基于共享表示的跨领域中文模糊限制语识别

周惠巍[1]， 宁时贤[1]， 杨云龙[1]， 刘 壮[1]， 林英玉[1]， 李思嘉[2]

(1. 大连理工大学 计算机科学与技术学院 辽宁 大连 116024;
2. 台湾逢甲大学 资讯电机学院 台湾 台中 40743)

摘要: 为充分利用源领域的标注数据,减少目标领域的标注代价,提出一种基于共享表示的跨领域模糊限制语识别方法.该方法利用双向长短期记忆网络,通过参数共享机制交替地学习源领域和目标领域的训练数据,同时引入对抗学习,把各领域私有特征从共享特征中剥离,从而获得不同领域间的共享语义表示.在中文生物医学和维基百科两个领域上的实验表明,基于共享表示的方法在跨领域中文模糊限制语识别性能上明显优于基于实例和基于特征的迁移学习方法.

关键词: 中文模糊限制语识别; 跨领域; 共享表示; 对抗学习

中图分类号: TP391.1 **文献标志码:** A **文章编号:** 1671-6841(2019)02-0034-06

DOI: 10.13705/j.issn.1671-6841.2018249

0 引言

模糊语言是一种常见的语言现象,模糊限制语(hedges)用来指“把一些事情弄得模模糊糊的词语”,表示不确定性的观点[1].由模糊限制语所引导的信息称为模糊限制信息.2010 年国际计算语言学协会将模糊限制信息检测定为 CoNLL(conference on computational natural language learning)共享任务[2],极大促进了英文模糊限制语的识别研究.

中文模糊限制语在不同领域中的作用存在差异.传统的机器学习方法假设训练数据和测试数据分布相同.但是由于中文模糊限制语存在领域特性,使得现有的基于某个领域训练得到的识别模型很难直接应用于其他领域.同时,中文模糊限制语语料缺乏,语料标注费时费力,为每个领域都标注大量训练语料是不现实的.文献[3]指出可以利用资源丰富的领域(源领域)的模糊限制语语料,辅助资源贫乏的领域(目标领域)的模糊限制语的识别,从而减少目标领域的数据标注代价.

早期的模糊限制语识别是基于词典匹配的方法,该方法取得了较高的召回率,但是精确率却很低.机器学习的方法弥补了这个缺点.基于分类的 passive aggressive 方法在新闻领域获得了 70.53%的模糊限制性句子识别 F 值[4].基于序列标注方法识别中文模糊限制语,在构建的《计算机学报》语料上获得 43.2%的 F 值[5].在科技文献、股市和产品评论 3 个领域,构建基于特征的序列标注模型,分别获得 73.27%、70.29%和 68.57%的 F 值[6].

上述模糊限制语识别方法的训练数据和测试数据均采用同领域的语料,即假定训练数据与测试数据具有相同的分布.然而,模糊限制语的使用具有领域特性.文献[3]将迁移学习用于跨领域英文模糊限制语识别.当训练数据与测试数据分布不一致时,迁移学习能够在不增加标注成本的情况下,提高系统在测试数据中的检测性能.迁移学习主要分为两种:基于特征的迁移学习[7]和基于实例的迁移学习[8].文献[7]的特征迁移算法 FruDA 引入源领域和目标领域的公共特征,实现源领域知识向目标领域的迁移.文献[8]的实例迁移学习算法 TrAdaBoost 通过迭代,调整源领域与目标领域训练样例的权重,从而挑选出与目标领域数据分布相似的源领域训练样例.

近年来,随着深度学习的兴起,神经网络被用于领域间共享特征表示的学习,并取得了较好的结果.文

收稿日期: 2018-09-06
基金项目: 国家自然科学基金项目(61772109,61272375);教育部人文社科项目(17YJA740076).
作者简介: 周惠巍(1969—),女,吉林长春人,副教授,主要从事生物医学信息处理和中文模糊限制信息检测研究,E-mail: zhouhuiwei@dlut.edu.cn.

