利用词表示和深层神经网络抽取蛋白质关系*

李丽双，蒋振超，万佳，黄德根

（大连理工大学 计算机科学和技术学院，辽宁 大连 116023）

摘要：蛋白质关系抽取是生物医学信息抽取领域的重要分支。目前研究中，基于特征和核函数方法的蛋白质关系抽取已被充分研究，并且达到了很高的 F-值，通过改进特征和核函数进一步优化实例表示变得十分困难。本文结合词表示和深层神经网络，提出了一种实例表示模型。该模型能够充分利用词表示的语义表示能力和深层神经网络的表示优化能力；同时引入主成分分析和特征选择进行特征优化，并且通过比较多种传统的分类器，寻找适合蛋白质关系抽取的分类器。该方法在 AIMed 语料、BioInfer 语料和 HPRD50 语料上的 F-值分别取得了 70.5%、82.2%和 80.0%，在蛋白质关系抽取任务上达到了目前最好的抽取水平。

关键词：蛋白质关系抽取；词表示；深层神经网络

中图分类号：TP391　　**文献标识码**：A

Extracting Protein-Protein Interactions with Word Representation and Deep Neural Network

LI Li-shuang, JIANG Zhen-chao, WAN Jia, HUANG De-gen

(School of Computer Science and Technology, Dalian University of Technology, Dalian, Liaoning 116023, China)

Abstract: Protein-Protein Interaction extraction (PPIE) is a fundamental part of biomedical text mining technology. Most of the current researches on PPI are based on kernels and features that have been achieved higher F-scores, which leave little room for further improvement of instance representation. This paper presents an instance representation model integrating word representation and deep neural network. The model takes advantages of word representation and deep neural network to improve the representation of PPI instances. Meanwhile, the model incorporates feature selection, PCA and different kinds of classifiers, and finds the best combinations for PPI extraction by comparing them. Experimental results show that the method makes great improvement on three public PPI corpora: AIMed, BioInfer, HPRD50, achieving the F-scores of 70.5%, 82.2% and 80%, that is better than other state-of-art methods.

Keywords: Protein-Protein Interaction extraction; word representation; deep neural network

1. 引言

蛋白质是基因表达的产物，承担了大部分的生命活动。研究蛋白质相互作用关系，对于探究生物进程存在的分子体制、分析机体细胞的生命活动具有重要的基础研究意义，进而用以分析疾病的起因，提出针对性的预防和治疗手段。因此如何高效而又准确地从生物医学文本中自动抽取蛋白质关系（Protein-Protein Interaction（PPI））成为生物医学领域文本挖掘的主要任务之一，具有重要的研究意义。

目前，蛋白质关系抽取方法主要包括基于规则的方法和基于统计机器学习的方法。基于规则的方法是利用模式匹配思想，根据已知信息预先制定好详尽的规则，然后进行规则匹配。Yakushiji[1]等人利用解析器生成的谓词参数结构，提出了自动构建特定应用抽取规则的方法。Fundel[2]等人开发了从自由文本中进行关系抽取的 RelEx，其主要思想是利用自然语言预处理产生依存解析树，并结合规则进行关系抽取。基于规则的方法可以取得较高的准确率，但是规则的泛化能力较差，并且规则的定义需要大量的人力物力。

*收稿日期：　　定稿日期：

基金项目：国家自然科学基金项目(61672126，61173101，61173100)

基于统计机器学习方法是目前蛋白质关系抽取的主流方法，其中特征向量和核函数的方法近年来得到了广泛的应用，即利用大量的特征并构建各种核函数表示关系实例。Miwa[3]等人提出了丰富特征向量，融入了词袋、最短路径和图特征，并应用加入语料权重的 SVM 来进行多语料的 PPI 抽取。Tikk[4]等人详细分析了 13 个在 PPI 上使用的核函数之间的差别和共同特征，得出使用相同输入表示的核函数抽取出的蛋白质关系类似，而不同的核函数的组合则能带来性能的提升，并且指出容易抽取错误的蛋白质关系有极少的共同特征，继续在 PPI 抽取上使用基于核函数的方法难以有大的突破。由此，本文将从词表示和深层神经网络入手，不采用大量的人工特征及复杂的核函数，完成蛋白质关系抽取。

