引用格式:肖雪,刘云. 嵌入式多标签分类算法的优化研究[J]. 北京化工大学学报(自然科学版),2019,46(5):94 - 100.
XIAO Xue, LIU Yun. Optimization of embedding-based multi-label classification [J]. Journal of Beijing University of Chemical Technology (Natural Science), 2019,46(5):94 - 100.

嵌入式多标签分类算法的优化研究

肖 雪 刘 云*
(昆明理工大学 信息工程与自动化学院, 昆明 650500)

摘 要: 多标签分类中如何有效处理具有许多实例和大量标签的大规模数据集、补偿训练集中缺失标签以及利用未标记实例改进预测性能等问题已成为重要研究方向。提出嵌入式多标签分类(EMC)算法,首先从伪实例参数化的高斯过程(GP)中提取两组随机变换来模拟特征向量、潜在空间表示向量和标签向量之间的非线性关系映射,其次引入一组辅助变量结合专家集成(EEOE)方法补偿缺失标签,最后利用未标记实例学习随机函数的平滑映射提高预测性能。仿真结果表明,与特征识别隐式标签空间编码的多标签分类(FaLE)算法和半监督低秩映射多标签分类(SLRM)算法相比,EMC 算法优化了处理大规模数据集、补偿缺失标签及利用未标记数据的能力,从而提高了类标签的预测性能,且具有良好的可扩展性,训练时间短。

关键词: 多标签分类; 缺失标签; 嵌入式; 可扩展性
中图分类号: TP311 **DOI**: 10.13543/j.bhxbzr.2019.05.014

引 言

多标签分类不同于传统的单标签分类,通常每个实例都有一组标签分配[1-2],处理起来也更为复杂,其面临的主要问题包括处理具有许多实例和大量标签的大规模数据集、有效补偿训练集中缺失的标签分配以及利用未标记实例来改进预测性能等[3-4]。许多先进的嵌入式方法通过线性降维将标签分配映射到低维空间(潜在空间),然后在输入特征上使用独立回归模型预测潜在空间中实例的表示,可以有效处理大规模数据集[5-6]。多标签数据集中,标签之间的语义模糊性容易造成标签矩阵缺失部分数据,导致训练集不能完全提供所有有效的标签分配,此问题即为标签缺失。在标签缺失的情况下,若直接利用缺失标签的数据进行训练,则训练过程很难保证预测结果的准确性[6-7]。因此在使用缺失标签的数据之前,利用特定的策略对缺失标签进行恢复是非常有必要的。

Lin 等[8]提出通过特征识别隐式标签空间编码的多标签分类算法(multi-label classification via feature-aware implicit label space encoding, FaLE),该算法考虑了低维潜在空间表示和标签向量之间的线性关系,通过特征识别隐式标签空间编码来实现标签空间降维。FaLE 算法虽然能处理大规模数据集,但忽略了尾标签(不经常分配给实例的标签),而在许多实际应用场景中都有数千个尾标签,丢弃这些尾标签会直接影响预测性能。Jing 等[9]提出半监督低秩映射多标签分类(semi-supervised low-rank mapping learning for multi-label classification, SLRM)算法,该算法利用标签相关性补偿缺失标签,通过映射的核范数正则化有效获取标签相关性,同时在映射中引入多个正则化器来获取数据之间的内在结构,可有效补偿缺失标签。

为了有效补偿缺失标签和利用未标记实例来提升预测性能,本文提出嵌入式多标签分类(embedding-based multi-label classification,EMC)算法,首先使用高斯过程(GP)提取两组随机函数对特征向量、潜在空间表示向量和标签向量之间的非线性关系进行建模,再引入一组辅助变量结合专家集成(EEOE)方法[10]补偿缺失标签,最后利用未标记实例学习随机函数的平滑映射以提高预测性能。为了避免高斯过程的高计算成本和内存要求,通过伪实例对

收稿日期: 2019-05-05
基金项目: 国家自然科学基金(61761025)
第一作者: 女,1994 年生,硕士生
*通信联系人
E-mail: liuyun@kmust.edu.cn

随机函数进行参数化，从而能有效处理具有大量实例的大规模数据集。

1　模型建立

算法执行过程的具体步骤如图 1 ~ 4 所示。$\boldsymbol{X}$、$\boldsymbol{C}$ 和 $\boldsymbol{Y}$ 分别表示特征向量、潜在空间的向量和标签向量。嵌入式方法在低维空间（潜在空间）中表示每个实例的标签向量，构建独立回归模型从向量 $\boldsymbol{X}$ 中预测向量 $\boldsymbol{C}$ 和 $\boldsymbol{Y}$。

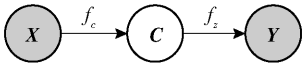

图 1　处理尾标签

Fig. 1　Handling tail labels

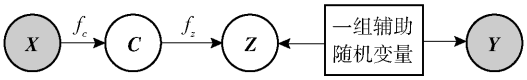

图 2　补偿缺失标签

Fig. 2　Handling missing labels

本文模型使用高斯过程生成随机函数 f_c 和 f_z 来模拟向量之间的非线性关系，即向量 $\boldsymbol{X}$ 映射到向量 $\boldsymbol{C}$ 的第 ℓ 维，记作 $\boldsymbol{c}_\ell \approx f_c^{(\ell)}(\boldsymbol{X})$；同理，向量 $\boldsymbol{C}$ 映射到向量 $\boldsymbol{Y}$ 的第 k 维，记作 $\boldsymbol{y}_k \approx f_z^{(k)}(\boldsymbol{C})$，其中 $f_c^{(\ell)}(\cdot)$，$f_z^{(k)}(\cdot)$ 是随机函数。这种方法考虑了向量 $\boldsymbol{C}$ 和 $\boldsymbol{Y}$ 之间的非线性关系，能够在训练阶段解决尾标签被忽略的问题。

