中文电子病历数据元抽取方法*

郭维嘉[1] 郭少友[2]

([1] 河南省图书馆 郑州 450052 [2] 郑州大学信息管理学院 郑州 450001)

〔摘要〕 **目的/意义** 提出基于国家标准的电子病历数据元抽取方法，以实现电子病历数据的细粒度共享。**方法/过程** 利用 ALBERT、BiLSTM 和 CRF 模型对电子病历进行序列标注，并根据标注结果生成一组候选数据元；针对每个候选数据元，采集其上下文信息并形成一个增强的键向量；计算该向量与标准向量之间的相似度，据此判断候选数据元是否有效。**结果/结论** 该方法 *F*1 值为 90.32%，效果较好。

〔关键词〕 电子病历；数据元；ALBERT；序列标注；token 向量

〔中图分类号〕 R-058 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2024.08.013

A Method for Extracting Data Elements from Chinese Electronic Medical Records

GUO Weijia[1], *GUO Shaoyou*[2]

[1]*Henan Provincial Library*, *Zhengzhou 450052*, *China*; [2]*School of Information Management*, *Zhengzhou University*, *Zhengzhou 450001*, *China*

〔**Abstract**〕 **Purpose/Significance** A method is proposed for extracting data elements from electronic medical records (EMR) based on national standards, helping to achieve fine-grained sharing of EMR data. **Method/Process** The ALBERT, BILSTM and CRF models are used to perform sequence labeling on EMR, and a set of candidate data elements based on labeling results are generated. For any candidate data elements, the contextual information is collected to form an enhanced key vector. Then the similarity between the vector and the standard vector is calculated to determine whether the candidate data element is valid. **Result/Conclusion** The *F*1 value is 90.32%, indicating the proposed method has a good performance.

〔**Keywords**〕 electronic medical records (EMR); data element; ALBERT; sequence labeling; token

1 引言

数据元是由一组属性规定其含义、标识、表示和允许值的数据单元[1]，通常用于构建语义正确、独立且无歧义的信息单元。电子病历语境下的数据元指电子病历中具有特定标识符、名称和值域的数据单元[2]。国家卫生主管部门于 2011 年发布卫生行业标准《卫生信息数据元目录》[3]，并在《电子病历共享文档规范》《电子病历基本数据集》等相关标准中使用[4-5]。医疗机构从电子病历中抽取数据元，据此生成标准化的共享文档和基本数据集，一方面可以提高医疗机构信息互联互通标准化成熟度，另一方面实现电子病历的数据元级交换与共享。以某患者电子病历中的“现病史”字段为例，可从中抽取 4 个数据元，其中第 4 个数据元为整个现病史字段，见图 1。

〔修回日期〕 2024-03-12

〔作者简介〕 郭维嘉，助理馆员，发表论文 9 篇；通信作者：郭少友，教授。

〔基金项目〕 国家社会科学基金一般项目（项目编号：20BTQ063）。

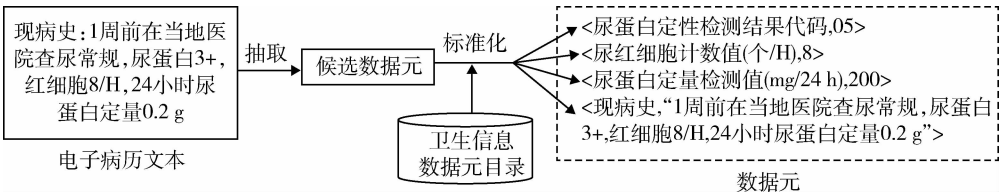

图 1 数据元抽取结果示例

2 相关研究

相关研究大致可分为两类：第 1 类，电子病历命名实体及其关系抽取；第 2 类，电子病历数据元抽取，以抽取自定义数据元或标准数据元为目标，抽取结果可能是数据元键，也可能是数据元键值对。

2.1 电子病历命名实体及其关系抽取

抽取的实体大致可归为实验室检验、影像检查、手术、疾病、症状、药物等几类，抽取结果虽然称为实体而不是数据元，但这些实体经过规范后可以作为数据元的组成部分。早期方法以基于规则与词典的方法、基于统计机器学习的方法为主，目前采用的主流方法是基于循环神经网络（recurrent neural network，RNN）和预训练语言模型的深度学习方法[6]。

