中国科学: 信息科学 2018 年 第 48 卷 第 1 期: 47–59
SCIENTIA SINICA Informationis
论文

利用辅助信息进行矩阵补全的核方法及其在多标记学习中的应用

徐淼[1,2], 周志华[1,2*]

1. 南京大学计算机软件新技术国家重点实验室, 南京 210023
2. 软件新技术与产业化协同创新中心, 南京 210023
* 通信作者. E-mail: zhouzh@lamda.nju.edu.cn

收稿日期: 2016–12–13; 接受日期: 2017–03–24; 网络出版日期: 2017–09–14
国家自然科学基金 (批准号: 61333014) 资助项目

摘要 现实机器学习任务中一个样本通常和多个标记相关, 但获取完整的标记信息需耗费大量人力物力, 因此多标记学习经常会遇到标记缺失的情况. 将未缺失的标记看作不完全的标记矩阵, 将样本特征作为辅助信息, 则可通过矩阵补全方法来解决该问题. 以往研究主要针对线性可分情形, 本文提出 KernelMaxide 方法, 在处理线性不可分多标记数据中缺失的监督信息的同时, 不仅能利用数据的非线性结构, 还能考虑标记之间的相互关系. 该方法依据矩阵核范数的表示定理, 构建了基于核矩阵的核范数最小化优化目标以及相应的优化算法, 并用 Nyström 方法缓解核矩阵的存储和计算开销问题. 实验显示出 KernelMaxide 的优越性能.

关键词 机器学习, 多标记学习, 矩阵补全, 核方法, Nyström 方法

1 引言

传统的机器学习通常认为一个样本仅仅和一个标记相关, 而在现实世界中一个样本通常和多个标记相关. 多标记学习 [1] 因为假设一个样本和多个标记相关, 所以相对于传统的单标记学习, 它可以更精确地描述现实, 因而在图像分类 [2]、文本分类 [3]、社交网络 [4]、生物信息学 [5] 和中医诊断 [6] 等多个领域得到了广泛的应用.

相对于传统的单标记学习, 标记多标记数据时, 需要大量的人工来检查所有的候选标记是否和当前样本相关, 非常费时费力; 因此在实际应用中, 很容易遇到多标记数据标记不全的情况 [7~9], 即每个训练样本都有部分标记被标出, 剩余的标记缺失. 对于这种每个样本都有部分标记缺失的问题, 并不能用半监督学习来处理, 因为半监督学习假设了未标记样本的所有标记都是缺失的. 针对这个问题, 最简单的处理方法就是把一个多标记问题看成若干个二分类问题, 每个标记对应一个二分类问题, 这样

引用格式: 徐淼, 周志华. 利用辅助信息进行矩阵补全的核方法及其在多标记学习中的应用. 中国科学: 信息科学, 2018, 48: 47–59, doi: 10.1360/N112016-00279
Xu M, Zhou Z-H. Kernel method for matrix completion with side information and its application in multi-label learning (in Chinese). Sci Sin Inform, 2018, 48: 47–59, doi: 10.1360/N112016-00279

在学习时, 就不需要考虑缺失的标记, 而只要用未缺失的标记对应的样本做训练数据就可以了 [10]. 这种方法简单直接, 但是它忽略了标记之间的关系, 而在多标记学习中, 标记关系对于学习效率和效果有很大作用 [1]. 当把所有标记看成一个有缺失的标记矩阵, 并且假设标记之间有相关性, 因此形成的矩阵是低秩的, 这样就可以利用矩阵补全技术 [11] 解决监督信息缺失问题.

矩阵补全可以看成张量分解的特殊形式, 即矩阵是一个 2 阶的张量. 经典的矩阵补全技术假设目标矩阵是低秩的, 并且矩阵内的元素随机缺失: 对于大小为 $n \times n$ 的 r 秩矩阵, 样本复杂度 (即为了完美恢复该矩阵, 需要观测到的最少元素个数) 为 $O(n \log^2 n)$ [11]. 而 Xu 等 [8] 提出的 Maxide 利用多标记的属性信息和标记关系作为辅助信息来帮助矩阵补全; 在有了辅助信息的帮助之后, 样本复杂度减少到 $O(\log n)$ [8], 相对于经典的矩阵补全技术 [11], 在样本复杂度上有了显著提升. 利用辅助信息的矩阵补全方法在一些监督信息缺失的多标记数据上虽然取得了不错的效果, 但是它针对数据是线性可分的情况展开研究, 而实际上有些多标记数据不是线性可分的, 一些工作 [5,12~14] 显示, 对于线性不可分的数据, 利用数据非线性结构的核方法会比传统的线性方法取得更好的效果. 因此, 如何在利用矩阵补全方法处理缺失监督信息的同时, 利用数据的非线性结构以提升分类精度就成了亟待解决的问题.

针对数据监督信息缺失且线性不可分的情况, 一种简单的解决方法是将原来的多标记问题拆成多个二分类问题, 为每个标记训练一个二分类的核分类器, 比如 KernelSVM, 这种方法通常被称为 KernelBSVM [10]. 但是, 这种方法忽略了标记之间的相关性. 在不能利用标记之间的相关性时, 为了训练出表现良好的分类器, 就要求每个标记都有充分的训练数据, 因此当标记缺失情况较严重而不能保证每个标记都有充分的训练数据时, KernelBSVM 就很难训练出一个表现良好的分类器. 矩阵补全方法则通过利用矩阵的低秩性质, 考虑了标记之间的相关性, 即使某些标记没有充分的训练数据, 也可以利用其与其他标记的相关性来获得较好的分类器. 因此, 用矩阵补全方法解决数据监督信息缺失且线性不可分的问题是很有必要的. 但是, 在考虑使用矩阵补全方法时面临两个问题, 一是如何在矩阵补全方法中利用核矩阵; 二是利用核矩阵时会面临数据量很大以致核矩阵无法存储或计算的情况, 应该如何寻找核矩阵的近似以减轻算法的存储和计算负担.