修改图 1，补偿缺失标签。由于训练集不提供某些标签分配，因此每个实例都有两个标签向量：①观察到的不完整标签向量 $\boldsymbol{Y} \in \{0,1\}^K$；②未观察到的完整标签向量 $\boldsymbol{Z} \in \mathbb{R}^K$。完整的标签向量 $\boldsymbol{Z}$ 是二进制值，本文提出的概率模型中假设向量 $\boldsymbol{Z}$ 是实值，向量 $\boldsymbol{Z}$ 的第 k 维表示该实例的第 k 个标签适应性。通过 EEOE 方法模拟向量 $\boldsymbol{Z}$ 和 $\boldsymbol{Y}$ 之间的关系。

EMC 算法类似于 SLRM 算法[9]，利用未标记实例学习“平滑”映射的随机函数 f_c 和 f_z，将向量 $\boldsymbol{X}$ 转换到向量 $\boldsymbol{Z}$。图 3 中 $\boldsymbol{X}_u$，$\boldsymbol{C}_u$ 和 $\boldsymbol{Z}_u$ 表示特征向量、嵌入的标签向量和未标记实例的完整标签向量。所提算法若直接使用图 3 模型会使均值场变分推断难以处理，因此引入随机变量 $\hat{\boldsymbol{C}}$ 和 $\hat{\boldsymbol{Z}}$，如图 4 所示，同时使用高斯过程隐变量模型（GP－LVM）[11] 中的一个证据下限。向量 $\boldsymbol{C}$ 和 $\hat{\boldsymbol{C}}$（向量 $\boldsymbol{Z}$ 和 $\hat{\boldsymbol{Z}}$）近似相同。

根据图 4，使用随机函数 $f_c^{(l)}(\cdot)$ 模拟向量 $\boldsymbol{X}$ 和向量 $\hat{\boldsymbol{C}}$ 的第 ℓ 维（即 $\hat{\boldsymbol{c}}_\ell$）之间的关系，假设 $\hat{\boldsymbol{c}}_\ell \approx f_c^{(\ell)}(\boldsymbol{X})$，则

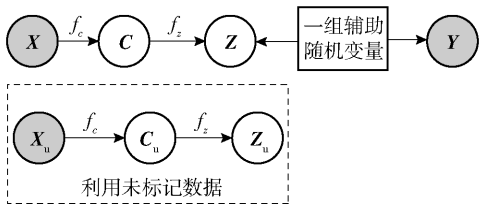

图 3　利用未标记实例

Fig. 3　Exploiting unlabeled instances

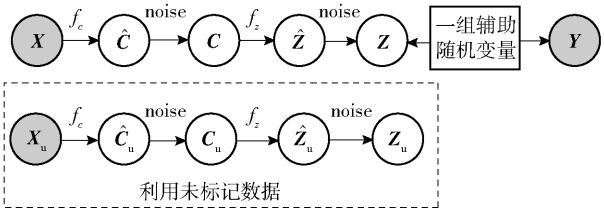

图 4　近似推断

Fig. 4　Approximate inference

$$f_c^{(\ell)}(\cdot) \sim GP(0, \kappa(\cdot,\cdot;\sigma_c))$$
$$\hat{\boldsymbol{c}}_\ell \sim N(\hat{\boldsymbol{c}}_\ell | f_c^{(\ell)}(\boldsymbol{X}), \alpha_c^2) \tag{1}$$

假设有 M 个伪实例（将 M 设置为较小的值），需要计算 $M \times M$ 矩阵的逆，通过一些伪实例对 f_c 和 f_z 中的函数进行参数化，可避免高斯过程[12] 的高计算成本和内存要求，则可以处理具有大量实例的大规模数据集。

针对以上步骤（图 1 ~ 4）建模如图 5 所示。

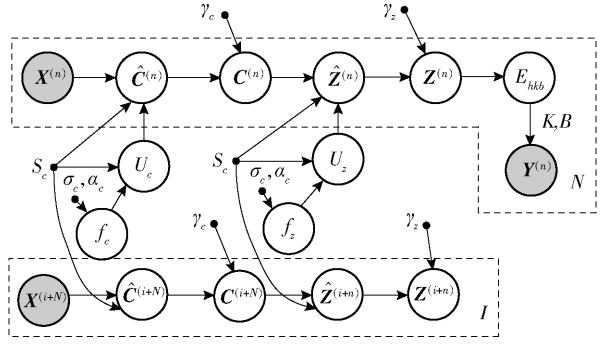

图 5　EMC 算法模型

Fig. 5　EMC algorithm model

2　EMC 算法

2.1　补偿缺失标签

在 EEOE 方法中引入一组辅助随机变量（称为专家 B）处理缺失标签。辅助变量对应于图 5 中变量 $\{E_{nkb}\}_{k,b}$，给定 $\boldsymbol{Z}^{(n)}$ 的第 k 维，生成观察到的标签向量的第 k 维（$\boldsymbol{y}_k^{(n)}$）如下。

