, Zhang M L. Partial label dimensionality reduction via confidence-based dependence maximization. In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, 2021. 46–54

44 Song L, Smola A, Gretton A, et al. A dependence maximization view of clustering. In: Proceedings of the 24th International Conference on Machine Learning, Corvallis, 2007. 815–822

45 He X, Niyogi P. Locality preserving projections. In: Proceedings of Advances in Neural Information Processing Systems, Vancouver & Whistler, 2003. 153–160

46 Zhuge W Z, Fan R D, Luo T J, et al. Incomplete multi-view clustering via independent self-representation learning. Sci Sin Inform, 2022, 52: 1186–1203 [诸葛文章, 范瑞东, 罗廷金, 等. 基于独立自表达学习的不完全多视图聚类. 中国科学: 信息科学, 2022, 52: 1186–1203]

47 Spyromitros-Xioufis E, Tsoumakas G, Groves W, et al. Multi-target regression via input space expansion: treating targets as inputs. Mach Learn, 2016, 104: 55–98

48 Cheng W, Dembczyński K, Hüllermeier E. Graded multi-label classification: the ordinal case. In: Proceedings of the

27th International Conference on Machine Learning, Haifa, 2010. 223–230
49 Liu C, Zhao P, Huang S J, et al. Dual set multi-label learning. In: Proceedings of the 32nd AAAI Conference on Artificial Intelligence, New Orleans, 2018. 3635–3642
50 Dua D, Graff C. UCI machine learning repository [http://archive.ics.uci.edu/]. Irvine: University of California
51 Huiskes M J, Lew M S. The MIR flickr retrieval evaluation. In: Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval, Vancouver, 2008. 39–43
52 Mavadati S M, Mahoor M H, Bartlett K, et al. DISFA: a spontaneous facial action intensity database. IEEE Trans Affective Comput, 2013, 4: 151–160
53 Yeh I C, Lien C H. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Syst Appl, 2009, 36: 2473–2480
54 Fan R E, Chang K W, Hsieh C J, et al. LIBLINEAR: a library for large linear classification. J Mach Learn Res, 2008, 9: 1871–1874
55 Demsar J. Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res, 2006, 7: 1–30
56 Zhou Z H. Open-environment machine learning. Natl Sci Rev, 2022, 9: nwac123
57 Parmar J, Chouhan S, Raychoudhury V, et al. Open-world machine learning: applications, challenges, and opportunities. ACM Comput Surv, 2023, 55: 1–37
58 Huang T, Jia B-B, Zhang M-L. Progressive label propagation for semi-supervised multi-dimensional classification. In: Proceedings of the 32nd International Joint Conference on Artificial Intelligence, Macau, 2023. 3821–3829
59 Hou C, Zhou Z H. One-pass learning with incremental and decremental features. IEEE Trans Pattern Anal Mach Intell, 2018, 40: 2776–2792

Supervised dimensionality reduction for multi-dimensional classification

Bin-Bin JIA[1,2] & Min-Ling ZHANG[1,3*]

1. *School of Computer Science and Engineering, Southeast University, Nanjing* 210096*, China*;
2. *College of Electrical and Information Engineering, Lanzhou University of Technology, Lanzhou* 730050*, China*;
3. *Key Laboratory of Computer Network and Information Integration* (*Southeast University*)*, Ministry of Education, Nanjing* 210096*, China*
* Corresponding author. E-mail: zhangml@seu.edu.cn

Abstract Compared to traditional multi-class classification, each object in multi-dimensional classification is also represented by a single instance while associated with multiple class variables. Here, each class variable corresponds to one heterogeneous class space characterizing an object's semantics from one dimension. Dimensionality reduction effectively alleviates the curse of dimensionality and expedites model training. Existing multi-dimensional classification studies aim at designing learning algorithms with better performance, while the problem of dimensionality reduction for multi-dimensional classification has not been investigated. According to the correlation between feature space and semantic space, this paper makes a first attempt at designing a supervised linear dimensionality reduction method called SDeM for multi-dimensional classification. SDeM measures the correlation between two spaces with the Hilbert-Schmidt independence criterion and determines the projection matrix by maximizing the correlation between the projected feature space and the semantic space under this metric. Experimental results show that the reduced features obtained by SDeM are more conducive than those obtained by unsupervised dimensionality reduction methods to achieve better generalization performance for multi-dimensional classification methods.

Keywords machine learning, multi-dimensional classification, dimensionality reduction, dependence between spaces, Hilbert-Schmidt independence criterion