本文提出了利用辅助信息进行矩阵补全的核方法 KernelMaxide, 依据矩阵范数的表示定理 [15], 构建了基于核矩阵的核范数最小化优化目标, 以及相应的优化算法, 在矩阵补全方法中利用了核矩阵, 并且针对核矩阵的存储和计算负担较重问题, 进一步将 Nyström 方法 [16] 整合到了基于核矩阵的优化目标中. 实验显示 KernelMaxide 方法在监督信息缺失的多标记学习数据集上取得了较好的效果, 无论相对于针对线性可分数据的 Maxide 方法 [8], 还是相对于不考虑标记相关性的 KernelBSVM 方法 [10], 其分类精度都有很大提升.

本文后续部分的组织如下: 第 2 节对多标记学习、监督信息缺失的多标记学习、矩阵补全以及核方法的相关工作进行总结; 第 3 节提出了利用辅助信息进行矩阵补全的核方法 KernelMaxide; 第 4 节展示了 KernelMaxide 方法在监督信息缺失的多标记数据上的实验效果; 最后总结全文, 并对未来可能的研究方向进行展望.

2 相关工作

2.1 多标记学习

多标记学习假设一个样本和多个标记相关, 它吸引了集成学习 [2]、度量学习 [17] 和在线学习 [18] 等机器学习相关领域专家的兴趣; 在图像标记 [2]、文本分类 [3] 和生物信息学 [5] 等多个方面取得了广

泛的应用; 在海量标记 [19] 与新类问题 [20] 等方面开拓了新的研究领域, 并成为近几年机器学习研究的热门方向. 多标记学习也有多种度量指标来度量学习效果, 比较常用的如 Hamming Loss [21] 计算了分类错误的标记所占的比例; Ranking Loss [5] 度量了排错的相关标记 – 不相关标记对的个数; Average Precision [22] 度量了排在相关标记前的相关标记的平均个数. 更多关于多标记学习及其损失函数的介绍, 可以参考相关综述 [1].

解决多标记学习问题最直接的方法是 BinaryRelevance (BR) [10]. BR 将一个多标记学习问题看成若干个二分类问题, 为每个标记单独学一个分两类的分类器. 这个方法在多标记学习研究的早期取得了一定的效果. 随着多标记学习领域的发展, BR 因为不能利用标记之间的关系而受诟病 [1]. Label Powerset (LP) 方法 [23] 是最直接的利用标记之间相关性的方法, 它将每种可能的标记组合都看成一个新的标记, 用传统的多类分类器解决多标记学习问题; 但是, 不是每种标记组合都能找到对应的训练数据, 所以 LP 方法常常面临数据缺失问题. 为了解决 LP 方法面临的问题, 一种方法是利用外来的数据学习标记之间的关系 [24]; 更多的方法从训练数据本身学习标记之间的关系 [22,25]. 这些算法因为考虑了标记之间的相互关系, 在很多多标记学习任务中都取得了不错的效果.

除了利用标记间的关系之外, 另一种提高多标记学习分类精度的方法就是利用非线性分类器. 比如 Elisseeff 等 [5] 提出了一种大间隔分类器 RankSVM 方法, 可以利用各种核矩阵来提高多标记学习的分类精度; 文献 [12∼14] 提出了针对多标记学习的多核方法, 即利用多个核的线性加和来学习一个新的非线性分类器以提高学习精度. 因此, 相对于简单的线性分类器, 本文将着重研究如何构建一个利用核矩阵的非线性分类器, 以及如何在利用核矩阵的同时利用标记之间的相互关系.

2.2 监督信息缺失的多标记学习和矩阵补全方法

近几年来, 很多研究者意识到, 得到多标记数据完整的监督信息是一件很难的事, 需要耗费大量的人力物力去标注所有样本的所有标记, 于是开始研究怎样获得廉价的标记数据, 比如利用众包技术 [26] 和主动学习技术 [27]. 除此之外, 更多的研究则侧重于直接利用监督信息缺失的多标记数据来做学习 [7∼9,28]. 这些方法通常假设标记矩阵是低秩的, 然后利用低秩矩阵的性质来补全标记. 矩阵补全方法 [11] 就是一种专门用于补全不完整低秩矩阵的方法, 其作为张量分解的特殊形式, 具有一定的理论依据: 对于大小为 $n \times n$ 的 r 秩矩阵, 样本复杂度 (即为了完美恢复该矩阵, 需要观测到的最少元素个数) 为 $O(n \log^2 n)$ [11].

Xu 等 [8] 提出了利用辅助信息进行矩阵补全的方法 Maxide, 相对于经典的矩阵补全技术, 除了在样本复杂度上有了显著提升外, Maxide 在每一轮循环中只需要对一个相对较小的矩阵做 SVD 分解, 相对于经典的矩阵补全方法同时提高了时间效率和空间效率. 虽然 Maxide 在理论和实验中都取得了较好的效果, 然而它主要针对数据是线性可分的情况, 对于线性不可分的数据应该如何学习, Maxide 并没有定论. 本文将提出利用辅助信息进行矩阵补全的核方法, 在利用辅助信息补全缺失监督信息的同时, 可以利用数据的非线性结构以取得更好的分类精度.

2.3 核方法

核方法 [29], 比如 KernelSVM [30], 通常被认为是最有效的机器学习方法之一. 这些方法把原始数据点投影到高维甚至无限维的空间, 在新的高 (无限) 维空间里学习一个线性分类器. 这相当于在原始空间里学习了一个非线性分类器, 在原始学习任务不是线性可分的情况下, 通常能取得较好的效果. 虽然核方法能取得不错的分类效果, 但其同时面临着计算和存储效率问题. 当样本的个数是 n 的时候, 使用

核方法通常需要对一个大小为 $n\times n$ 的矩阵进行操作; 当 n 比较大的时候, 存储和计算这个 $n\times n$ 的矩阵就会带来很多的负担. 一些工作可以有效地解决这个问题, 比如随机 Fourier 特征 (random Fourier feature) [31] 或者 Nyström 方法 [16]. 随机 Fourier 特征采样一些 cos 和 sin 函数来近似原始核矩阵; Nyström 方法从训练样本中随机采样一小部分样本, 并利用这一小部分样本与其他样本之间的核来近似 $n\times n$ 的核矩阵. Yang 等 [32] 的研究显示, 在核矩阵的特征值谱 (eigen-spectral) 的差异 (gap) 比较大的时候, Nyström 方法有更低的泛化错误率. 本文将考虑如何利用 Nyström 方法, 减轻所提出的核方法的计算和存储负担.