命题 1　若 $1 \leqslant b \leqslant B$：$E_{nkb} \sim Bernoulli(\sigma(\lambda z_k^{(n)}))$，$\lambda \in R$ 是常数，可得

$$\boldsymbol{y}_k^{(n)} = E_{nk1}\frac{\sum_{b=1}^{B} E_{nkb}}{B} \tag{2}$$

2.2 利用未标记实例及处理大规模数据

为利用未标记实例改善预测性能,图 3 中未标记实例通过学习平滑映射函数 f_c 和 f_z,实现特征向量 $\boldsymbol{X}_u$ 到潜在空间向量 $\boldsymbol{C}_u$、潜在空间向量到完整的未观察到的标签向量 $\boldsymbol{Z}_u$ 的非线性映射。图 5 模型中 I 部分对应于图 3 的虚线框,通过引入变量 $\boldsymbol{X}_u$,$\boldsymbol{C}_u$ 和 $\boldsymbol{Z}_u$,所提出模型学习随机函数 f_c 和 f_z 具有平滑性,即如果向量 $\boldsymbol{X}^{j_1}$ 和 $\boldsymbol{X}^{j_2}$ 彼此接近,则向量 $\boldsymbol{C}^{j_1}$ 和 $\boldsymbol{C}^{j_2}$,及向量 $\boldsymbol{Z}^{j_1}$ 和 $\boldsymbol{Z}^{j_2}$ 最有可能彼此接近。

为了避免图 3 中传统高斯过程的高计算和空间成本,使用伪实例将随机函数 f_c、f_z 参数化。

命题 2 若

$$1 \leqslant \ell \leqslant L: f_c^{(\ell)}(\cdot) \sim GP(0, \kappa(\cdot,\cdot;\sigma_c)) \tag{3}$$

从函数 $f_c = \{f_c^{(\ell)}(\cdot)\}_{\ell=1}^{L}$ 中生成伪类观察值 $\{(S_c^{(m)}, u_{cm}^{(\ell)})\}_{m=1}^{M_c}$,且变量 $\{S_c^{(m)}\}_{m=1}^{M_c}$ 被视为模型参数(而不是随机变量),生成变量 $\{u_{cm}^{(\ell)}\}_{m=1}^{M_c}$。

命题 3 若 $1 \leqslant \ell \leqslant L, 1 \leqslant m \leqslant M_c$:

$$u_{cm}^{(\ell)} \sim N(f_c^{(\ell)}(s_c^{(m)}), \alpha_c^2) \tag{4}$$

伪实例 M_c 的观察值 $\{(S_c^{(m)}, u_{cm}^{(\ell)})\}_{m=1}^{M_c}$,后验概率 $p(f_c^{(l)}(\cdot) \mid \{(S_c^{(m)}, u_{cm}^{(l)})\}_{m=1}^{M_c})$ 将是具有以下平均函数的高斯过程[13]。

$$\boldsymbol{\mu}_c^{(\ell)}(\boldsymbol{X}^*) = \kappa(\boldsymbol{X}^*, S_c)[\kappa(S_c, S_c) + \alpha^2 I_{M_c \times M_c}]^{-1} \times [\boldsymbol{\mu}_{c1}^{(\ell)}, \cdots, \boldsymbol{\mu}_{cM_c}^{(\ell)}]^{\mathrm{T}} \tag{5}$$

式中,$S_c = \{S_c^{(m)}\}_{m=1}^{M_c}$,$\kappa(\cdot,\cdot)$ 是核函数。

当给定 $\{(S_c^{(m)}, u_{cm}^{(\ell)})\}_{m=1}^{M_c}$ 和 $\{\boldsymbol{X}^{(n)}\}_{n=1}^{N+I}$,生成潜在空间表示 $\hat{\boldsymbol{C}}$。

命题 4 $1 \leqslant n \leqslant (N+I), 1 \leqslant \ell \leqslant L$:

$$\hat{\boldsymbol{c}}_{\ell}^{(n)} \sim N(\kappa(X^{(n)}, S_c)[\kappa(S_c, S_c) + \alpha^2 I_{M_c \times M_c}]^{-1} \times [\boldsymbol{\mu}_{c1}^{(\ell)}, \cdots, \boldsymbol{\mu}_{cM_c}^{(\ell)}]^{\mathrm{T}}, \beta_c^2) \tag{6}$$

式(6)证明了通过特征向量 $\boldsymbol{X}^{(n)}$ 和伪实例生成 $\hat{\boldsymbol{C}}^{(n)}$ 的 ℓ 维,本文不直接使用函数 $f_c^{(l)}(\cdot)$,可以直观地看出 $\hat{\boldsymbol{c}}_{\ell}^{(n)} \approx f_c^{(\ell)}(\boldsymbol{X}^{(n)})$。

通过伪样本参数化函数 $f_z = \{f_z^{(k)}(\cdot)\}_{k=1}^{K}$,生成变量 $f_z^{(k)}(\cdot)$、$u_{zm}^{(k)}$、$\hat{\boldsymbol{Z}}$ 的过程与生成变量 $f_c^{(l)}(\cdot)$、$u_{cm}^{(l)}$、$\hat{\boldsymbol{C}}$ 的过程非常相似,可同理得出。

命题 5 若 $1 \leqslant k \leqslant K$:

$$f_z^{(k)}(\cdot) \sim GP(0, \kappa(\cdot,\cdot;\sigma_z)) \tag{7}$$

命题 6 若 $1 \leqslant k \leqslant K, 1 \leqslant m \leqslant M_z$:

$$u_{zm}^{(k)} \sim N(f_z^{(k)}(s_z^{(k)}), \alpha_z^2) \tag{8}$$

利用 $\{(S_z^{(k)}, u_{zm}^{(k)})\}_{m=1}^{M_z}$ 和 $\boldsymbol{C}^{(n)}$ 生成 $\hat{\boldsymbol{Z}}^{(n)}$ 的第 k 维。