2.1.1 基于 RNN 的方法 采用 RNN 或其变种如双向长短期记忆（bidirectional long short term memory，BiLSTM）抽取电子病历中的实体及关系，例如，Chowdhury S 等[7]提出面向中文电子病历命名实体识别的多任务双向 RNN 模型，曾青霞等[8]提出结合自注意力和条件随机场（conditional random field，CRF）的电子病历命名实体识别方法。

2.1.2 基于预训练语言模型的方法 采用双向编码器表征（bidirectional encoder representations from transformers，BERT）或其变种实现实体及关系抽取，例如，Qi R 等[9-10]提出基于 BERT－BiLSTM－CRF 的中文电子病历命名实体抽取方法，陈杰等[11-12]提出基于 ALBERT－BiLSTM－CRF 的中文医疗病历命名实体抽取方法，张芳丛等[13-14]提出基于 RoBERTa－BiLSTM－CRF 的电子病历命名实体识别方法，其中基于转换器的精简双向编码表征（a lite bidirectional encoder representations from transformers，ALBERT）属于 BERT 的精简版，稳健优化的 BERT 方法（robustly optimized BERT approach，RoBERTa）属于 BERT 的改进方法。

2.2 电子病历数据元抽取

这类研究可进一步分为两种情况。一是从电子病历中抽取实体并据此形成自定义的数据元。例如，Kuo T 等[15-16]采用自然语言处理技术从心力衰竭、川崎病、充血性心力衰竭等疾病文本中抽取数据元，Margaret M 等[17]提出基于 pyConTextNLP 算法的数据元抽取框架 tbiExtractor，可以从患者的放射检查报告中抽取创伤性脑损伤数据元。二是抽取遵循国家标准的数据元。例如，汤学民等[18]采用基于规则的方法从电子病历中提取数据元，肖晓霞等[19]采用隐马尔可夫模型（hidden Markov model，HMM）、CRF 等机器学习算法提取中医临床症状数据元，孙静等[20]从识别对象类、构建诊断信息模型、提取数据元、提取结果等 4 个方面阐述中医诊断信息数据元手工提取过程。综合来看，第 1 种研究虽然成果丰富，但抽取结果大都是一些医疗实体，并不是真正意义上的数据元，陈杰等[11]虽然与本文一样也采用 ALBERT－BiLSTM－CRF 模型，但仅限于利用该模型抽取命名实体，并未进一步实现数据元的抽取。第 2 种研究相对较少，依据《卫生信息数据元目录》进行中文电子病历数据元抽取的成果更少，在最相关的 3 篇文献中，汤学民等[18]采用基于规则的传统抽取方法，孙静等[20]提出的方法需要借助人工完成，肖晓霞等[19]虽然在抽取过程中采用传统机器学习方法，但整体上仍以人工抽取为主。本文拟借鉴现有的实体抽取方法，采用 AL-

BERT、BiLSTM 等深度学习模型从电子病历中抽取候选数据元；再通过进一步考虑候选数据元的上下文向量完全自动地确定候选数据元是否有效。

3 数据元抽取方法

本文所提方法的基本思路：首先以句子为单位，利用 ALBERT[21]、BiLSTM[22]、CRF[23] 对输入的电子病历文本进行序列标注；然后根据标注结果生成一组候选数据元，通过计算候选数据元键、值的 token 向量均值，得到每个候选数据元的键向量、值向量，进一步考虑键的上下文，生成增强的键向量；最后计算候选数据元键向量与标准数据元键向量集合中所有向量之间的相似度并取最大值，当最大值大于指定阈值时，即可将相应的候选数据元转换为一个最终的抽取结果，见图 2，其中标注标签共有 5 个：B－Key、I－Key、B－Val、I－Val、O，分别表示数据元键第 1 个字符、数据元键后续字符、数据元值第 1 个字符、数据元值后续字符、非数据元键/值字符。