综上所述, 本文将提出利用辅助信息进行矩阵补全的核方法, 并将其应用到多标记学习中. 这种方法针对多标记学习中监督信息缺失的情况, 不仅考虑了标记之间的相关性, 也考虑了在原始特征空间非线性可分的情况下, 如何利用核方法训练一个非线性分类器, 以提高分类精度.

3 KernelMaxide 方法

3.1 经典的利用辅助信息进行矩阵补全算法在多标记学习中的应用

本节将讨论利用辅助信息进行矩阵补全的算法 Maxide 及其在多标记学习中的应用 [8].

$M\in\mathbb{R}^{n\times m}$ 表示秩为 r 的待恢复标记矩阵, 其中 n 表示样本的个数, m 表示标记的个数. $\|\cdot\|_{\mathrm{tr}}$ 表示矩阵的核范数, 即其奇异值的和; $\|\cdot\|_{\mathrm{F}}$ 表示矩阵的 Frobenius 范数, 即所有元素的平方和的平方根. $\Omega\subset\{1,\dots,n\}\times\{1,\dots,m\}$ 表示矩阵 M 中观测到的元素的位置下标的集合, 即 Ω 以外的元素在矩阵 M 中是缺失的. 给定 Ω, 则可以定义线性算子 $\mathcal{R}_\Omega(M):\mathbb{R}^{n\times m}\to\mathbb{R}^{n\times m}$,

$$[\mathcal{R}_\Omega(M)]_{i,j}=\begin{cases}M_{i,j}, & (i,j)\in\Omega,\\ 0, & (i,j)\notin\Omega.\end{cases}$$

传统的利用辅助信息进行矩阵补全的算法 [8] 进行多标记学习时有如下优化目标:

$$\min_{W\in\mathbb{R}^{k\times m}}\mathcal{L}(W)=\lambda\|W\|_{\mathrm{tr}}+\frac{1}{2}\|\mathcal{R}_\Omega(AW-Y)\|_{\mathrm{F}},\tag{1}$$

其中 $Y\in\mathbb{R}^{n\times m}$ 是由未缺失的标记组成的矩阵 (缺失部分为 0), A 来源于多标记数据的特征矩阵 $X\in\mathbb{R}^{n\times d}$, d 是特征个数. 一般情况下 A 是标准正交矩阵, 于是有 $\|W\|_{\mathrm{tr}}=\|AW\|_{\mathrm{tr}}$; 在 A 不是标准正交矩阵的情况下, 可以对原特征 X 做 SVD 分解, 并将 A 设定为 X 的右特征值矩阵 (即 X 的列空间). 优化式 (1), 即为利用样本特征 X 作为辅助信息进行矩阵补全的算法 Maxide, 由于这里假设了目标矩阵 $M=AW$, 而 A 是 X 所在的空间, 所以 Maxide 算法利用了目标矩阵 M 和特征 X 之间的线性关系.

3.2 利用辅助信息进行矩阵补全的核方法

本节将提出利用辅助信息进行矩阵补全的核方法. 一般的核方法通常假设有一个投影函数 $\phi(\cdot)$, 将低维特征投影到一个高维空间 (甚至无穷维空间). 假设目标矩阵 W 是低秩的, 将辅助信息矩阵 A 投影到 $\phi(A)$ 之后原始优化目标式 (1) 变成

$$\min_{W\in\mathbb{R}^{k\times m}}\mathcal{L}(W)=\lambda\|W\|_{\mathrm{tr}}+\frac{1}{2}\|\mathcal{R}_\Omega(\phi(A)W-Y)\|_{\mathrm{F}}.\tag{2}$$

对式 (2) 使用矩阵核范数的表示定理 (the representor theorem of matrix norm) [15], 于是目标 W 的第 t 列 w_t 可以被表示为 $\phi(a_i)$ 的线性加和, 其中 $\phi(a_i)$ 表示矩阵 $\phi(A)$ 的第 i 行, 即第 i 个样本被投影之后的特征表示. 假设在多个标记中有任意标记是非缺失的样本系数所在的集合为 Uniform Training (UT), 即 $\mathrm{UT}=\{i:\exists j,(i,j)\in\Omega\}$, 则 w_t 和 W 用 $\phi(a_i)$ 的线性加和来表示, 可以分别形式化为

$$w_t=\sum_{i\in\mathrm{UT}}c_i\phi(a_i),\quad W=[w_1,w_2,\ldots,w_m]=\phi(A_{\mathrm{UT}})^{\mathrm{T}}C. \tag{3}$$

用式 (3) 中的 $\phi(A_{\mathrm{UT}})$ 和 C 表示 W 之后, 有

$$\phi(A)W=\phi(A)\phi(A_{\mathrm{UT}})^{\mathrm{T}}C=\boldsymbol{K}_{\mathrm{all,UT}}C, \tag{4}$$

其中 $\boldsymbol{K}_{\mathrm{all,UT}}$ 是全部样本 (共 n 个) 和 UT 里的样本两两之间的核组成的矩阵. 将式 (3) 代入式 (4), 则有

$$\min_{C\in\mathbb{R}^{n_{\mathrm{UT}}\times m}}\mathcal{L}(C)=\lambda\|\phi(A_{\mathrm{UT}})^{\mathrm{T}}C\|_{\mathrm{tr}}+\frac{1}{2}\|\mathcal{R}_\Omega(\boldsymbol{K}_{\mathrm{all,UT}}C-Y)\|_{\mathrm{F}}, \tag{5}$$

其中 $n_{\mathrm{UT}}=|\mathrm{UT}|$. 由于标记之间的相关性, 假设线性分类系数 W 是低秩的, 而 $\phi(A_{\mathrm{UT}})$ 是原低维空间投影到高维空间之后的结果, 通常是满秩的, 所以新的目标矩阵 C 需要是低秩的. 在 C 是低秩时, 式 (5) 转换成如下的优化问题:

$$\min_{C\in\mathbb{R}^{n_{\mathrm{UT}}\times m}}\mathcal{L}(C)=\lambda\|C\|_{\mathrm{tr}}+\frac{1}{2}\|\mathcal{R}_\Omega(\boldsymbol{K}_{\mathrm{all,UT}}C-Y)\|_{\mathrm{F}}. \tag{6}$$

3.3 利用 Nyström 方法解决大数据问题

在式 (6) 中, 当核矩阵较小时, 可以直接优化; 而通常情况下, 核矩阵都比较大, 甚至大到内存无法存储的情况. 接下来, 将使用 Nyström 方法 [16] 来处理较大的核矩阵的计算和存储问题, 以提高算法的时间和空间效率.