词表示能够捕捉词语语义信息，已经作为额外特征或者直接作为输入在许多文本挖掘任务中得到了广泛应用，且已被证实对系统性能具有一定提升作用；而深层神经网络算法能够对原始数据逐层进行表示优化，使得数据表示对分类更有利，从而提升系统性能。但这些技术在 PPI 抽取中尚未得到充分开发利用。Li[5]等在组合核函数中用到了词表示、布朗聚类等词表示技术，Zhao[6]等运用了堆叠自动编码器的深度学习模型，实验表明这些技术能很好地应用于 PPI，由此可见，词表示和深度学习方法在蛋白质关系抽取方面还有广阔的研究空间，有望提升蛋白质关系抽取的性能。

虽然使用特征和核函数的系统在蛋白质关系抽取任务上的性能得到了有效提升，然而，上述系统都没有考虑对特征集合进行优化。很多研究表明特征优化能够更好地选取特征，从而提升数据表示质量。例如，Landeghem[7]等运用了基于信息增益的特征选择方法，Li[8]等运用了半监督的特征耦合泛化（Feature Coupling Generalization，FCG）框架对特征集合进行优化，等等。此外，现有的 PPI 系统大多都采用 SVM 机器学习算法，其优势在过去的 PPI 研究中得到了证实，但任何一种机器学习算法都具有特定的优势和劣势，除了 SVM 之外，其他机器学习算法也应该得到验证。

本文提出了一种蛋白质关系抽取的实例表示模型，该模型首先提取骨架特征，利用词表示的语义表示能力，通过向量组合得到实例表示，然后，采用特征优化策略对输入向量进行优化，最后，通过实验寻找合适的分类器完成分类。此模型在蛋白质关系抽取任务上达到了目前最好的抽取水平。

2 基于实例表示的蛋白质关系抽取模型

2.1 蛋白质关系抽取模型

图 1 为 PPI 抽取模型示意图，主要由三部分构成：实例表示、特征优化和分类器。在实例表示部分，选取蛋白质关系实例的骨架特征（简单的词特征），通过查表将特征转换为词向量，经过向量组合和拼接后得到蛋白质关系的输入向量；随后，采用 PCA 或者特征选择对输入进行优化；最后，选取合适的机器学习算法作为分类器。

2.2 实例表示

实例表示的目的在于将蛋白质关系实例表示为空间向量。传统的做法是利用人工定制的特征集合抽取特征，然后采用独热编码等方式将特征向量转换为空间向量，或通过设计核函数的方式计算实例之间的内积。而本文仅仅利用骨架特征[9]和其对应的词表示将蛋白质关系实例表示为空间向量。本文使用的实例表示方法分四步：

第一步，提取骨架特征。传统的特征往往经过加工，而骨架特征只包含基础的词语特征，以经过切词之后的文本“N - SH2 and SH3 + N - SH2 interact only with IR beta .”为例，骨架特征包括：

a.蛋白质词语特征。本例中目标蛋白质对为“SH3 + N – SH2”和“IR beta”，蛋白质词语特征分别为[SH3, +, N, -, SH2]和[IR, beta]。

b.蛋白质周围词特征。选取到目标蛋白质对距离小于 r 的词作为周围词特征。在本例中当 r=1 时，周围词特征分别是[and, interact]和[with, .]。

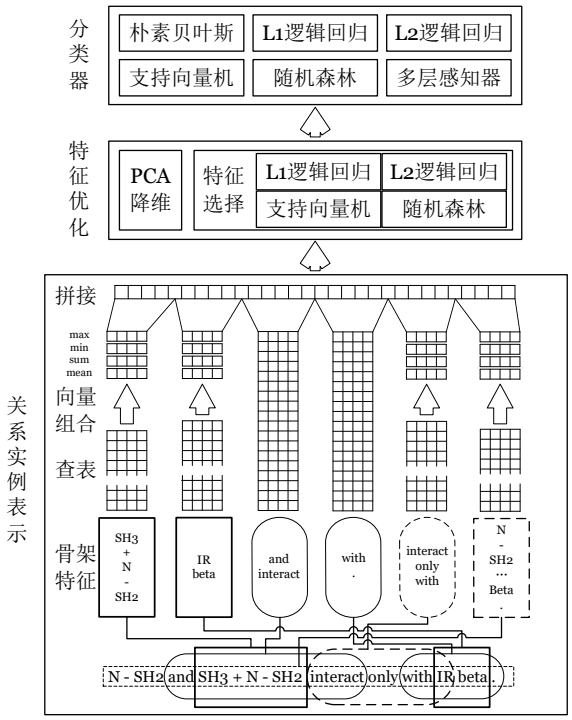


图 1　蛋白质关系抽取模型

c.蛋白质中间词特征。蛋白质之间的词往往预示着蛋白质对之间是否存在关系，如在该例句中，[interact, only, with]预示了两个蛋白质之间存在交互关系。