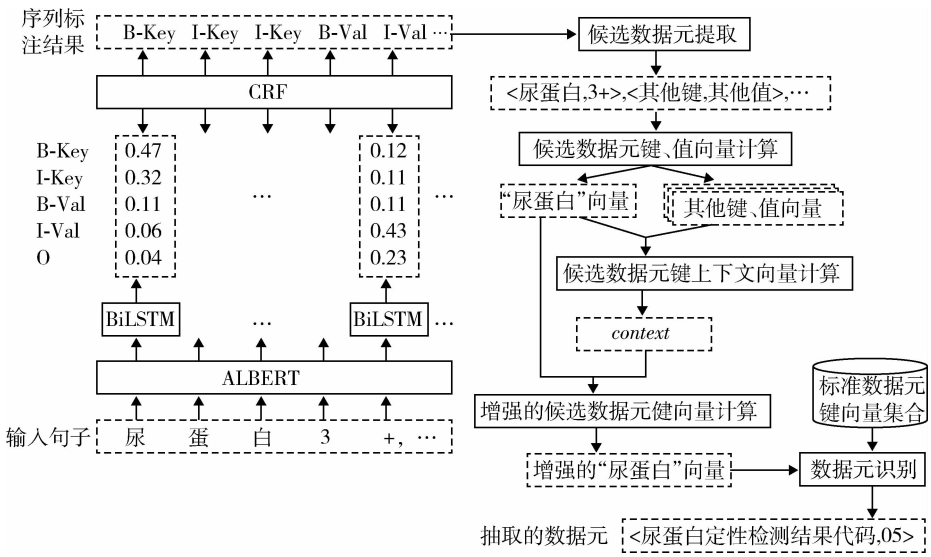

图 2 数据元抽取方法

3.1 电子病历文本序列标注

通过 ALBERT、BiLSTM、CRF 3 个模型完成输入句子的序列标注。首先通过 ALBERT 为句子中的每个字符生成一个 128 维的 token 向量，然后通过 BiLSTM 计算每个字符的标注得分向量，最后通过 CRF 计算每个可能的标注序列的概率得分，并将具有最高得分的标注序列作为输入句子的最优标注结果。如前文图 2 所示，针对输入句子“尿蛋白 3 +, …”，BiLSTM 在 ALBERT 输出结果的基础上计算每个字符的标注得分，如“尿”字在 B－Key 标签上的得分是 0.47，在 I－Val 标签上的得分是 0.06；CRF 计算出最优的序列标注结果即（B－Key，I－Key，I－Key，B－Val，I－Val，…）。

3.2 向量计算

3.2.1 候选数据元提取 设输入句子 $t = [t_1, \cdots, t_i, \cdots, t_n]$，其中 t_i 表示第 i 个 token（即字符）；序列标注结果 $ta = [ta_1, \cdots, ta_i, \cdots, ta_n]$，其中 ta_i 表示 t_i 的标注结果。候选数据元提取方法是：遍历 t，参考 5 个标签的合法逻辑组配情况和 ta 的实际标注结果，从 t 中提取 0 至多个候选数据元。若提取结果中存在候选数据元 $<ck_i, cv_i>$，其中 ck_i 由 t 中的 token 序列 $t_k t_{k+1} \cdots t_{k+q}$ 组成，则 t_k 在 ta 中对应的标注结果 ta_k 为 B－Key，其余 token 的标注结果均为 I－Key；cv_i 由 t 中的 token 序列 $t_r t_{r+1} \cdots t_{r+z}$ 组成，则 t_r 在 ta 中对应的标注结果 ta_r 为 B－Val，其余 token 的标注结果均为 I－Val。如前文图 2 所示，可从序列标注结果中提取一个候选数据元，

即 < 尿蛋白，3 + >。

3.2.2 候选数据元键、值向量计算 对于提取的任意一个候选数据元 $< ck_i, cv_i >$，键 ck_i、值 cv_i 的向量都可通过计算各自所包含字符的 token 向量均值得到。以 ck_i 向量为例，设 ck_i 由 $q+1$ 个 token 组成，即 $ck_i = t_k t_{k+1} \cdots t_{k+q}$，则 ck_i 向量 v_i 计算方式如下。其中 x_{k+u} 表示 ck_i 中第 $k+u$ 个字符的 token 向量。

