快速收敛截断核范数矩阵填充方法的远监督关系抽取

王烨　张百强
（中国科学院长春光学精密机械与物理研究所　吉林省长春市　130000）

摘　要：本文使用截断核范数代替核范数，进行基于低秩矩阵填充技术的关系抽取，改善远监督关系抽取存在较多噪声数据的问题。该方法具有准确率高、容噪性好的特点，能够更好的保留矩阵的主要成分，并且对于矩阵的秩函数有更好的逼近效果。本文利用具有快速收敛特性的 TNNR-ADMMAP 算法求解最小化截断核范数的凸优化子问题。

关键词：远监督学习；关系抽取；快速收敛；低秩矩阵填充；截断核范数

1 引言

关系抽取技术能够从海量数据中挖掘出反映实体之间关系的结构化数据。为了应对当今大数据时代的海量异构数据 [1]，远监督关系抽取方法被提出，该方法通过将一个已有的知识库和文本集进行启发式匹配生成训练数据。这些训练数据的产生是基于这一假设条件——“任何包含已知知识库中关系的实体对的句子，都是在用某种方式潜在的表达了这种关系”[2]。实际上存在提到某实体对的句子并未表达该实体对在知识库中对应的关系的情况，对这样的句子进行特征提取，就导致了噪声特征的产生。

Fan[3] 等人提出了基于低秩矩阵填充方法的远监督关系抽取，来应对远监督关系抽取存在较多噪声特征的问题。该方法将低秩矩阵填充技术应用在远监督关系抽取当中，准确率比通过训练分类器进行远监督关系抽取的方法有了显著的提高。然而这种基于核范数的低秩矩阵填充技术本身也存在一定局限：核范数的大小与秩的大小无法完全等价，因此对于秩函数不能取得很好的逼近效果，导致了优化目标不够明确的问题。

本文使用截断核范数代替核范数 [4] 进行远监督关系抽取，选取奇异值开始快速衰减的位置进行截断，不改变前 r 个最大的奇异值的大小，通过最小化奇异值序列中剩余的 min(m,n)-r 个奇异值得和——即截断核范数，来进行最小化秩函数的求解，从而进行基于低秩矩阵填充的远监督关系抽取。

2 相关工作

远监督学习是弱监督学习的一种，Craven[5] 等人通过将酵母蛋白质数据于 PubMed 目录进行匹配，得到的训练数据用来训练朴素贝叶斯分类器，这是远监督学习的思想第一次被提出。Snow[6] 等人使用 WorldNet 知识库作为监督来提取文本中实体之间的上下位关系。Hoffman[7] 等人提出了使用多标签的框架来适应一个实体对对应多个关系的情况。噪声特征的存在影响了远监督关系抽取的准确率，Fan[3] 等人提出了将低秩矩阵填充技术应用于远监督关系抽

使用的角度来说，其操作流程如图 3 所示。

4 结束语

本篇论文描述了一个基于 B/S 架构的随机组卷考试系统，并对教师部分所涉及到的技术做了较为详细的描述。B/S 架构对于考试之类的较复杂场景有着很大的便携性优势，另一方面，更多的诸如微信小程序之类的跨平台应用可能会更为简化校园事务 [16]，这值得我们去进一步研究实现。

参考文献

[1] 黄海明．防舞弊在线考试系统的设计与实现 [D]．江西师范大学，2015.
[2] 申田静，陈俊．国内在线考试系统研究综述 [J]．中国教育技术装备，2015 (14)：19-22.
[3] 黄春．基于 JSP 的在线考试系统的开发与设计 [J]．信息通信，2018 (2)：163-164.
[4] 金圣道．在线考试及试卷分析系统的设计与实现 [J]．电子技术与软件工程，2018 (7).
[5] 王洪祥，韩天亮，白彦国．关于在线试题系统的研究与分析 [J]．黑龙江科学，2018, v. 9; No. 124 (9)：150-152.
[6] 桑国珍．在线考试系统的设计 [J]．信息技术，2015 (9)：56-59.
[7] 蔡行，王海春，邓珊．一种基于 JSP 的题库系统设计 [J]．数字技术与应用，2015 (12)：172-173.
[8] 梁浩．基于 B/S 的在线考试系统的设计与实现 [J]．考试系统，2015.
[9] 韦小凤，顾平．关系数据库上机考试系统——组卷算法的研究与探讨 [J]．信息通信，2016 (8)：93-95.
[10] 龚利．网络考试系统中组卷算法比较及应用 [J]．黄冈职业技术学院学报，2015 (3)：94-96.
[11] 陈国彬，张广泉．基于改进遗传算法的快速自动组卷算法研究 [J]．计算机应用研究，2015, 32 (10)：2996-2998.
[12] 王悦．遗传算法在函数优化中的应用研究 [J]．电子设计工程，2016, 24 (10)：74-76.
[13] 王莉．基于遗传算法的高校在线考试系统研究 [J]．电子设计工程，2015 (24)：29-31.
[14] 卞勇．基于遗传算法在线考试系统题库的设计与实现 [J]．宁波职业技术学院学报，2016, 20 (6)：87-89.
[15] 曲志坚，张先伟，曹雁锋．基于自适应机制的遗传算法研究 [J]．计算机应用研究，2015 (11)：3222-3225.
[16] 王天泥．当图书馆遇上微信小程序 [J]．图书与情报，2016 (6)：83-86.