命题 7 若 $1 \leqslant n \leqslant (N+I), 1 \leqslant k \leqslant K$:

$$\hat{z}_k^{(n)} \sim N(\kappa(C^{(n)}, S_z)[\kappa(S_z, S_z) + \alpha_z^2 I_{M_z \times M_z}]^{-1} \times [\boldsymbol{\mu}_{z1}^{(k)}, \cdots, \boldsymbol{\mu}_{zM_z}^{(k)}]^{\mathrm{T}}, \beta_z^2) \tag{9}$$

利用伪实例参数化函数 $f_c = \{f_c^{(\ell)}(\cdot)\}_{\ell=1}^{L}$ 和 $f_z = \{f_z^{(k)}(\cdot)\}_{k=1}^{K}$,所提算法可以处理具有大量实例的大规模数据集。利用本文算法计算 $M_c \times M_c$ 的逆矩阵(式(6))和 $M_z \times M_z$ 的逆矩阵(式(9))需要的计算时间分别为 $O(M_c^3)$ 和 $O(M_z^3)$,内存分别为 $O(M_c^2)$ 和 $O(M_z^2)$。

2.3 近似推断

图 5 中,若直接连接节点 $\boldsymbol{X}^{(n)}$ 到节点 $\boldsymbol{C}^{(n)}$,节点 $\boldsymbol{C}^{(n)}$ 到节点 $\boldsymbol{Z}^{(n)}$,会使 $P(\boldsymbol{C} \mid \boldsymbol{X}, f_c) \times P(\boldsymbol{Z} \mid \boldsymbol{C}, f_z)$ 项在联合分布中的高斯均值场变分推断难以处理。图 4、5 中给定特征向量 $\boldsymbol{X}^{(n)}$ 和函数 f_c,首先生成潜在空间表示(即向量 $\hat{\boldsymbol{C}}^{(n)}$),随后生成向量 $\boldsymbol{C}^{(n)}$。实际上,$\boldsymbol{C}^{(n)}$ 和 $\hat{\boldsymbol{C}}^{(n)}$ 具有相同的作用,都是第 n 个实例的潜在空间表示向量,同理,$\boldsymbol{Z}^{(n)}$ 和 $\hat{\boldsymbol{Z}}^{(n)}$ 都是第 n 个实例的完整标签向量。证明如下。

如图 3 所示,$P(\boldsymbol{C}^{(n)} \mid the\ rest) \propto P(\boldsymbol{C}^{(n)} \mid \boldsymbol{X}^{(n)}, U_c) \times P(\boldsymbol{Z}^{(n)} \mid \boldsymbol{C}^{(n)}, U_c)$ 中 $P(\boldsymbol{C}^{(n)} \mid the\ rest)$ 包含 $\exp(\kappa(\boldsymbol{C}^{(n)}, S_c))$(式(6))和 $\exp(\|\boldsymbol{C}^{(n)}\|^2)$(式(9)),忽略变量 $\hat{\boldsymbol{C}}$ 和 $\hat{\boldsymbol{Z}}$,高斯均值场变分推断难以实现,在指数族中计算 $\boldsymbol{C}^{(n)}$ 的条件分布十分困难。改进为图 4,引入随机变量 $\hat{\boldsymbol{C}}$ 和 $\hat{\boldsymbol{Z}}$ 并使用结构化变分推理,采用 GP-LVM 中调整证据下限值的方法。

由于本文引入辅助随机变量,所以对变分分布作出假设,以使这个下限适用于本文的概率模型,假设如下。

假设 1 变分布 q 可以根据贝叶斯网络分解为式(10)所示的分布族

$$q = \left[\prod_{\ell=1}^{L} q(f_c^{(\ell)})\right]\left[\prod_{k=1}^{K} q(f_z^{(k)})\right]\left[\prod_{\ell=1}^{L} q([u_{c1}^{(\ell)}, \cdots, u_{cM_c}^{(\ell)}])\right]\left[\prod_{k=1}^{K} q([u_{z1}^{(k)}, \cdots, u_{zM_z}^{(k)}])\right]\left[\prod_{n=1}^{N+I} q(\hat{\boldsymbol{C}}^{(n)}) \mid U_c, \boldsymbol{X}^{(n)}\right]\left[\prod_{n=1}^{N+I} q(\hat{\boldsymbol{Z}}^{(n)}) \mid \boldsymbol{C}^{(n)}, U_z\right]\left[\prod_{n=1}^{N+I} q(\boldsymbol{Z}^{(n)})\right]\left[\prod_{n=1}^{N}\prod_{k=1}^{K}\prod_{b=1}^{B} q(E_{nkb})\right] \tag{10}$$

假设 2 以下等式成立

$$q(\hat{\boldsymbol{C}}^{(n)} \mid U_c, \boldsymbol{X}^{(n)}) = P(\hat{\boldsymbol{C}}^{(n)} \mid U_c, \boldsymbol{X}^{(n)}) \tag{11}$$

$$q(\hat{\boldsymbol{Z}}^{(n)} \mid U_z, \boldsymbol{C}^{(n)}) = P(\hat{\boldsymbol{Z}}^{(n)} \mid U_z, \boldsymbol{C}^{(n)}) \tag{12}$$