$$v_i = \frac{1}{q+1} \sum_{u=0}^{q} x_{k+u} \tag{1}$$

3.2.3 候选数据元键上下文向量计算 对于抽取的任意一个候选数据元 $< ck_i, cv_i >$，其中 ck_i 的语义不但与其包含的各字符的 token 向量有关，还与 cv_i 以及 ck_i 所在句子中其他候选数据元有关，后者统称为 ck_i 的上下文。计算 ck_i 上下文向量 context 方式如下。其中，N 为输入句子中所有候选数据元的键、值总个数，v_i 为 ck_i 的键向量，v_j 为第 j 个键（或值）的向量，sim（v_i，v_j）为 v_i 与 v_j 之间的余弦相似度，a_j 表示经过 softmax 函数归一化后的相似度。

$$a_j = \text{softmax}(\text{sim}(v_i, v_j)) = \frac{\exp(\text{sim}(v_i, v_j))}{\sum_{j=1}^{N} \exp(\text{sim}(v_i, v_j))} \tag{2}$$

$$\text{context} = \sum_{j=1}^{N} (a_j \times v_j) \tag{3}$$

3.2.4 增强的候选数据元键向量计算 对于抽取的任意一个候选数据元 $< ck_i, cv_i >$，其中 ck_i 的最终向量由 ck_i 的键向量 v_i 和 ck_i 的上下文向量 context 相加得到，称为 ck_i 的增强向量，计算方式如下。

$$ev_i = v_i + \text{context} \tag{4}$$

3.3 数据元识别

对于抽取的任意一个候选数据元 $< ck_i, cv_i >$，可通过如下方法判断其是否可认定为有效的数据元：事先根据 ALBERT 为《卫生信息数据元目录》中的每个标准数据元生成一个键向量，形成一个标准数据元键向量集合；计算 ev_i 与集合中所有键向量之间的余弦相似度，当最高相似度大于给定阈值 β 时，用相应的标准数据元名称替换候选数据元 $< ck_i, cv_i >$ 中的 ck_i，在对 cv_i 进行必要的数据格式转换后，即可将该候选数据元认定为有效的数据元。

4 数据元抽取实验

4.1 数据准备

实验数据来自郑州大学附属医院 410 份脱敏后的肾病患者入院记录。从每份记录中提取既往史、现病史、专科检查、辅助检查 4 个字段的内容并进行数据清洗之后，参考《卫生信息数据元目录》，由人工对提取内容所包含的实验室检查、体格检查、临床辅助检查、药品 4 类数据元标注，生成 3 个数据集：训练集、验证集、测试集。训练集用于对 ALBERT 模型微调训练，验证集用于调整模型参数，测试集用于测试模型性能，记录数量分别为 246、82 和 82，包含句子数量分别为 6 642、2 356 和 2 415。标注策略采用常用的 BIO 方案，即对于候选数据元键，开始字符标注为 B - Key，其余字符标注为 I - Key；对于候选数据元值，开始字符标注为 B - Val，其余字符标注为 I - Val；非数据元字符标注为 O。

4.2 评价指标

采用准确率（P）、召回率（R）和 $F1$ 值作为评价指标，P 表示正确识别出的数据元与全部识别出数据元的比值，R 表示正确识别出的数据元与应识别出数据元的比值，$F1$ 值计算方式如下。

$$F1 = \frac{2\text{PR}}{\text{P} + \text{R}} \times 100\% \tag{5}$$

4.3 实验环境及参数

实验环境：Windows10，Intel（R）Core i7 - 10750H CPU @ 2.60 GHz2.59 GHz，内存 128 GB（DDR4），Python3.6.8，Tensorflow1.15.0，ALBERT - base 版本。主要的实验参数设置：优化器参数 optimizer 值为 adam，句子序列长度参数 max_seq_len 值为 256，参数 Dropout 值为 0.5；相似度阈值参数 β 初值设为 0.8，以 0.01 为增幅重复实验，最终确定 β 的最优取值为 0.92。

4.4 实验结果分析

4.4.1 抽取结果及分析 抽取结果，见表 1。虽

然本文方法的 *F*1 为 90.32%，整体效果良好，但不同类型数据元的抽取效果相差较大，实验室检查类和临床辅助检查类抽取效果明显优于体格检查类和药品类，其中药品类的准确率低于实验室检查类和临床辅助检查类 11 个百分点，原因主要有两点。一是不同类型数据元的易识别性存在差异。实验室检查类和临床辅助检查类文本大都同时包含较为明确的数据元键和数据元值，数据元的易识别性相对较高。这一特点既有利于人工标注以提高 ALBERT 的训练效果，也有利于测试阶段的候选数据元抽取。而一些体格检查类和药品类文本只包含候选键或候选值，数据元的易识别性相对较低，需要借助于前期较精细的人工标注和足够多的微调训练才能正确抽取。二是数据集规模较小。3 个数据集共有 11 413 个句子，实际包含 2 262 个数据元，其中体格检查类和药品类数据元分别只有 389 个和 202 个，占比较低，导致 ALBERT 微调训练不足、不同类型数据元的训练程度不均衡。