作者简介

顾亚文（1979-），女，江苏省阜宁县人。硕士研究生，副教授。研究方向为嵌入式。

取当中，过滤噪音数据，恢复潜在的低秩矩阵，准确率和召回率都较之前的方法有了明显的提高，但是在进行最小化矩阵的秩的过程中，矩阵的有效信息和噪音数据一同被减小。

本文提出的能够快速收敛的基于截断核范数矩阵填充技术的远监督关系抽取，利用 TNNR-ADMMAP 算法进行凸优化子问题的求解，能够较好地保留矩阵中对远监督关系抽取有效的成分，同时降低噪声数据对结果的影响。

3 基于TNNR低秩矩阵填充的远监督关系抽取过程

本文将基于截断核范数快速收敛的低秩矩阵填充技术，应用于远监督关系抽取当中。TNNR（Truncated Nuclear Norm Regularization）方法是通过针对截断核范数（奇异值序列中最小的 k 个奇异值的和）进行最小化求解，来代替最小化矩阵的秩的问题的求解，不改变前 r 个最大的奇异值的大小，其中 k=min(m,n)-r。

3.1 将远监督关系抽取转化为矩阵填充问题

将远监督关系抽取问题转化为矩阵填充问题，通过对矩阵中未知部分的填充，实现关系抽取的目的。根据训练数据集，构建一个包含 n 个实体对、t 个标签和 d 维特征向量的基于远监督关系抽取规则的待填充矩阵。

可以表示为以下形式：

$$X=\begin{bmatrix} F_{train} & L_{train} \\ F_{test} & L_{test} \end{bmatrix} \quad (1)$$

$F_{train}\in R^{n\times d}$ 代表训练数据的特征矩阵，$L_{train}\in R^{n\times t}$ 代表训练数据的标签矩阵，$F_{test}\in R^{m\times d}$ 代表测试数据的特征矩阵，$L_{test}\in R^{m\times t}$ 代表测试数据的标签矩阵。利用已观测到的 F_{train}，L_{train} 和 F_{test} 将线性分类问题转化为填充 L_{test} 中未知部分的问题。可以将低秩矩阵填充过程表示为对最小化矩阵秩函数的问题的求解。

3.2 截断长度的选取

本文通过截断的核范数代替核范数来近似表示矩阵的秩。矩阵经过奇异值分解，得到呈快速衰减趋势的奇异值序列。矩阵秩的大小等于奇异值的个数，因此对于矩阵秩的大小来说，奇异值无论大小，重要性是相同的。核范数的大小等于所有奇异值的和。因此，与核范数相比，从奇异值快速衰减处进行截断，仅保留最小的 r 个奇异值的截断核范数可以更好地逼近矩阵的秩。本文使用 ISD[8] 方法寻找奇异志序列的截断位置——以奇异值的最后显著跳跃点作为截断位置，即 r 的取值。

3.3 最小化截断核范数的问题求解

本文通过最小化截断核范数代替核范数，来求解最小化矩阵的秩的问题。以下是截断核范数的定义：

定义 3.1([4]). 对于给定矩阵 $X\in R^{m\times n}$，截断的核范数 $\|X\|_r$ 为奇异值序列中最小的 k 个奇异值的和，其中 k=min(m,n)-r。矩阵的截断核范数可表示为如下形式：