献 [9] 利用两种语言间拼写的相似之处，学习两种语言的共享字符表示，同时学习各语言的私有词表示，用于跨语言序列标注任务.文献 [9] 共享字符表示学习方法，难以学习到没有共同字符的两种语言间的共享特征.为了克服这一问题，文献 [10] 采用一个共享的 BLSTM(bidirectional long short-term memory) 模块和多个语言特定的私有 BLSTM 模块分别学习多语言间的共享表示和各语言的私有表示.同时，在共享 BLSTM 模块中引入了对抗学习，使得共享模块变得与语言无关，从而获得不含有私有特征的更纯净的共享表示.文献 [11] 利用多个中文分词语料库学习共享表示，并引入对抗训练方法抽取不同分词标准间的共享特征，有效提高在各个语料上的分词性能.

本文研究跨领域中文模糊限制语的识别，针对目标领域训练数据非常稀少的情况(仅 200 个标注样例) ，提出一种基于共享表示的跨领域中文模糊限制语识别方法.训练时，利用源领域大量标注数据和目标领域少量标注数据，交替学习各个领域的数据; 同时引入对抗训练 [12] 获得更纯净的共享表示.本文提出的方法能够有效利用源领域和目标领域信息，取得了比传统的迁移学习方法更好的跨领域识别性能.

1　基于共享表示的跨领域中文模糊限制语识别模型

文献 [11] 提出融合对抗训练的共享-私有模型，本文称其为 Sh-pri 模型.我们借鉴文献 [11] 的方法，基于跨领域中文模糊限制语识别的实际问题，提出一种共享-对抗(Sh-adv) 模型，用于跨领域模糊限制语识别，如图 1 所示.

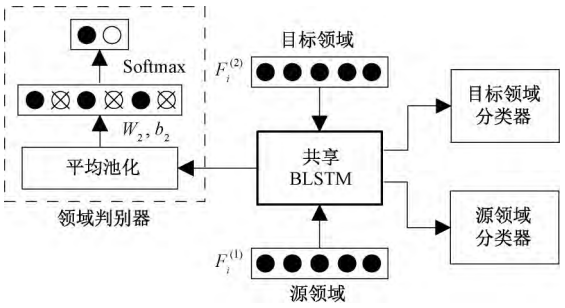

图 1　共享-对抗模型

Fig.1　The architecture of Sh-adv model

两个模型均采用了共享表示学习和对抗训练的思想，利用共享 BLSTM 模块学习源领域和目标领域间的共享语义表示.Sh-pri 模型在学习共享语义表示的同时，学习了各领域私有语义表示.Sh-adv 模型未学习私有语义表示，而是直接引入领域判别器模块，与共享 BLSTM 进行对抗训练，获得剥离领域私有特征的更纯净的共享表示.因为本文假设目标领域训练数据极其稀少(仅 200 个标注样例) ，在整体模型中引入私有 BLSTM 模块无法充分学习到目标领域的私有语义表示.而源领域训练数据远远大于目标领域，共享语义表示可能受到私有语义表示的影响，降低目标领域的模糊限制语识别性能.

1.1　数据处理及特征抽取

给定训练样本数据集 $\{x_i^{(m)}, y_i^{(m)}\}_{i=1}^{N_m}$，其中: $x_i^{(m)}$ 是领域 m 的输入数据; $y_i^{(m)}$ 是 $x_i^{(m)}$ 对应的模糊限制语分类标签; N_m 是领域 m 的训练样本个数.数据预处理采用文献 [13] 的方法，即先基于模糊限制语词典，利用最大字长匹配，获得候选模糊限制语.然后抽取候选模糊限制语的相关特征作为模型输入.相关特征包括窗口大小为 2 的上下文特征 $x_i^{(m)} = (\boldsymbol{w}_{i-2}, \boldsymbol{w}_{i-1}, \boldsymbol{w}_i, \boldsymbol{w}_{i+1}, \boldsymbol{w}_{i+2}) \in \mathbf{R}^{5\times d_1}$，$d_1$ 为词向量 $\boldsymbol{w}$ 的维度; 词性特征 $pos_i^{(m)} = (\boldsymbol{p}_{i-2}, \boldsymbol{p}_{i-1}, \boldsymbol{p}_i, \boldsymbol{p}_{i+1}, \boldsymbol{p}_{i+2}) \in \mathbf{R}^{5\times d_2}$，$d_2$ 为词性向量 $\boldsymbol{p}$ 的维度; 以及候选模糊限制语 w_i 对应的共现特征 $cooccurrence_i^{(m)} \in \mathbf{R}^{d_3}$，$d_3$ 为共现特征的维度.共现特征是指一个句子中是否有两个或两个以上的候选模糊限制语同时出现，比如“不是 … 就是 …”. 如果一个句子中存在两个或两个以上模糊限制语候选词，$cooccurrence_i^{(m)} = 1$，否则 $cooccurrence_i^{(m)} = 0$. 拼接上述特征 $F_i^{(m)} = (x_i^{(m)}; pos_i^{(m)}; 5 * cooccurence_i^{(m)}) \in \mathbf{R}^{5\times(d_1+d_2+d_3)}$ 得到模型的训练数据 $\{F_i^{(m)}, y_i^{(m)}\}_{i=1}^{N_m}$.