表 1 本文方法抽取结果

数据元类型	实际数据元数（个）	正确抽取数（个）	错误抽取数（个）	抽取数（个）	P（%）	R（%）	*F*1（%）
实验室检查类	1 125	1 032	84	1 116	92.47	91.73	92.10
体格检查类	389	330	37	367	89.92	84.83	87.30
临床辅助检查类	546	493	41	534	92.32	90.29	91.29
药品类	202	175	41	216	81.02	86.63	83.73
合计	2 262	2 030	203	2 233	90.91	89.74	90.32

4.4.2 *对比实验结果及分析* 部分研究[18-20]虽与本文方法相似度较高，但均未提供抽取方法技术细节，无法对比；另有研究[15-17]与本文相似度较低，不适合对比。鉴于此，本文设计了两个基于 ALBERT－BiLSTM－CRF 的数据元抽取实验，用于说明本文所提方法的效果。（1）ABC（即 ALBERT－BiLSTM－CRF）方法实验。根据 ALBERT－BiLSTM－CRF 抽取候选数据元，直接计算候选数据元键与标准数据元集合中各数据元名称之间的相似度，进而确认候选数据元是否有效。（2）ABC＋token 向量方法实验。根据 ALBERT－BiLSTM－CRF 抽取候选数据元，计算候选数据元中各字符 token 向量的均值并将其作为候选数据元键向量，计算该键向量与标准数据元键向量集合中各向量之间的相似度，进而确认候选数据元是否有效。实验结果，见表 2。

表 2 不同方法实验结果对比

实验方法	P（%）	R（%）	*F*1（%）
ABC 方法	85.94	87.37	86.65
ABC＋token 向量方法（简称 ABCT 方法）	86.02	89.38	87.67
本文方法	90.91	89.74	90.32

相较于 ABC 方法，ABCT 方法的 *F*1 值提高 1.18%；相较于 ABC 方法和 ABCT 方法，本文方法的 *F*1 值分别提高了 4.24% 和 3.02%。可能的原因：ABCT 方法根据输入句子动态生成各字符的 token 向量，并在此基础上生成候选数据元键向量，在一定程度上降低了具体场合下候选数据元一词多义的可能性，因此，ABCT 方法的 *F*1 值略高于 ABC 方法；从同一个输入句子中抽取的若干个候选数据元，彼此之间互为上下文，有助于进一步消除候选数据元的一词多义情况，本文方法考虑了这一情况，因此 *F*1 值高于 ABCT 方法。

5 结语

本文提出一种基于 ALBERT－BiLSTM－CRF 和上下文向量的电子病历数据元抽取方法。实验结果显示，在自定义数据集上，*F*1 值为 90.32%，抽取效果良好。不足之处：自定义数据集规模较小，且数据元类型分布不均衡，药品类数据元存在数据稀疏问题，在一定程度上影响整体抽取效果；数据只来源于肾病患者入院记录中的既往史、现病史、专科检查、辅助检查 4 个字段，且只抽取其中的实验室检查、体格检查、临床辅助检查、药品 4 类数据元，抽取方法的普适性有待进一步验证。后续将通过增加样本类型和数量、扩大抽取范围解决上述问题。