式(11)和(12)右边分别对应于式(6)和(9)中的条件分布，$L(q)$作为 GP-LVM 的下限值。

定义 1

$$L(q) \triangleq \int q\ln\left(\frac{P}{q}\right)\mathrm{d}l \tag{13}$$

式中，l 表示潜在空间变量。

通过式(11)和(12)，可以忽略$\frac{P}{q}$中的 $P(\hat{\boldsymbol{C}}^{(n)} \mid U_c, \boldsymbol{X}^{(n)})$和 $P(\hat{\boldsymbol{Z}}^{(n)} \mid U_z, \boldsymbol{C}^{(n)})$以及分母中的 $q(\hat{\boldsymbol{C}}^{(n)} \mid U_c, \boldsymbol{X}^{(n)})$和 $q(\hat{\boldsymbol{Z}}^{(n)} \mid U_z, \boldsymbol{C}^{(n)})$，本文模型生成过程中用函数 $G(t,\xi)$来近似函数 $\sigma(t)$[14]。

定义 2

$$G(t,\xi) \triangleq \sigma(\xi)\exp\left\{\frac{t-\xi}{2} - \frac{\tanh\left(\frac{\xi}{2}\right)}{4\xi}(t^2-\xi^2)\right\} \tag{14}$$

近似计算

$$p(E_{nkb} \mid z_k^{(n)}) = \sigma(\lambda z_k^{(n)})^{E_{nkb}}(1-\sigma(\lambda z_k^{(n)}))^{1-E_{nkb}} = \left(\frac{\sigma(\lambda z_k^{(n)})}{1-\sigma(\lambda z_k^{(n)})}\right)^{E_{nkb}}(1-\sigma(\lambda z_k^{(n)})) = \exp(E_{nkb}\lambda z_k^{(n)})\sigma(-\lambda z_k^{(n)}) \approx \exp(E_{nkb}\lambda z_k^{(n)})G(-\lambda z_k^{(n)}, \xi_{nk}) \tag{15}$$

式中，$\{\xi_{nk}\}_{n=1k=1}^{N\ K}$是辅助变分参数。当$\frac{\partial L(q)}{\partial \xi_{nk}} = 0$时，得 $\xi_{nk} = \pm\lambda E_{Z^{(n)}\sim q(Z^{(n)})}[z_k^{(n)}]$。

如果 $\boldsymbol{y}_k^{(n)} = 1$，设置 $q(\boldsymbol{Z}^{(n)})$的 k 维均值变分分布，设置伪实例 $\{S_c^{(m)}\}_{m=1}^{M_c}$为特征向量的子集(即 $\{\boldsymbol{X}^{(n)}\}_{n=1}^{N+I}$)，随机选择一些特征向量设置伪实例 $\{S_c^{(m)}\}_{m=1}^{M_c}$，在变分推理的每次迭代中更新伪实例 $\{S_z^{(m)}\}_{m=1}^{M_z}$为

$$S_z^{(m)} \leftarrow E[f_c(S_c^{(m)})] \tag{16}$$

通过以上近似推断证明并使用结构化变分推断引入变量 $\hat{\boldsymbol{C}}$ 和 $\hat{\boldsymbol{Z}}$，最后生成向量 $\boldsymbol{C}$ 和 $\boldsymbol{Z}$。

若 $1 \leqslant n \leqslant (N+I)$，则

$$\boldsymbol{C}^{(n)} \sim N(\hat{\boldsymbol{C}}^{(n)}, \gamma_c^2 I_{L\times L}) \tag{17}$$

$$\boldsymbol{Z}^{(n)} \sim N(\hat{\boldsymbol{Z}}^{(n)}, \gamma_z^2 I_{K\times K}) \tag{18}$$

EMC 算法伪代码如下。

输入：X

输出：Y

(1)初始化 N, I, F, K, L；

(2)引入 B 和 $\{E_{nkb}\}_{k,b} \longrightarrow \boldsymbol{y}_k^{(n)}$；

(3)伪实例参数化随机函数 f_c，计算出 $\hat{\boldsymbol{c}}_{\prime}^{(n)} \approx f_c^{(\prime)}(\boldsymbol{X}^{(n)})$；

(4)同理，伪实例参数化随机函数 f_z，计算 $\hat{\boldsymbol{z}}_k^{(n)}$；

(5)生成潜在空间 $\hat{\boldsymbol{C}}^{(n)}$，利用结构化变分推理生成 $\boldsymbol{C}^{(n)} \sim N(\hat{\boldsymbol{C}}^{(n)}, \gamma_c^2 I_{L\times L})$；

(6)生成第 n 个实例的完整标签向量 $\hat{\boldsymbol{Z}}^{(n)}$，再生成 $\boldsymbol{Z}^{(n)} \sim N(\hat{\boldsymbol{Z}}^{(n)}, \gamma_z^2 I_{K\times K})$。

3　仿真分析与结果

3.1　数据集和参数设置

为了验证所提算法，选取 Mulan Library[15] 中的 CAL500 和 NUS-WIDE 多标签数据集进行仿真，其中 NUS-WIDE 是大规模数据集，用来进行可扩展性分析。表 1 为仿真数据集详细信息。

表 1　仿真数据集

Table 1　The simulation dataset

数据集	N	F	K	测试集个数	每个样本平均标签个数
CAL500	400	68	174	102	26.044
NUS-WIDE	161 789	500	81	107 859	1.865 5