1.2　共享-对抗(Sh-adv) 模型

文献 [11] 的 Sh-pri 模型考虑了不同领域间数据的共性和差异性，但并不能保证共享语义表示不受到私

有语义表示的影响.因此,我们提出了 Sh-adv 模型,不考虑各领域的私有语义表示,仅使用共享语义表示进行分类.首先,Sh-adv 模型利用一个共享 BLSTM 模块学习领域间的共享语义表示 $h^s = BLSTM_{sh}(F_i^{(m)}, \overrightarrow{h}_{i-1}, \overleftarrow{h}_{i+1}, \theta_s)$(图1中的加粗黑色矩形框),其中 $\overrightarrow{h}_i$ 和 $\overleftarrow{h}_i$ 分别表示前向和后向 LSTM 在位置 i 的隐层状态,θ_s 表示共享模块的参数.然后对共享语义表示 h^s 进行平均池化操作,并输入到 Softmax 分类器中,预测当前输入的模糊限制语分类标签 $\hat{y}_i^{(m)} = Softmax(W_1^{\mathrm{T}} h^s + b_1)$. 模糊限制语识别任务的损失函数定义为

$$J_{\text{hedge}} = -\sum_{i=1}^{N_m}\sum_{j=1}^{C}\left(y_{i,j}^{(m)}\log\left(\hat{y}_{i,j}^{(m)}\right)\right),$$

其中: C 是模糊限制语类别的数量; $\hat{y}_{i,j}$ 表示第 i 个输入样例的第 j 个类别的预测概率; $m = 1$ 表示源领域,$m = 2$ 表示目标领域.

同时,为减少私有语义表示的噪声混入,在共享 BLSTM 模块中引入一个领域判别器模块与共享表示学习形成对抗.其主要思想是设定一个生成器 G(generator)和一个判别器 D(discriminator),生成器的目的是尽量去学习真实的数据分布,而判别器的目的是尽量正确判别输入数据是来自真实数据还是来自生成器.二者相互对抗并交替优化,不断提高各自的生成能力和判别能力.具体来说,Sh-adv 模型将共享 BLSTM 模块视为生成器 $G(F_i^{(m)}, \theta_s) = BLSTM_{sh}(F_i^{(m)}, \overrightarrow{h}_{i-1}, \overleftarrow{h}_{i+1}, \theta_s)$,求出共享语义表示 $h^s = (h_1^s, h_2^s, \cdots, h_n^s)$,其中 n 表示 BLSTM 的隐层时间步数量.生成的共享表示 h^s 将会输入到领域判别器中,判断当前的输入样例属于哪个领域,$D(h^s, \theta_d) = Softmax(W_2^{\mathrm{T}} f(h^s) + b_2)$,其中: $f(h^s) = \frac{1}{n}\sum_{i=1}^{n} h_i^s$ 为共享表示的隐层状态平均值; θ_d 为领域判别器的参数.针对不同领域数据输入,引入对抗损失 J_{adv} 训练模型,让共享 BLSTM 模块生成领域间的共享语义表示误导领域判别器; 同时,领域判别器要尽可能准确可靠地预测数据所属领域.两者形成对抗,借此剥离共享表示中的领域私有特征,获得更纯净的共享特征.定义对抗损失 J_{adv} 为

$$J_{adv} = \min_{\theta_s}\left(\lambda \max_{\theta_d}\left(\sum_{m=1}^{2}\sum_{i=1}^{N_m} t_i^{(m)}\log\left[D(G(F_i^{(m)}, \theta_s), \theta_d)\right]\right)\right),$$

其中: $t_i^{(m)}$ 是当前输入样例 $F_i^{(m)}$ 所属的正确领域标签(源领域或目标领域); λ 是权重系数,用于控制对抗损失中判别器和共享模块间的比重.