作者贡献：郭维嘉负责论文撰写、实验数据准备与系统测试；郭少友负责研究设计和实验结果分析。
利益声明：所有作者均声明不存在利益冲突。

参考文献

1 信息技术 - 元数据注册系统（MDR）第一部分：框架 [EB/OL]. [2024 - 02 - 18]. https://std.samr.gov.cn/gb/search/gbDetailed?id = 71F772D7CE6ED3A7E05397BE0A0AB82A.
2 卫生信息数据元目录第 1 部分：总则 [EB/OL]. [2024 - 02 - 18]. http://www.nhc.gov.cn/wjw/s9497/202311/879235ee175344c8ab6c02f8d02cd255/files/bed9599f67aa42d0b573801ab39d5735.pdf.
3 卫生信息数据元目录 [EB/OL]. [2023 - 05 - 18]. http://www.nhc.gov.cn/wjw/s9497/wsbz_2.shtml.
4 电子病历共享文档规范 [EB/OL]. [2023 - 05 - 20]. http://www.nhc.gov.cn/wjw/s9497/wsbz_5.shtml.
5 电子病历基本数据集 [EB/OL]. [2023 - 05 - 18]. http://www.nhc.gov.cn/wjw/s9497/wsbz_11.shtml.
6 崔博文，金涛，王建民. 自由文本电子病历信息抽取综述 [J]. 计算机应用，2021，41（4）：1055 - 1063.
7 CHOWDHURY S, DONG X, QIAN L, et al. A multitask bi - directional RNN model for named entity recognition on Chinese electronic medical records [J]. BMC Bioinformatics, 2018, 19（S17）: 76 - 84.
8 曾青霞，熊旺平，杜建强，等. 结合自注意力的 BiLSTM - CRF 的电子病历命名实体识别 [J]. 计算机应用与软件，2021，38（3）：159 - 162.
9 QI R, WU X, ZHANG Q, et al. A named entity recognition method for Chinese electronic medical records based on BERT [J]. Theory and practice of science and technology, 2020, 1（1）: 1 - 12.
10 ZHANG W, JIANG S, ZHAO S, et al. A BERT - BiLSTM - CRF model for Chinese electronic medical records named entity recognition [C]. Xiangtan: IEEE 2019 International Conference on Intelligent Computation Technology and Automation（ICICTA）, 2019.
11 陈杰，奚雪峰，皮洲，等. 基于 ALBERT 的中文医疗病历命名实体识别 [J]. 南京师范大学学报（工程技术版），2021，21（1）：36 - 43.
12 马诗语，黄润才. 基于 ALBERT 与 BILSTM 的糖尿病命名实体识别 [J]. 中国医学物理学杂志，2021，38（11）：1438 - 1443.
13 张芳丛，秦秋莉，姜勇，等. 基于 RoBERTa - WWM - BiLSTM - CRF 的中文电子病历命名实体识别研究 [J]. 数据分析与知识发现，2022，6（Z1）：251 - 262.
14 张云秋，汪洋，李博诚. 基于 RoBERTa - wwm 动态融合模型的中文电子病历命名实体识别 [J]. 数据分析与知识发现，2022，6（Z1）：242 - 250.
15 KUO T, HUH J, KIM J, et al. The Impact of automatic pre - annotation in clinical note data element extraction - the CLEAN tool [EB/OL]. [2023 - 12 - 22]. https://arxiv.org/abs/1808.03806.
16 KUO T, RAO P, MAEHARA C, et al. Ensembles of NLP tools for data element extraction from clinical notes [C]. Chicago: AMIA 2016 Annual Symposium, 2016.
17 MARGARET M, DANIEL R, HANNAH C, et al. TbiExtractor: a framework for extracting traumatic brain injury common data elements from radiology reports [J]. Plos one, 2020, 15（7）: e0214775.
18 汤学民，吴晓云. 基于前馈神经网络电子病历辅助诊断系统的研究 [J]. 中国数字医学，2023，18（3）：42 - 48.
19 肖晓霞. 基于机器学习的中医临床症状数据元研究 [D]. 长沙：湖南中医药大学，2019.
20 孙静，邓文萍，毛树松. 中医诊断信息数据元提取研究 [J]. 医学信息学杂志，2015，36（2）：61 - 64.
21 LAN Z, CHEN M, GOODMAN S, et al. ALBERT: a lite BERT for self - supervised learning of language representations [EB/OL]. [2023 - 03 - 10]. https://arxiv.org/pdf/1909.11942.
22 GRAVES A, SCHMIDHUBER J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures [J]. Neural networks the official journal of the international neural network society, 2005, 18（5/6）: 602 - 610.
23 LAFFERTY J, MCCALLUM A, PEREIRA F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data [C]. Williamstown: International Conference on Machine Learning（ICML）, 2001.