d.句中所有词特征。在例句中，所有词特征为[N, -, SH2, and, SH3, +, N, -, SH2, interact, only, with, IR, beta, .]。

第二步，查表，即在抽取出骨架特征之后，利用词表示将词语转换为向量。词表示由 Skip-gram、CBOW、GloVe 和 ELB[10]等模型训练得到。由于骨架特征均是词语特征，因此，骨架特征经过查表后均可转换为词表示。

第三步，向量组合。在骨架特征中，除了周围词特征由 r 固定长度之外，其他特征长度均是不定的，因此需要对除周围词特征之外的骨架特征进行统一。以中间词特征为例，经过查表之后[interact, only, with]分别对应词向量$C(interact)$，$C(only)$**错误!未找到引用源。**，和$C(with)$。采用四种常见的向量组合方法，分别取$C(interact)$，$C(only)$和$C(with)$每一维度取值的和、平均值、最大值和最小值，得到 4 个组合向量，其维度与词向量维度一致。除了取平均之外，每种组合方式的分布与原始词向量的分布不同，因此四种向量组合方式在信息表示上不会产生冗余。

第四步，向量拼接。经过向量组合后拼接所有特征对应的向量，得到最终的输入向量。经过骨架特征提取、查表、向量组合和拼接等步骤之后，关系实例就转换成了空间中的向量。至此便完成了关系实例的初步表示，接下来将对该表示进行优化。

2.3 特征优化

本文在特征优化阶段使用了主成分分析（Principal component analysis（PCA））和特征选择，这两种方法是运用最广泛的两类特征优化方法。PCA 是一种分析、简化数据集的技术，在降低维度的同时保持数据集中对方差贡献最大的特征。

特征选择是指从特征集合中选取一个子集的过程，本文采用的特征选择方法有 L1 逻辑回归、L2 逻辑回归、SVM 和随机森林四种基于模型的特征选择方法。

L1 逻辑回归和 L2 逻辑回归是在逻辑回归的损失函数中分别加入 L1 正则化项和 L2 正则化项，通过对正则项的引入减少模型的过拟合。基于 L1 逻辑回归的特征选择提供了较好的

解释性，但在关联特征存在的情况下稳定性较差，而 L2 逻辑回归刚好相反，对关联特征的评价具有较好的稳定性，但特征之间得分往往较为接近，导致特征的区分度相对较弱。SVM 算法思想是使距离分类超平面最近的样本与分类超平面的距离尽可能远，即间隔最大化原则，其符合结构风险最小化原则。SVM 在许多自然语言处理任务中的表现非常出色，因此，SVM 作为分类器和特征选择工具都具有十分重要的研究价值。随机森林是在决策树算法基础上应用自举汇聚法（bootstrap aggregating）得到的。使用随机森林进行特征选择时，采用构建决策树过程中特征的信息增益作为特征的得分。与逻辑回归和 SVM 的模型参数不同，这种方式从信息论的角度来衡量特征的重要度，对前几种特征选择方法形成了较好的补充。

2.4 分类器

本文总共采用 6 种分类器：朴素贝叶斯、L1 逻辑回归、L2 逻辑回归、支持向量机、随机森林和多层感知机。朴素贝叶斯分类器是在假设特征之间独立的条件下运用贝叶斯定理的分类器，对于给出的待分类项，求解在此项出现的条件下各个类别出现的概率，将概率最大的类别确定为此项的类别；逻辑回归分类器是是一种常见的广义线性分类器，在逻辑回归的损失函数中分别引入 L1、L2 正则项即可得到 L1 逻辑回归和 L2 逻辑回归分类器；支持向量机是一种基于最大间隔思想的分类器，目前在蛋白质关系抽取中应用最为广泛；随机森林是在决策树基础上进行扩展的方法，能够处理高维度的数据，适合做分类问题；多层感知机是一种前向结构的人工神经网络，映射一组输入向量到一组输出向量。可以被看作是一个有向图，由多个节点层所组成，每一层都全连接到下一层。除了输入节点，每个节点都是一个带有非线性激活函数的神经元。神经网络包含神经元的一个或多个隐层，并且隐层的节点使神经网络从输入模式中不断获取有意义的特性，从而学会处理高度非线性的复杂任务。