2 实验结果与分析

2.1 实验数据及设置

实验采用文献[14]构建的中文模糊限制语语料库(https://github.com/DUT-NLP/CHScope),包含维基百科、生物医学文献的实验结果、摘要和讨论4部分语料,共24 414句,约100万词.各部分模糊限制语的个数分别是1 958、1 622、2 759和4 674.维基百科中,33.78%的句子包含模糊限制信息; 生物医学文献中,实验结果中27.8%的句子、摘要中25.28%的句子和讨论中47.69%的句子包含模糊限制信息.为检测维基百科和生物医学两个领域间的跨领域中文模糊限制语识别性能,共设置了6组实验,如表1所示.

表1 实验设置

Tab.1 Experiment setup

组号	源领域	目标领域	组号	源领域	目标领域
1	维基百科	生物医学(讨论)	4	生物医学(讨论)	维基百科
2	维基百科	生物医学(摘要)	5	生物医学(摘要)	维基百科
3	维基百科	生物医学(实验结果)	6	生物医学(实验结果)	维基百科

为减小偶然性,每组数据进行实验时,我们都做了五折交叉实验,将目标领域数据平均分为5份,取每份中的200个实例作为训练数据,其余4份作为测试数据.实验采用 F 值对模型进行评价,公式为

$$F = 2PR/(P + R),$$

其中: P 表示准确率; R 表示召回率.

我们从万方数据库下载了 6.19 MB 的生物医学文献摘要和 106 MB 的中文维基百科语料库，加上实验所用的 4.16 MB 语料，共计 117 MB 的语料用于训练词向量.采用 Word2vec 工具训练词向量.词性向量和共现特征向量均为随机初始化，通过模型训练进行调整.词向量、词性向量和共现特征向量分别是 100 维、50 维和 10 维.模型采用随机梯度下降策略进行参数更新，对抗学习的权重系数 $\lambda = 0.05$.

2.2　基线方法

为了探知共享表示对跨领域中文模糊限制语识别的影响，我们比较了下列 4 种基线方法，分别是: 线形核函数的支持向量机 SVM，单层的无共享机制的双向长短期记忆神经网络 BLSTM_NO，以及两种性能优异的迁移学习的方法: FruDA[7] 特征迁移学习和 TrAdaBoost[8] 实例迁移学习.基线方法使用 Target Only(TO) 、Source Only(SO) 、Target+Source(T+S) 3 种数据形式.

Target Only(TO) : 仅使用目标领域的 200 个标注数据训练获得识别模型.

Source Only(SO) : 仅使用源领域的标注数据训练获得识别模型.

Target+Source(T+S) : 同时使用 TO 数据和 SO 数据训练获得识别模型.

测试时，模型对目标领域测试数据进行检测.另外，FruDA 和 TrAdaBoost 方法使用 T+S 数据进行训练，所用的特征与本文其他方法相同，参数设置全部采用默认值.4 种基线方法在 6 组实验的平均 F 值如表 2 所示.

表 2　基线方法的跨领域中文模糊限制语识别 F 值

Tab.2　F-value of cross-domain Chinese hedge cue detection by baseline methods　%

方法	数据	组 1	组 2	组 3	组 4	组 5	组 6	平均
SVM	TO	38.29	37.55	42.86	32.41	32.41	32.41	35.99
	SO	64.26	55.54	60.01	74.51	63.58	61.78	63.28
	T+S	67.01	64.22	63.93	75.04	66.51	63.82	66.76
BLSTM_NO	TO	31.02	43.94	40.04	53.90	57.89	54.55	46.89
	SO	63.00	71.30	65.10	75.84	79.10	74.57	71.46
	T+S	67.59	73.05	68.63	76.89	79.13	76.74	**73.67**
FruDA	T+S	67.26	65.04	68.59	72.28	64.33	69.00	67.75
TrAdaBoost	T+S	72.77	65.18	64.16	74.24	64.44	65.63	67.74

注: 黑体表示 F 值的最高平均值.