Nyström 方法首先从大小为 n 的样本空间 $D=\{x_1,\ldots,x_n\}$ 中随机采样 n_r 个样本, $\hat{D}=\{\hat{x}_1,\ldots,\hat{x}_{n_r}\}$, 则 $\hat{D}$ 上的核矩阵为 $\boldsymbol{K}_{n_r\times n_r}=[\kappa(\hat{x}_i,\hat{x}_j)]_{n_r\times n_r}$. 在本文中主要使用 RBF 核, 即

$$\kappa(x_1,x_2)=\exp\left\{-\|x_1-x_2\|_2^2/2\sigma^2\right\}.$$

$\hat{D}$ 中的样本和原始样本空间中的样本组成的核矩阵为 $\boldsymbol{K}_{n\times n_r}=[\kappa(x_i,\hat{x}_j)]_{n\times n_r}$, 则原始的核矩阵 $\boldsymbol{K}_{n\times n}=[\kappa(x_i,x_j)]_{n\times n}$ 可以被近似为

$$\widetilde{\boldsymbol{K}}_{n\times n}\approx\boldsymbol{K}_{n\times n_r}\boldsymbol{K}^{\dagger}_{n_r\times n_r}\boldsymbol{K}^{\mathrm{T}}_{n\times n_r}, \tag{7}$$

其中 $\boldsymbol{K}^{\dagger}_{n_r\times n_r}$ 表示矩阵 $\boldsymbol{K}_{n_r\times n_r}$ 的伪逆矩阵. Yang 等 [32] 理论证明了使用 Nyström 方法后核矩阵的谱范数近似错误率是 $O(n_r^{-1/2})$. 本文将直接利用式 (7) 来计算全部样本的核矩阵 $\widetilde{\boldsymbol{K}}_{n\times n}$, 并取 $\widetilde{\boldsymbol{K}}_{n\times n}$ 中在 Ω 中出现过的样本对应的列组成 $\widetilde{\boldsymbol{K}}_{\mathrm{all,UT}}$, 即 $\widetilde{\boldsymbol{K}}_{\mathrm{all,UT}}$ 是由 $\widetilde{\boldsymbol{K}}_{n\times n}$ 的第 t 列组成的矩阵, $t\in\{j:(i,j)\in\Omega\}$.

3.4 优化

为了高效地优化式 (6), 将采用加速梯度下降方法 (accelerated gradient descend) [33]. 这个方法利用了目标函数的平滑特征, 获得了 $O(1/T^2)$ 的收敛率, 其中 T 是循环进行的次数. 优化式 (6) 所得的 KernelMaxide 算法如算法 1 所示, 其中 ϵ 通常设置为一个很小的常数. 为了解决第 4 行的优化问题, 通常用 Singluar Value Thresholding 方法 [34].

算法 1 利用辅助信息进行矩阵补全的核方法 KernelMaxide

算法输入: Ω, Y, λ, $\widetilde{\boldsymbol{K}}_{\mathrm{all,UT}}$

算法过程:

初始化: $C_1 = C_2$, L, $\gamma > 1$, $\theta_1 = \theta_2 \in (0, 1]$, ϵ, $k = 2$
while $\widetilde{\mathcal{L}}(C_{k+1}) \leqslant (1-\epsilon)\widetilde{\mathcal{L}}(C_k)$ **do**
  $Z_k = C_k + \theta_k(1/\theta_{k-1} - 1)(C_k - C_{k-1})$
  $C_{k+1} = \mathrm{argmin}\ \ \lambda\|C\|_{\mathrm{tr}} + Q_k(C)$
  while $\mathcal{E}(C_{k+1}) - \mathcal{E}(C_k) \geqslant H_k(L)$ **do**
    $L = L\gamma$
    $C_{k+1} = \mathrm{argmin}\ \lambda\|C\|_{\mathrm{tr}} + Q_k(C)$
  end while
  $\theta_{k+1} = (\sqrt{\theta_k^4 + 4\theta_k^2} - \theta_k^2)/2$
  $k = k + 1$
end while

算法输出: C_{k+1}

在算法 1 中,

$$\widetilde{\mathcal{L}}(C) = \lambda\|C\|_{\mathrm{tr}} + \frac{1}{2}\|\mathcal{R}_\Omega(\widetilde{\boldsymbol{K}}_{\mathrm{all,UT}}C - Y)\|_{\mathrm{F}},$$

$$\mathcal{E}(C) = \frac{1}{2}\|\mathcal{R}_\Omega(\widetilde{\boldsymbol{K}}_{\mathrm{all,UT}}C - Y)\|_{\mathrm{F}}^2,$$

$$Q_k(C) = \frac{L}{2}\left\|C - \left(Z_k - \frac{1}{L}\widetilde{\boldsymbol{K}}_{\mathrm{all,UT}}^{\mathrm{T}}\mathcal{R}_\Omega(AY_k - Y)\right)\right\|_F^2,$$

$$H_k(L) = \mathrm{Tr}\left((C_{k+1} - Z_k)^{\mathrm{T}}\widetilde{\boldsymbol{K}}_{\mathrm{all,UT}}^{\mathrm{T}}\mathcal{R}_\Omega(\widetilde{\boldsymbol{K}}_{\mathrm{all,UT}}Z_k - Y)\right) + \frac{L}{2}\|C_{k+1} - Z_k\|_{\mathrm{F}}^2.$$

对于算法 1 的相关优化方法的更多细节, 可以参考 Tseng [33] 的文章.

4 实验

本节将在多标记数据集上评估所提出的利用辅助信息进行 “矩阵补全的核方法” (KernelMaxide). 所有代码在 Matlab 平台上实现, 并在 CPU 2.53 GHz 和内存 48 G 的 Linux 服务器上完成实验.

本实验中使用了 “yahoo.com” 网页分类数据集 [35], 此数据集包含 11 个分类任务, 每个分类任务包含 5000 个训练样本. 这 11 个分类任务的特征数和标记数都有较大差异, 其特征数从 438 ∼ 1047 不等, 标记数从 21 ∼ 40 不等. 关于这些数据集的详细信息, 可以参考表 1.