仿真时，FaLE 算法中的参数 α_{FaLE} 从集合 $\{10^{-1}, 10^0, \cdots, 10^4\}$ 中选择，SLRM 算法中的参数 λ_{SLRM}和 γ_{SLRM}从集合 $\{10^{-3}, 10^0, \cdots, 10^3\}$ 中选择，本文 EMC 算法中设置的内核参数 σ_c 是 FaLE 中特征向量间平均欧氏距离的两倍，参数 λ、ρ 和 σ_z 从集合 $\{10^1, \cdots, 10^5\}$ 中选择，参数 $B = \min\left\{\frac{E_0}{E_1}, 50\right\}$，$E_0$ 表示矩阵 $\boldsymbol{Y}$ 中值为 0 的元素个数，E_1 表示矩阵 $\boldsymbol{Y}$ 中值为 1 的元素个数，伪实例数量 $M_c = M_z = 500$。

仿真指标使用 3 个基于秩的评估度量，即 ROC 下面积曲线(AUC)[16]，覆盖率[16]和精度@ k[17]，这些评估指标广泛用于多标签分类。

3.2　仿真结果分析

3.2.1　补偿缺失标签的性能

将实例随机分为训练数据和测试数据，在训练集中随机移除标签分配(10% ~ 50%)。图 6、7 中横坐标表示标签移除率，移除率为 0 时表现为 FaLE、SLRM 和 EMC 算法在对应数据集中的分类性能。由图可以看出，EMC 算法补偿缺失标签的能力明显优于 FaLE 算法和 SLRM 算法，表明预测性能得到提升。

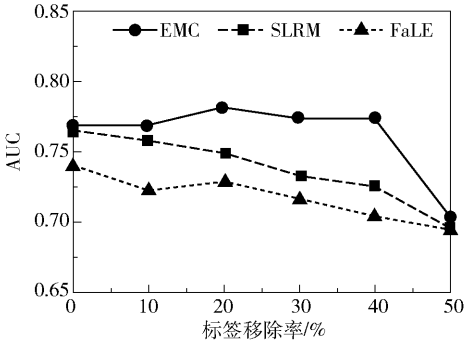

图 6 CAL500 数据集中不同算法的 AUC

Fig. 6 AUC of different algorithms in CAL500

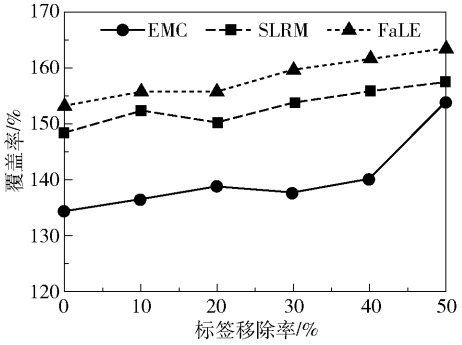

图 7 CAL500 数据集中不同算法的覆盖率

Fig. 7 Coverage of different algorithms in CAL500

3.2.2 利用未标记实例的性能

选择一部分(10% ~20%)标签向量用于训练阶段,将实例随机分为训练数据和测试数据。图 8、9 中横坐标表示训练集中使用的标签向量分数。在此设置中,随机选择一些训练数据并仅使用这些实例的标签向量(其他实例可视为未标记数据),选择矩阵 $\boldsymbol{Y}$ 中一部分元素并设置为 0 以评估补偿缺失标签的能力。由图 8、9 可以看出,EMC 算法利用未标记实例的能力明显优于 FaLE 算法和 SLRM 算法,从而提升了预测性能。

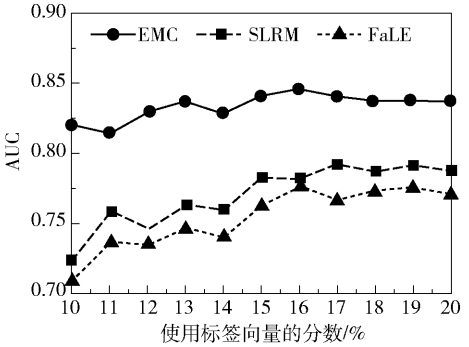

图 8 CAL500 数据集中不同算法的 AUC

Fig. 8 AUC of different algorithms in CAL500

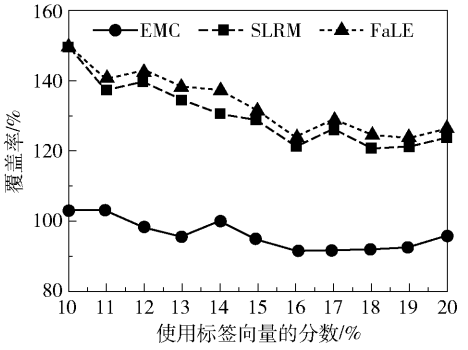

图 9 CAL500 数据集中不同算法的覆盖率

Fig. 9 Coverage of different algorithms in CAL500

3.2.3 处理尾标签的性能

对于每个数据集,根据标签频率(即分配标签的实例数量)对标签进行分类,对每个标签计算 $\sum_{j=1}^{k} R_k@50$,其中 $R_k@n$ 定义为

$$R_k@n = \frac{TP(n,j)}{f(j)} \tag{19}$$

式中,$TP(n,j)$ 为当分配第 j 个标签给预测矩阵时,排名靠前的 n 个最相关实例的真阳性预测数,$f(j)$ 为第 j 个标签已分配的实例数,$f(j) < f(j+1)$。由图 10、11可知,对于250个不频繁标签,FaLE、SLRM