从表 2 的实验结果可以看出: 1) TO 数据的实验结果最差，SO 数据有较大提升，T+S 数据表现最佳.说明使用源领域数据辅助学习能够获得两个领域间相似的数据分布，为生物医学领域和维基百科领域的数据迁移提供了可行性; 2) FruDA 方法和 TrAdaBoost 方法在 T+S 数据的平均识别结果均低于 BLSTM_NO 方法，说明 BLSTM 模型能够更好地学习深层语义信息，帮助模型进行跨领域的模糊限制语识别.

2.3　共享表示方法

表 3 比较了 Sh-pri 模型和我们提出的 Sh-adv 模型在 6 组跨领域中文模糊限制语识别实验的 F 值.另外为验证共享表示方法中判别器的效果，我们去掉了 Sh-pri 模型和 Sh-adv 模型中的判别器模块，进行了 Sh-pri-only和 Sh-only 模型的实验.

表 3　共享表示方法的跨领域中文模糊限制语识别 F 值

Tab.3　F-value of cross-domain Chinese hedge cue detection by shared representation methods　%

方法	数据	组 1	组 2	组 3	组 4	组 5	组 6	平均
Sh-pri-only	T+S	69.19	72.40	70.29	76.79	79.46	77.91	74.34
Sh-pri	T+S	69.63	73.75	70.81	77.16	79.79	77.42	74.76
Sh-only	T+S	70.14	74.94	71.61	76.98	79.32	77.57	75.09
Sh-adv	T+S	72.02	74.38	71.58	76.62	79.76	78.22	**75.43**

注: 黑体表示 F 值的最高平均值.

从表 3 可以看出共享表示方法均好于基线方法，说明共享表示在跨领域中文模糊限制语识别中的有效性.另外，比起带有私有语义表示学习模块的 Sh-pri 模型和 Sh-pri-only 模型，仅使用共享语义表示的 Sh-adv 模型和 Sh-only 模型的识别性能更好.在目标领域训练数据量稀少的情况下，很难学习获得目标领域的私有

语义表示.同时,在整体模型训练中引入源领域私有语义表示学习,会影响目标领域的模糊限制语识别性能.相反从不同领域间抽取共性特征能够更好地实现跨领域模糊限制语识别.在融合对抗训练后,Sh-adv 模型取得了模糊限制语识别实验 75.43%的最高平均 F 值(表 3 中黑体表示),均好于其无对抗训练的模型.但是对抗机制所带来的提升并不明显,其主要原因可能是共享模块试图通过共享参数来保持共有特征的不变,然而目标领域训练数据太少,无法使得私有特征完全从共享表示中剥离,也就无法获得更纯净的源领域和目标领域共享表示.

3 结论与展望

本文提出了一种基于共享表示的跨领域中文模糊限制语识别方法.通过大量的源领域训练数据和少量的目标领域训练数据(200 个),利用对抗学习策略学习源领域和目标领域间的共享语义表示.在生物医学和维基百科领域的实验中,共享表示方法取得了较好的跨领域中文模糊限制语识别性能.本文仅研究了两个领域间的跨领域中文模糊限制识别,如何利用多个源领域的数据,辅助目标领域的模糊限制语识别,是本文下一步的主要研究工作.

参考文献:

[1] LAKOFF G. Hedges: a study in meaning criteria and the logic of fuzzy concepts [J]. Journal of philosophical logic, 1973, 2(4): 458-508.

[2] FARKAS R, VINCZE V, MÓRA G, et al. The CoNLL-2010 shared task: learning to detect hedges and their scope in natural language text [C]// Proceedings of the Fourteenth Conference on Computational Natural Language Learning (CoNLL 2010): Shared Task. Uppsala, 2010: 1-12.

[3] SZARVAS G, VINCZE V, FARKAS R, et al. Cross-genre and cross-domain detection of semantic uncertainty [J]. Computational linguistics, 2012, 38(2): 335-367.

[4] 计峰, 邱锡鹏, 黄萱菁. 中文不确定性句子的识别研究 [C]//第六届全国信息检索学术会议. 哈尔滨, 2010: 594-601.

[5] CHEN Z, ZOU B, ZHU Q, et al. Chinese negation and speculation detection with conditional random fields [C]// Natural Language Processing and Chinese Computing: Second CCF Conference. Chongqing, 2013: 30.