本实验采用和以往工作 [8] 相同的实验方法, 即对每个数据集, 随机采样 10% 的数据作为测试数据, 剩下的 90% 数据作为训练数据. 为了使得训练数据 “不完全”, 对于每个标记, 让 ω% 的正样本和负样本作为可观测到的标记, 而剩下的训练样本标记作缺失处理. ω% 将从 $\{10\%, 20\%, 30\%, 40\%\}$ 中取值. 每个实验被重复 10 次, 呈递的结果将是 10 次实验在测试数据上的平均结果. 对于本文提出的

表 1 11 个 "yahoo.com" 分类数据的详细信息 [a)]

Table 1 Eleven "yahoo.com" datasets [a)]

Dataset	Dimension	Class	MaxL
Arts	462	26	14
Business	438	30	12
Computers	581	33	17
Education	550	33	7
Entertainment	640	21	17
Health	612	32	13
Recreation	606	22	17
Reference	793	33	12
Science	743	40	9
Social	1047	39	10
Society	636	27	16

a) "Dimension" is the number of features. "Class" is the number of candidate labels. "MaxL" denotes the maximum number of relevant labels per instance.

KernelMaxide 方法, 其正则化参数 λ 将通过交叉验证从 $2^{\{-10,-9,\ldots,9,10\}}$ 中选取. 实验中采用 RBF-kernel, 其参数 σ 将以交叉验证从 $2^{\{-5,-4,\ldots,4,5\}}$ 中选取. KernelMaxide 中的参数 γ 和 ϵ 将被分别设置成 2 和 10^{-5}. 最大循环轮数被设置为 100. 根据 Nyström 方法采样的样本数目不同, 本文将对两种 KernelMaxide 方法进行测试, 即 KernelMaxide-f(ull) 以及 KernelMaxide-n(yström). 在 KernelMaxide-f 中, 所有的样本都用来计算核矩阵, 即 $n_r = 5000$. 在 KernelMaxide-n 中, n_r 设置为 1000, 即只用原始数据的 20% 计算核矩阵. 因为 KernelMaxide 及其比较方法 Maxide 仅能为每个标记输出一个实值, 而不能确定这个标记是否相关, 本实验将利用 Ranking Loss 指标来比较这 11 个多标记学习任务上的标记排序结果, 而将分类结果的比较留到后续工作中.

实验中 KernelMaxide 方法将和 2 个方法比较, 一个是传统的利用辅助信息进行矩阵补全的方法, 即 Maxide [8]; 另一个是为每个标记单独训练一个 KernelSVM 分类器的方法, 即 KernelBSVM [10]. 第一个比较方法 Maxide 虽然利用了辅助信息和标记矩阵的低秩性质, 但是并没有考虑特征矩阵的核. KernelBSVM 方法虽然考虑了特征矩阵的核, 但并没有考虑标记之间的关系; 并且 KernelBSVM 方法并没有利用 Nyström 方法减轻数据的计算和存储开销, 仍然利用了所有的数据来计算核矩阵. 对于 Maxide, 实验中使用了和原工作 [8] 同样的调参数方式. 对于 KernelBSVM, 实验中利用 LibSVM [36] 作为基分类器, 其正则化参数和 RBF-kernel 参数的选取, 则使用了和 KernelMaxide 同样的方式.

表 2 展示了 KernelMaxide-f 和比较方法在监督信息缺失的多标记数据上的实验结果. 在所有情况下, KernelMaxide-f 在 Ranking Loss 上的表现都是最优的, 且比两种比较方法显著的好. 一方面, 这说明在 "yahoo.com" 网页分类数据集上, KernelMaxide 方法通过对原始特征的高维投影, 输出了一个精度更高的非线性分类器; 另一方面, 相对于 KernelBSVM, KernelMaxide 方法也利用了标记矩阵的低秩性, 即考虑了标记之间的关系, 从而提高了分类精度.

表 3 展示了 KernelMaxide-n 及比较方法在监督信息缺失的多标记数据上的实验结果. 结果显示, 在用 Nyström 方法对核矩阵做了近似之后, 即使只采样了原始数据的 20% 计算核矩阵, 在多数情况下 (44 次实验中的 41 次), KernelMaxide-n 在 Ranking Loss 上的表现依然比两种比较方法 Maxide 和 KernelBSVM 要显著的好. 对比表 2 和 3, 虽然采用 Nyström 方法对核矩阵做近似部分损坏了算法的

表 2　使用了全部数据计算核矩阵的 KernelMaxide-f 方法在监督信息缺失的多标记数据上的 Ranking Loss (越小越好) 实验结果 [a)]

Table 2　Ranking Loss (the smaller the better) results of KernelMaxide-f on multi-label learning with incomplete supervised information [a)]

Dataset	Algorithm	$\omega\% = 10\%$	$\omega\% = 20\%$	$\omega\% = 30\%$	$\omega\% = 40\%$
Arts	KernelMaxide-f	**0.1514**	**0.1365**	**0.1264**	**0.1266**
	Maxide	0.1596	0.1500	0.1421	0.1422
	KernelBSVM	0.2325	0.2262	0.2138	0.2110
Business	KernelMaxide-f	**0.0424**	**0.0428**	**0.0359**	**0.0366**
	Maxide	0.0457	0.0488	0.0444	0.0458
	KernelBSVM	0.0785	0.0833	0.0722	0.0724
Computers	KernelMaxide-f	**0.0957**	**0.0889**	**0.0785**	**0.0861**
	Maxide	0.1033	0.0974	0.0902	0.0988
	KernelBSVM	0.1676	0.1588	0.1475	0.1565
Education	KernelMaxide-f	**0.0944**	**0.0833**	**0.0839**	**0.0806**
	Maxide	0.1025	0.0917	0.0904	0.0893
	KernelBSVM	0.2249	0.2037	0.2011	0.1915
Entertainment	KernelMaxide-f	**0.1186**	**0.1080**	**0.1064**	**0.0991**
	Maxide	0.1233	0.1199	0.1177	0.1119
	KernelBSVM	0.1911	0.1902	0.1797	0.1707
Health	KernelMaxide-f	**0.0597**	**0.0540**	**0.0491**	**0.0491**
	Maxide	0.0691	0.0652	0.0609	0.0607
	KernelBSVM	0.1130	0.1144	0.1124	0.1127
Recreation	KernelMaxide-f	**0.1674**	**0.1508**	**0.1419**	**0.1389**
	Maxide	0.1824	0.1609	0.1549	0.1526
	KernelBSVM	0.2347	0.2184	0.2143	0.2056
Reference	KernelMaxide-f	**0.0870**	**0.0724**	**0.0720**	**0.0703**
	Maxide	0.1052	0.0946	0.0897	0.0828
	KernelBSVM	0.1567	0.1472	0.1489	0.1448
Science	KernelMaxide-f	**0.1260**	**0.1186**	**0.1117**	**0.1086**
	Maxide	0.1423	0.1329	0.1316	0.1268
	KernelBSVM	0.2140	0.2054	0.1993	0.1983
Social	KernelMaxide-f	**0.0687**	**0.0588**	**0.0569**	**0.0552**
	Maxide	0.0773	0.0718	0.0727	0.0727
	KernelBSVM	0.1155	0.1048	0.1070	0.1086
Society	KernelMaxide-f	**0.1445**	**0.1330**	**0.1298**	**0.1287**
	Maxide	0.1546	0.1454	0.1416	0.1421
	KernelBSVM	0.2141	0.2030	0.2054	0.2014