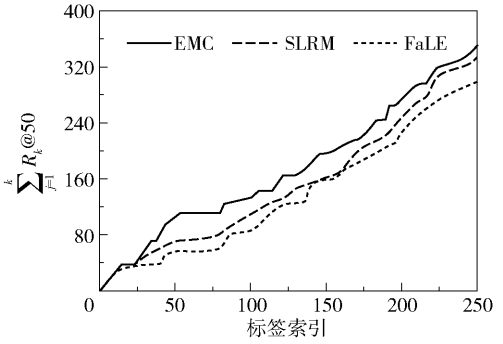

图 10 CAL500 数据集中预测性能对比

Fig. 10 Predicting performance in CAL500

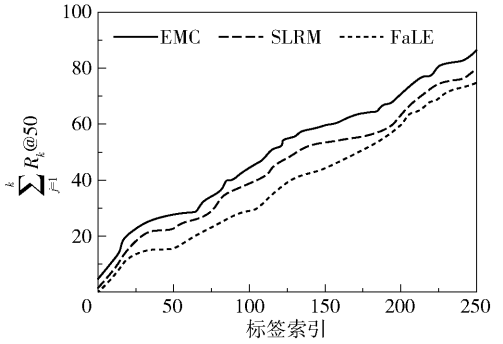

图 11 NUS-WIDE 数据集中不同算法的预测性能

Fig. 11 Predicting performance of different algorithms in NUS-WIDE

算法的预测性能比 EMC 算法差。

3.2.4　可扩展性

为了评估 EMC 算法处理大规模数据集的能力，选用大型数据集 NUS-WIDE 进行仿真，选取精度 $P@1$、精度 $P@3$ 和训练时间指标进行对比，其中精度 $P@k$ 表示 Top-K 的精度[18]。

从表 2 可看出，EMC 算法的可扩展性优于 FaLE、SLRM 算法。

表 2　NUS-WIDE 数据集中不同算法性能对比

Table 2　Performance comparison of different algorithms in NUS-WIDE

算法	$P@1$/%	$P@3$/%	训练时间/s
EMC	40.93	28.52	6 650
SLRM	31.36	20.45	1 202
FaLE	20.76	16.04	1 681

4　结论

本文提出了一种嵌入式多标签分类算法(EMC)，用伪实例对高斯过程(GP)提取的随机函数进行参数化，可以有效处理具有大量实例的大规模数据集，避免了高的计算成本和内存需求；引入辅助变量结合 EEOE 方法可以有效补偿缺失标签，利用未标记实例学习随机函数的平滑映射可提高预测性能。仿真结果表明，与 FaLE 算法和 SLRM 算法相比，所提 EMC 算法优化了处理大规模数据集、补偿缺失标签及利用未标记数据等能力，从而提高了类标签的预测性能，且具有良好的可扩展性，训练时间短。

符　号　说　明

N—训练集中被标记实例数

I—未标记实例数

F—特征数

K—标签集的基数

L—潜在空间的维数($L \ll K$)

$\boldsymbol{X}^{(n)} \in \mathbb{R}^F$—训练集中第 n 个标记实例的特征向量

$\boldsymbol{X}^{(i+N)} \in \mathbb{R}^F$—训练集中第 i 个未标记实例的特征向量

$\boldsymbol{Y}^{(n)} \in \{0,1\}^K$—第 n 个实例的标签向量

$\boldsymbol{C}^{(n)} \in \mathbb{R}^L$—第 n 个标签向量 $\boldsymbol{Y}^{(n)}$ 的嵌入表示向量

$\boldsymbol{C}^{(i+N)} \in \mathbb{R}^L$—第 i 个未标记实例未观察到的标签向量的潜在空间表示

$\boldsymbol{Z}^{(n)} \in \mathbb{R}^K$—向量 $\boldsymbol{Z}^{(n)}$ 第 k 维表示第 k 个标签对第 n 个标记实例的适应性

$\boldsymbol{Z}^{(i+N)} \in \mathbb{R}^K$—向量 $\boldsymbol{Z}^{(i)}$ 第 k 维表示第 k 个标签对第 i 个未标记实例的适应性

$\boldsymbol{c}_{\ell}^{(n)} \in \mathbb{R}$—向量 $\boldsymbol{c}^{(n)}$ 第 ℓ 个元素

$\lambda \in \mathbb{R}$—调整平滑函数斜率的一个常数，用 $\sigma(\lambda \cdot)$ 代替 $\sigma(\cdot)$

$f_c = \{f_c^{(\ell)}\}_{\ell=1}^{L}$—$f_c^{(\ell)}: \mathbb{R}^F \to \mathbb{R}$ 是随机函数，将向量 $\{\boldsymbol{X}^{(n)}\}_{n=1}^{N+I}$ 映射到向量 $\{\boldsymbol{C}^{(n)}\}_{n=1}^{N+I}$ 的 ℓ 维

$f_z = \{f_z^{(k)}\}_{k=1}^{K}$—$f_z^{(k)}: \mathbb{R}^L \to \mathbb{R}$ 是随机函数，将向量 $\{\boldsymbol{C}^{(n)}\}_{n=1}^{N+I}$ 映射到向量 $\{\boldsymbol{Y}^{(n)}\}_{n=1}^{N+I}$ 的 k 维

M_c, M_z—f_c, f_z 中随机函数的伪实例数

$S_c = \{S_c^{(m)}\}_{m=1}^{M_c}$—$f_c$ 函数的伪样本集，其中 $S_c^{(m)} \in \mathbb{R}^F$，$1 \leqslant m \leqslant M_c$

$S_z = \{S_z^{(m)}\}_{m=1}^{M_z}$—$f_z$ 函数的伪样本集，其中 $S_z^{(m)} \in \mathbb{R}^L$，$1 \leqslant m \leqslant M_z$

$U_c = \{u_{cm}^{(\ell)}\}_{m=1,\ell=1}^{M_c\ L}$—$f_c$ 函数中伪实例值的集合，$u_{cm}^{(\ell)} \approx f_c^{(\ell)}(s_c^{(m)})$

$U_z = \{u_{zm}^{(k)}\}_{m=1,k=1}^{M_z\ K}$—$f_z$ 函数中伪实例值的集合，$u_{zm}^{(k)} \approx f_z^{(k)}(s_z^{(m)})$

$\kappa(\cdot,\cdot;\sigma)$—具有平滑参数 σ 的径向基核函数

参考文献：

[1] SUN Z W, HU K Y, HU T, et al. Fast multi-label low-rank linearized SVM classification algorithm based on approximate extreme points[J]. IEEE Access, 2018, 6: 42319-42326.