$$\|X\|_r=\sum_{i=r+1}^{\min(m,n)}\sigma_i(X) \quad (2)$$

σ_i 为 X 的第 i 个大的奇异值。r 值即核范数的截断位置。截断核范数是非凸函数，我们将最小化截断核范数问题转化为如下形式：

$$\min\|X\|_*-\max_{AA^T=I,BB^T=I}\mathrm{Tr}(AXB^T)$$

$$\text{s.t. } P_\Omega(X)=P_\Omega(M) \quad (3)$$

采用两步迭代机制求解式（3），在第 l 次迭代中：

（1）固定 X_l，通过对矩阵进行奇异值分解得到 A_l 和 B_l。

（2）固定 A_l 和 B_l，然后通过求解凸优化子问题得到 X_{l+1}，该凸优化子问题描述如下：

$$\min \|X\|_*-\mathrm{Tr}(AXB^T)$$

$$\text{s.t. } P_\Omega(X)=P_\Omega(M) \quad (4)$$

3.4 TNNR-ADMMAP优化方法

虽然 TNNR-ADMM (alternating direction method of multipliers) 可以用来求解两步迭代算法中的凸优化子问题，但由于远监督关系抽取问题计算规模较大，TNNR-ADMM 难以求得使其收敛的最优解。本文利用 ADMMAP(alternating direction method of multipliers with adaptive penalty) 求该解凸优化子问题，以达到加速收敛的目的。

TNNR-ADMMAP 算法通过对凸优化字问题的约束条件进行放松，使用自适应惩罚算法加速收敛。

首先，将式（7）的两个线性约束条件：X=W 和 $P_\Omega(W)=P_\Omega(M)$，写成如下形式：

$$X_{l+1}=\arg\min_X\|X\|_*-\mathrm{Tr}(AWB^T)$$

$$\text{s.t.}A(X)+B(M)=C \quad (5)$$

其中，A 和 $B\in R^{m\times n}\rightarrow R^{2m\times 2n}$，A 和 B 是线性映射。

因此，对应的增广拉格朗日函数为：

$$L_{AP}(X,Y,W,\beta)=\|X\|_*-\mathrm{Tr}\left(A_lWB_l^T\right)+\frac{\beta}{2}\|A(X)+B(W)-C\|_F^2+\langle Y,A(X)+B(W)-C\rangle \quad (6)$$

其中 $Y=\begin{pmatrix} Y_{11} & Y_{12} \\ Y_{21} & Y_{22}\end{pmatrix}\epsilon R^{2m\times 2n}$，为拉格朗日乘子矩阵。在 ADMM 算法中，惩罚参数 β 为定值，β 值设定得过大或过小都会显著增加计算成本。在 ADMMAP 算法中，令 β 为动态惩罚参数，采用以下自适应的原则更新 β：

$$\beta_{k+1}=\min(\beta_{max},\rho\beta_k) \quad (7)$$

其中，β_{max} 为 β 的上界，ρ 的定义如下：

$$\rho=\begin{cases}\rho_0, & \text{if } \dfrac{\beta_k\max\{\|X_{k+1}-X_k\|_F,\|W_{k+1}-W_k\|_F\}}{\|C\|_F}<\varepsilon,\\ 1, & \text{otherwise.}\end{cases}$$

其中，ρ_0 是常数，且 $\rho_0>1$，ε 为近端参数。

通过线性 ADMM，求解式（6）：

$$X_{k+1}=\arg\min_X L_{AP}(X,Y_k,W_k,\beta)=\arg\min_X\|X\|_*+\frac{\beta}{2}\left\|A(X)+B(W)-C+\frac{1}{\beta}Y_k\right\|_F^2 \quad (8)$$

$$W_{k+1}=\arg\min_W L_{AP}(X_{k+1},W,Y_k,\beta)=\arg\min_W\frac{\beta}{2}\left\|A(X_{k+1})+B(W)-C+\frac{1}{\beta}Y_k\right\|_F^2-\mathrm{Tr}(A_lWB_l^T) \quad (9)$$

$$Y_{k+1}=Y_k+\beta[A(X_{k+1})+B(W_{k+1})-C] \quad (10)$$

定义 A^*，B^*：$R^{2m\times 2n}\rightarrow R^{m\times n}$ 为 A，B 的伴随矩阵。

TNNR-ADMMAP 具体算法如下：

Algorithm1

Input:M_Ω,A_l,B_l, tolerance ϵ

Initialize:$X_1=M_\Omega$,$W_1=X_1$,$Y_1=X_1$,$\varepsilon=10^{-3}$，ρ_0，and β_0.