a) $\omega\%$ represents the percentage of training instances with observed label assignment for each label. The best results and its comparable ones (t-tests at 95% confidence level) are bolded.

效果, 但是绝大多数情况下其表现依然比其他两种比较方法优异.

图 1 比较了两种 KernelMaxide 及其比较方法的时间效率. 在用 Nyström 方法选取了 20% 的数

表 3 使用了 20% 数据计算核矩阵的 KernelMaxide-n 方法在监督信息缺失的多标记数据上的 Ranking Loss (越小越好) 实验结果 [a)]

Table 3 Ranking Loss (the smaller the better) results of KernelMaxide-n on multi-label learning with incomplete supervised information [a)]

Dataset	Algorithm	$\omega\% = 10\%$	$\omega\% = 20\%$	$\omega\% = 30\%$	$\omega\% = 40\%$
Arts	KernelMaxide-n	**0.1527**	**0.1382**	**0.1278**	**0.1289**
	Maxide	0.1596	0.1500	0.1421	0.1422
	KernelBSVM	0.2325	0.2262	0.2138	0.2110
Business	KernelMaxide-n	**0.0434**	**0.0446**	**0.0386**	**0.0393**
	Maxide	0.0457	0.0488	0.0444	0.0458
	KernelBSVM	0.0785	0.0833	0.0722	0.0724
Computers	KernelMaxide-n	**0.0957**	**0.0902**	**0.0811**	**0.0879**
	Maxide	0.1033	0.0974	0.0902	0.0988
	KernelBSVM	0.1676	0.1588	0.1475	0.1565
Education	KernelMaxide-n	**0.0955**	**0.0853**	**0.0861**	**0.0833**
	Maxide	0.1025	0.0917	0.0904	0.0893
	KernelBSVM	0.2249	0.2037	0.2011	0.1915
Entertainment	KernelMaxide-n	**0.1224**	**0.1113**	**0.1124**	**0.1036**
	Maxide	**0.1233**	0.1199	0.1177	0.1119
	KernelBSVM	0.1911	0.1902	0.1797	0.1707
Health	KernelMaxide-n	**0.0616**	**0.0562**	**0.0519**	**0.0503**
	Maxide	0.0691	0.0652	0.0609	0.0607
	KernelBSVM	0.1130	0.1144	0.1124	0.1127
Recreation	KernelMaxide-n	**0.1704**	**0.1568**	**0.1489**	**0.1484**
	Maxide	0.1824	**0.1609**	0.1549	**0.1526**
	KernelBSVM	0.2347	0.2184	0.2143	0.2056
Reference	KernelMaxide-n	**0.0896**	**0.0764**	**0.0749**	**0.0727**
	Maxide	0.1052	0.0946	0.0897	0.0828
	KernelBSVM	0.1567	0.1472	0.1489	0.1448
Science	KernelMaxide-n	**0.1328**	**0.1248**	**0.1175**	**0.1145**
	Maxide	0.1423	0.1329	0.1316	0.1268
	KernelBSVM	0.2140	0.2054	0.1993	0.1983
Social	KernelMaxide-n	**0.0719**	**0.0624**	**0.0609**	**0.0600**
	Maxide	0.0773	0.0718	0.0727	0.0727
	KernelBSVM	0.1155	0.1048	0.1070	0.1086
Society	KernelMaxide-n	**0.1451**	**0.1341**	**0.1326**	**0.1314**
	Maxide	0.1546	0.1454	0.1416	0.1421
	KernelBSVM	0.2141	0.2030	0.2054	0.2014

a) The same as in Table 2.

据对核矩阵做了近似之后, KernelMaxide 方法的时间效率获得显著提升. 由图中可见, KernelMaxide-n 的运行时间相对于 KernelMaxide-f 方法有较大提升, 且接近甚至优于最快的 Maxide 方法. 在用所有