[6] ZOU B, ZHU Q, ZHOU G. Negation and speculation identification in Chinese language [C]// Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Beijing, 2015: 656-665.

[7] DAUMÉ III H. Frustratingly easy domain adaptation [C]// The 45th Annual Meeting of the Association of Computational Linguistics. Prague, 2007: 256-263.

[8] DAI W Y, YANG Q, XUE G R, et al. Boosting for transfer learning [C]// Proceedings of the 24th International Conference on Machine Learning. Corvallis, 2007: 193-200.

[9] YANG Z, SALAKHUTDINOV R, COHEN W W. Transfer learning for sequence tagging with hierarchical recurrent networks [C]// Proceedings of the International Conference on Learning Representations. Toulon, 2017: 1-10.

[10] KIM J K, KIM Y B, SARIKAYA R, et al. Cross-lingual transfer learning for POS tagging without cross-lingual resources [C]// Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, 2017: 2822-2828.

[11] CHEN X, SHI Z, QIU X, et al. Adversarial multi-criteria learning for Chinese word segmentation [C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, 2017: 1193-1203.

[12] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets [C]// Proceedings of the 2014 Conference on Advances in Neural Information Processing Systems. Montréal, 2014: 2672-2680.

[13] ZHOU H, LI X, HUANG D, et al. Exploiting multi-features to detect hedges and their scope in biomedical texts [C]// Proceedings of the Fourteenth Conference on Computational Natural Language Learning: Shared Task. Uppsala, 2010: 106-113.

[14] 周惠巍, 杨欢, 张静, 等. 中文模糊限制语语料库的研究与构建 [J]. 中文信息学报, 2015, 29(6): 83-89.

Cross-domain Chinese Hedge Cue Detection Based on Shared Representations

ZHOU Huiwei[1], NING Shixian[1], YANG Yunlong[1], LIU Zhuang[1], LIN Yingyu[1], LI Sijia[2]
(1. *School of Computer Science and Technology, Dalian University of Technology, Dalian* 116024, *China*;
2. *College of Information and Electrical Engineering, Feng Chia University, Taichung* 40743, *China*)

Abstract: To make full use of out-of-domain data and minimize annotation costs to adapt to a new domain, a novel cross-domain approach based on shared representations was proposed for hedge cue detection. This approach used bidirectional long short-term memory network to alternately learn the training data in the source and target domain by using parameter-sharing mechanism. Meanwhile, it introduced adversarial learning to separate the private features of each domain from the shared features, for the purpose of obtaining the shared semantic representations across different domains. Experiments on Chinese biomedical domain and Wikipedia domain showed that the method based on shared representations could get a significant improvement on cross-domain Chinese hedge cue detection, compared to instance-based transfer learning and feature-based transfer learning methods.

Key words: Chinese hedge cue detection; cross-domain; shared representation; adversarial learning

(责任编辑：王浩毅)

(上接第 28 页)

DDoS Attack Detection Method Based on Combination Correlation Degree and Random Forest

LI Mengyang[1,2], TANG Xiangyan[1,2], CHENG Jieren[1,2,3], LIU Yifu[1,2]
(1. *Key Laboratory of Internet Information Retrieval of Hainan Province, Hainan University, Haikou* 570228, *China*;
2. *College of Information Science and Technology, Hainan University, Haikou* 570228, *China*;
3. *State Key Laboratory of Marine Resource Utilization in South China Sea, Haikou* 570228, *China*)

Abstract: A DDoS attack detection method based on combination correlation and random forest (RF) was proposed. The network flow combination correlation degree (CCD) was defined based on the non-symmetric and the semi-double interaction characterizes of attack flow; and the two tuples form of address correlation statistics (ACS) and unidirectional flow semi interaction (UFSI) was used as the feature of the network flow in CCD. Then the genetic algorithm with the CCD feature sequences was used for the optimization of two key parameters of the decision tree in the RF, namely, the number of maximum trees and the maximum depth of the decision tree. And the RF model within optimized parameters was applied to train the classification model which could be used for the DDoS attack detection. The experiment suggested that the proposed method was suitable for detecting the DDoS attack in big data environment with higher accuracy rate, lower false alarm rate, and missing alarm rate compared with existing DDoS attack detection methods.

Key words: DDoS attack detection; network flow feature extraction; optimization by genetic algorithm; random forest

(责任编辑：王浩毅)