Repeat

STEP 1. Set Y_k and W_k fixed

$$X_{k+1}=\mathcal{D}_{\frac{1}{\beta}}\left(W_k-\frac{1}{\beta}(Y_k)_{11}\right)$$

STEP 2. Set Y_k and X_{k+1} fixed

$$W_{k+1}=\frac{1}{\beta}\left(A_l^TB_l+(Y_k)_{11}\right)+X_{k+1}+\frac{1}{2\beta}P_\Omega[\beta(M-X_{k+1})-\left(A_l^TB_l+(Y_k)_{11}+(Y_k)_{22}\right)]$$

STEP 3.Set W_{k+1} and X_{k+1} fixed

$Y_{k+1}=Y_k+\beta[A(X_{k+1})+B(W_{k+1})-C]$

STEP 4. if $\frac{\beta_k\max\{\|X_{k+1}-X_k\|_F,\|W_{k+1}-W_k\|_F\}}{\|C\|_F}<\varepsilon$

$\rho=\rho_0$,

else $\rho=1$.

$\beta_{k+1}=\min(\beta_{max},\rho\beta_k)$.

until $\|X_{k+1}-X_k\|_F\le\epsilon$

表 1

	Top-100	Top-500	Top-1000	Average
DRMC-b	82.01%	70.18%	68.20%	73.46%
DRMC-1	80.00%	77.00%	77.18%	78.06%
TNNR-ADMM	81.02%	70.2%	67.60%	72.94%
TNNR-ADMMAP	87.50%	84.50%	80.60%	84.20%

4 实验设计

将本文中的 TNNR-ADMMAP 算法与 ADMM、APGL 优化算法以及 Miao Fan 等人提出的使用低秩矩阵填充方法进行远监督关系抽取的方法——DRMC-1 和 DRMC-b 进行比较。

4.1 实验数据

本文中使用 NYT’13 数据库需要将该数据库的内容以矩阵的形式表示出来。根据 NYT’13 数据库构造出来的矩阵为稀疏矩阵，共包含 10591 个实体对，该数据库中的特征为实体对之间的依存路径。

4.2 参数选取

在实验中，我们设定 $\beta=1$，$\epsilon=10^{-4}$。对数据进行五组交叉验证，减少因数据划分对实验结果产生的影响。

4.3 实验结果

表 1 在前 100、500、1000 个关系实例预测中 DRMC-1、DRMC-b、TNNR-ADMM 和 TNNR-ADMMAP 方法的准确率。

使用同一组数据（NYT’13）进行实验，针对置信度前 100、500、1000 个关系实例的预测准确率进行统计，使用 TNNR-ADMM 算法进行关系抽取的准确率未能比 DRMC-b 和 DRMC-1 有明显提高，这是由于 ADMM 不易收敛，难以得到使其收敛的最优解，对于计算规模较大的远监督关系抽取问题的求解没有明显优势。TNNR-ADMMAP 具有较快的收敛速度，截断核范数能够更好地逼近秩函数并且能够更好的保留矩阵的主要成分信息，因此基于截断核范数的 TNNR-ADMMAP 方法准确率较高，平均准确率能够达到 84.20%。

5 展望

本文对于远监督关系抽取存在噪声特征这一特点进行了有针对性的改进，降低噪声数据对结果的影响。在以后的工作中，可以针对以下两方面开展研究：

（1）研究特征稀疏对关系抽取效果的影响。

（2）研究应对标签不完整性的方法，进一步提高远监督关系抽取的准确率。

参考文献

[1]Li Guojie, Cheng Xueqi. Research Status and Scientific Thinking of Big Data. Bulletin of the Chinese Academy of Sciences, 2012,27(6):647-657.

[2]Mike Mintz, Steven Bills, Rion Snow, Dan Jurafsky. Distant supervision for relation extraction without labeled data.In: Proc. of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, 2009,1003-1011.

[3]Fan M, Zhao D, Zhou Q, et al. Distant Supervision for Relation Extraction with Matrix Completion. In: Proc. of the 52nd Annual Meeting of theACL, 2014:839-849.

[4]Zhang D, Hu Y, Ye J, et al. Matrix completion by truncated nuclear norm regularization.In: Proc. of Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE, 2012:2192-2199.

[5]Craven M, Kumlien J. Constructing biological knowledge bases by extracting information from text sources. In: Proc. of ISMB- 1999, 1999:77-86.

[6]Snow R, Jurafsky D, Ng A Y. Learning syntactic patterns for automatic hypernym discovery. Advances in Neural Information Processing Systems, 2005:1297-1304.

[7]Hoffmann R, Zhang C, Ling X, et al. Knowledge-based weak supervision for information extraction of overlapping relations. In: Proc. of the 49th Annual Meeting of the ACL, 2011:541-550.

[8]Wang Y, Yin W. Sparse signal reconstruction via iterative support detection. SIAM Journal on Imaging Sciences, 2010,3(3):462-491.

作者简介

王烨（1989-），女，硕士学位。研究方向为人工智能自然语言处理。
张百强（1989-），男，硕士学位。研究方向为导航制导与控制。