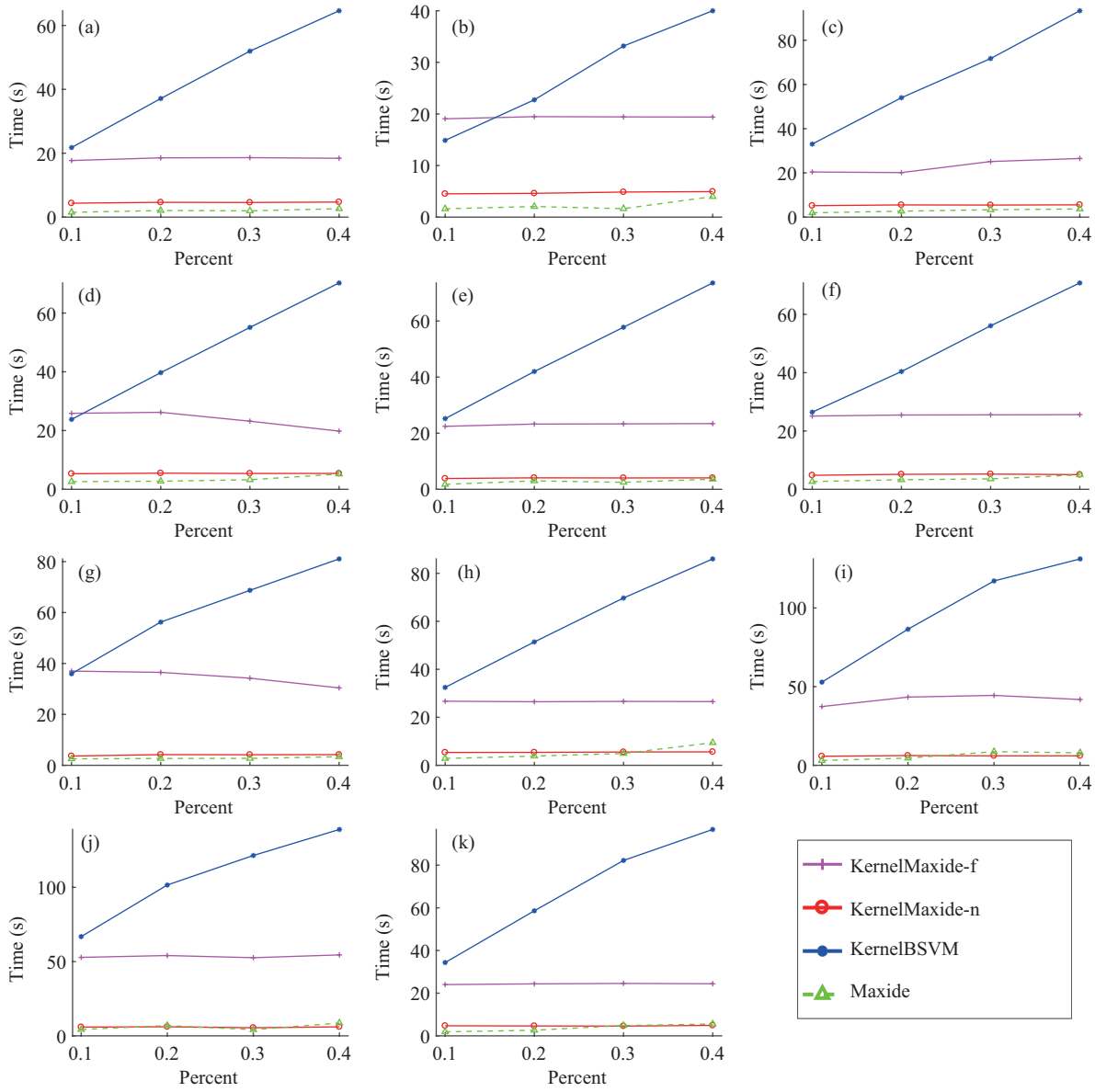

图 1 (网络版彩图) KernelMaxide 算法和其他算法在时间效率上的比较

Figure 1 (Color online) Time efficiency of the KernelMaxide algorithms compared to those of baselines. The X-axis is the observation rate (varies from 0.1 to 0.4) of the label matrix, while the Y-axis is the time cost measured in seconds. The less the running time, the higher the time efficiency. (a) Arts; (b) business; (c) computers; (d) education; (e) entertainment; (f) health; (g) recreation; (h) reference; (i) science; (j) social; (k) society

数据计算核矩阵而不做近似时, KernelMaxide-f 方法相对于 KernelBSVM 方法依然比较快, 且其运行时间并没有随着观测到元素的增多而显著增加. 这是由于 KernelMaxide 方法的计算瓶颈在于核矩阵的计算和矩阵的 SVD 分解, 这两种操作受标记矩阵中观测到元素增多的影响比较小. 注意到, 在有些实验中, 随着观测到元素的增多, KernelMaxide 的运行时间反而有轻微下降, 这可能是优化过程中收敛速度的不同造成的, 计划在未来的工作中研究这个异常现象.

5 总结

本文提出了利用辅助信息进行矩阵补全的核方法. 相对于传统的基于辅助信息进行矩阵补全的方法, 本文提出的 KernelMaxide 方法利用了数据的非线性结构提高了分类精度. 进一步利用 Nyström 方法解决了核矩阵的存储和计算开销问题. 监督信息缺失的多标记数据上的实验结果显示, 相对于传统的利用辅助信息进行矩阵补全的方法和忽略了标记之间相关性的核方法, KernelMaxide 方法都取得了较好的效果.

未来的工作将侧重于提高 KernelMaxide 算法的优化效率, 使之适合海量数据的运算. 同时, 未来的工作将研究在其他多标记学习指标上 KernelMaxide 算法的表现.

参考文献

1 Zhang M L, Zhou Z H. A review on multi-label learning algorithms. IEEE Trans Knowl Data Eng, 2014, 26: 1819–1837
2 Zhuang Y T, Han Y H, Wu F, et al. Stable multi-label boosting for image annotation with structural feature selection. Sci China Inf Sci, 2011, 54: 2508–2521
3 Nguyen V A, Boydgraber J L, Resnik P, et al. Learning a concept hierarchy from multi-labeled documents. In: Proceedings of the Advances in Neural Information Processing Systems, Montreal, 2014. 3671–3679
4 Chakrabarti D, Funiak S, Chang J, et al. Joint inference of multiple label types in large networks. In: Proceedings of the 31st International Conference on Machine Learning, Beijing, 2014. 874–882
5 Elisseeff A, Weston J. A kernel method for multi-labelled classification. In: Proceedings of the Advances in Neural Information Processing Systems, Vancouver, 2001. 681–687
6 Shao H, Li G, Liu G, et al. Symptom selection for multi-label data of inquiry diagnosis in traditional Chinese medicine. Sci China Inf Sci, 2013, 56: 052118
7 Cabral R S, de la Torre F, Costeira J P, et al. Matrix completion for weakly-supervised multi-label image classification. IEEE Trans Pattern Anal Mach Intell, 2015, 37: 121–135
8 Xu M, Jin R, Zhou Z H. Speedup matrix completion with side information: application to multi-label learning. In: Proceedings of the Advances in Neural Information Processing Systems Conference, Lake Tahoe, 2013. 2301–2309
9 Goldberg A B, Zhu X, Recht B, et al. Transduction with matrix completion: three birds with one stone. In: Proceedings of the Advances in Neural Information Processing Systems Conference, Vancouver, 2010. 757–765
10 Boutell M R, Luo J, Shen X, et al. Learning multi-label scene classification. Pattern Recog, 2004, 37: 1757–1771
11 Recht B. A simpler approach to matrix completion. J Mach Learn Res, 2011, 12: 3413–3430
12 Ji S, Sun L, Jin R, et al. Multi-label multiple kernel learning. In: Proceedings of the Advances in Neural Information Processing Systems Conference, Vancouver, 2008. 777–784
13 Tang L, Chen J, Ye J. On multiple kernel learning with multiple labels. In: Proceedings of the 18th International Joint Conference on Artificial Intelligenc, Pasadena, 2009. 1255–1260
14 Bucak S S, Jin R, Jain A K. Multi-label multiple kernel learning by stochastic approximation: application to visual object recognition. In: Proceedings of the Advances in Neural Information Processing Systems Conference, Vancouver, 2010. 325–333
15 Argyriou A, Micchelli C A, Pontil M. When is there a representer theorem? Vector versus matrix regularizers. J Mach Learn Res, 2009, 10: 2507–2529
16 Williams C K I, Seeger M W. Using the nyström method to speed up kernel machines. In: Proceedings of the Advances in Neural Information Processing Systems Conference, Denver, 2000. 682–688
17 Liu W, Tsang I W. Large margin metric learning for multi-label prediction. In: Proceedings of the 29th AAAI Conference on Artificial Intelligence, Austin, 2015. 2800–2806
18 Gentile C, Orabona F. On multilabel classification and ranking with bandit feedback. J Mach Learn Res, 2014, 15: 2451–2487
19 Bhatia K, Jain H, Kar P, et al. Sparse local embeddings for extreme multi-label classification. In: Proceedings of the Advances in Neural Information Processing Systems Conference, Montreal, 2015. 730–738