[2] 李锋, 杨有龙. 基于标签特征和相关性的多标签分类算法[J]. 计算机工程与应用, 2019, 55(4): 48-55.
LI F, YANG Y L. Multi-label classification algorithm based on label-specific features and label correlation[J]. Computer Engineering and Applications, 2019, 55(4): 48-55. (in Chinese)

[3] WU Q Y, TAN M K, SONG H J, et al. ML-FOREST: a multi-label tree ensemble method for multi-label classification[J]. IEEE Transactions on Knowledge and Data Engineering, 2016, 28(10): 2665-2680.

[4] NAZMI S, RAZEGHI-JAHROMI M, HOMAIFAR A. Multi-label classification with weighted labels using learning classifier systems[C]//2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA). Cancun, 2017: 275-280.

[5] 孙圣姿, 万源, 曾成. 自适应嵌入的半监督多视角特征降维方法[J]. 计算机应用, 2018, 38(12): 3391-3398.
SUN S Z, WAN Y, ZENG C. Semi-supervised adaptive multi-view embedding method for feature dimension reduction[J]. Journal of Computer Applications, 2018, 38(12): 3391-3398. (in Chinese)

[6] LI X, SHEN B, LIU B D, et al. Ranking-preserving low-

rank factorization for image annotation with missing labels [J]. IEEE Transactions on Multimedia, 2018, 20(5): 1169-1178.

[7] HUANG J, QIN F, ZHENG X, et al. Learning label-specific features for multi-label classification with missing labels[C]//2018 IEEE Fourth International Conference on Multimedia Big Data (BigMM). Xi'an, 2018: 1-5.

[8] LIN Z J, DING G G, HU M Q, et al. Multi-label classification via feature-aware implicit label space encoding [C]//Proceedings of the 31st International Conference on Machine Learning. Beijing, 2014: 325-333.

[9] JING L P, YANG L, YU J, et al. Semi-supervised low-rank mapping learning for multi-label classification[C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, 2015: 1483-1491.

[10] AKBARNEJAD A, BAGHSHAH M S. A probabilistic multi-label classifier with missing and noisy labels handling capability[J]. Pattern Recognition Letters, 2017, 89: 18-24.

[11] KO J, FOX D. Learning GP-Bayes filters via Gaussian process latent variable models[J]. Autonomous Robots, 2011, 30(1): 3-23.

[12] CHEN S Q, AMMAR H B, TUYLS K, et al. Optimizing complex automated negotiation using sparse pseudo-input Gaussian processes [C] // International Conference on Autonomous Agents & Multi-agent Systems. Taipei, 2013.

[13] RASMUSSEN C E, NICKISCH H. Gaussian processes for machine learning (GPML) toolbox[J]. Journal of Machine Learning Research, 2010, 11: 3011-3015.

[14] BISHOP C M, SVENSÉN M. Bayesian hierarchical mixtures of experts[C]//Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence. Acapulco, 2012: 57-64.

[15] TSOUMAKAS G, SPYROMITROS-XIOUFIS E, VILCEK J, et al. Mulan: a java library for multi-label learning [J]. Journal of Machine Learning Research, 2011, 12: 2411-2414.

[16] SOROWER M. A literature survey on algorithms for multi-label learning[D]. Corvallis: Oregon State University, 2010.

[17] YU H F, JAIN P, KAR P, et al. Large-scale multi-label learning with missing labels[C]//Proceedings of the 31st International Conference on Machine Learning. Beijing, 2014: 593-601.

[18] WU B Y, LIU Z L, WANG S F, et al. Multi-label learning with missing labels [C] // 2014 22nd International Conference on Pattern Recognition. Stockholm, 2014: 1964-1968.

Optimization of embedding-based multi-label classification

XIAO Xue LIU Yun*

(School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China)

Abstract: How to deal effectively with large-scale data sets with many instances and a large number of labels, compensate for missing labels in training sets, and improve prediction performance by using unlabeled instances in multi-label classification has become an important research direction. This paper proposes an embedding-based multi-label classification (EMC) algorithm. Firstly, two sets of random transformations were extracted from the pseudo-instance parameterized Gaussian process (GP) to model the nonlinear relationship mapping between feature vectors, latent space representation vectors and label vectors, and then a set of auxiliary variables combined with an expert ensemble with an overriding expert (EEOE) was introduced. The method compensates for the missing tags, and finally uses the unlabeled instance to learn the smooth mapping of the random function to improve the prediction performance. The simulation results show that compared with the FaLE and SLRM algorithms, the EMC algorithm optimizes the ability to process large data sets, compensate for missing tags, and utilize unlabeled data, thereby improving the predictive performance of class tags, with good scalability and short training time.

Key words: multi-label classification; missing labels; embedding-based; scalability

(责任编辑:吴万玲)