20 Zhu Y, Ting K M, Zhou Z H. Multi-label learning with emerging new labels. In: Proceedings of the IEEE International Conference on Data Mining, Barcelona, 2016. 1371–1376
21 Schapire R E, Singer Y. BoosTexter: a boosting-based system for text categorization. Mach Learn, 2000, 39: 135–168
22 Zhang M L, Zhou Z H. ML-KNN: a lazy learning approach to multi-label learning. Pattern Recog, 2007, 40: 2038–2048
23 Tsoumakas G, Katakis I, Vlahavas I P. Random k-labelsets for multilabel classification. IEEE Trans Knowl Data Eng, 2011, 23: 1079–1089
24 Bi W, Kwok J T. Bayes-optimal hierarchical multilabel classification. IEEE Trans Knowl Data Eng, 2015, 27: 2907–2918
25 Fürnkranz J, Hüllermeier E, Mencía E L, et al. Multilabel classification via calibrated label ranking. Mach Learn, 2008, 73: 133–153
26 Deng J, Russakovsky O, Krause J, et al. Scalable multi-label annotation. In: Proceedings of the CHI Conference on Human Factors in Computing Systems, Toronto, 2014. 3099–3102
27 Gao N, Huang S J, Chen S. Multi-label active learning by model guided distribution matching. Front Comput Sci, 2016, 10: 845–855
28 Yu H F, Jain P, Kar P, et al. Large-scale multi-label learning with missing labels. In: Proceedings of the 31st International Conference on Machine Learning, Beijing, 2014. 593–601
29 Schölkopf B, Smola A J. Learning With Kernels: Support Vector Machines, Cegularization, Optimization, and Beyond. Cambridge: MIT Press, 2011
30 Steinwart I, Christmann A. Support Vector Machines. New York: Springer, 2008
31 Rahimi A, Recht B. Random features for large-scale kernel machines. In: Proceedings of the Advances in Neural Information Processing Systems Conference, Vancour, 2007. 1177–1184
32 Yang T, Li Y F, Mahdavi M, et al. Nyström method vs random fourier features: a theoretical and empirical comparison. In: Proceedings of the Advances in Neural Information Processing Systems Conference, Lake Tahoe, 2012. 485–493
33 Tseng P. On Accelerated Proximal Gradient Methods for Convex-Concave Optimization. Technical Report. Seattle: University of Washington. 2008. http://www.mit.edu/ dimitrib/PTseng/papers/apgm.pdf
34 Cai J F, Candès E J, Shen Z. A singular value thresholding algorithm for matrix completion. SIAM J Optim, 2010, 20: 1956–1982
35 Ueda N, Saito K. Parametric mixture models for multi-labeled text. In: Proceedings of the Advances in Neural Information Processing Systems Conference, Vancour, 2002. 721–728
36 Fan R E, Chen P H, Lin C J. Working set selection using second order information for training support vector machines. J Mach Learn Res, 2005, 6: 1889–1918

Kernel method for matrix completion with side information and its application in multi-label learning

Miao XU[1,2] & Zhi-Hua ZHOU[1,2*]

1. *National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing* 210023*, China*;
2. *Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing* 210023*, China*
* Corresponding author. E-mail: zhouzh@lamda.nju.edu.cn

Abstract In practical machine learning, one instance is always associated with multiple labels. However, due to high cost, it is difficult to acquire the full supervised information for multi-label data. Thus, multi-label learning faces the problem of missing supervised information. By considering missing labels as unobserved entries in a matrix and features as side information, the matrix completion algorithm can be exploited to solve the missing-supervised-information problem in multi-label learning. While the previous research often focused on the case where data is linearly separable, in this paper, we propose the KernelMaxide algorithm, which not only exploits the nonlinear structure in the missing-supervised-information multi-label data, but also considers the correlation between labels. In particular, we construct a novel optimization objective based on the kernel matrix, using the Representer Theorem of Matrix Norm. We further use the Nyström method to reduce the memory and computational burden on the kernel matrix. Experiments show the merit of our proposal.

Keywords machine learning, multi-label learning, matrix completion, kernel method, Nyström method

Miao XU is a Ph.D. student in the Department of Computer Science & Technology of Nanjing University. She received the B.S. degree from Nanjing University, China, in 2009. Her main research interests include machine learning and data mining.

Zhi-Hua ZHOU was born in 1973. He received his Ph.D. degree in computer science from Nanjing University, China, in 2000. Currently, he is a professor at Nanjing University. His research interests mainly include artificial intelligence, machine learning, and data mining. He is a fellow of the ACM, AAAI, AAAS, IEEE, IAPR, IET/IEE, and CCF.