3. 实验与分析

3.1 语料及评价方法

本文在五个公共开放的蛋白质关系抽取任务上进行了实验。该任务包含 5 个公共的语料：AIMed[11]、BioInfer[12]、HPRD50[2]、IEPA[13]和 LLL[14]。蛋白质关系抽取的性能通过十倍交叉验证的方式进行评价。目前主流的十倍交叉验证方法分为文档级和实例级两种方式，而文档级的十倍交叉验证极易受到如何切分的影响，这一点在 AIMed 语料上尤为明显，在 5 个语料中，AIMed 语料实例数与文档数的比例是最大的，在这种情况下，当采用文档级切分时，AIMed 语料上取得的 F 值很容易受到切分方式的影响。因此，本文采用句子级十倍交叉验证计算 F 值的方式对模型进行评价。

3.2 实验结果及分析

3.2.1 词表示和分类器对系统的影响

首先，利用独热编码完成实例表示，即图 1 查表步骤中采用的向量是独热向量，并采用 SVM 分类器在 5 个蛋白质关系抽取语料上进行实验，以此作为基线系统。其次，利用 4 种词表示方法和 6 种分类器的组合进行实验。实验结果如表 1 所示。

从表 1 的对比实验中可以分析得出如下结论：

首先，词表示比独热编码的数据表示方法更有优势。在所有 5 个语料上，同样采用 SVM 作为分类器，ELB、Skip-gram 和 CBOW 均比独热编码取得了更高的 F 值，这验证了词表示方法能够捕捉词语语义信息的能力，证实了通过引入词表示可以有效提高系统性能。同时，实验中所有词向量的训练语料是从 MEDLINE 中挑选的与蛋白质相关的 50000 篇文献，窗口大小均为 5，词向量维度均是 400。从表 1 中可以看出，ELB 在 3 个语料上取得了较好的效果，而在 HPRD50 和 LLL 上，Skip-gram 和 GloVe 分别取得了最好的效果。

第二，分类器对模型具有很大影响。尽管可以从理论上分析每一种分类器各自的优缺点和适用场景，但从表 1 中可以看出，在 6 种分类器中，L2 逻辑回归、支持向量机和多层感知机在 PPIE 任务上是效果相对较好的。尽管 L1 逻辑回归和 L2 逻辑回归区别仅仅在于正则

项的不同，最终取得的 F 值也有较为明显的差距。另一方面，虽然现有大多数 PPIE 方法基

表 1 词表示和分类器对系统的影响

词表示+分类器	AIMed	BioInfer	HPRD50	IEPA	LLL
one hot+支持向量机	56.8	71.6	66.6	65.1	74.8
Skip-gram+朴素贝叶斯	37.5	53.3	66.2	66.4	66.1
CBOW+朴素贝叶斯	37.9	53.0	67.0	65.0	57.9
GloVe+朴素贝叶斯	36.7	49.6	60.8	65.2	66.2
ELB+朴素贝叶斯	38.2	52.5	67.7	65.8	70.5
Skip-gram+L1 逻辑回归	59.5	69.3	72.2	71.0	78.5
CBOW+L1 逻辑回归	59.1	69.0	72.2	70.5	78.6
GloVe+L1 逻辑回归	44.8	57.6	67.0	66.2	74.9
ELB+L1 逻辑回归	63.2	75.0	68.7	69.7	77.3
Skip-gram+L2 逻辑回归	63.5	71.3	76.3	72.6	80.4
CBOW+L2 逻辑回归	63.3	70.8	76.6	71.4	81.9
GloVe+L2 逻辑回归	51.1	61.7	68.0	68.4	78.4
ELB+L2 逻辑回归	66.9	76.6	79.1	72.8	83.7
Skip-gram+随机森林	44.1	70.9	65.9	65.6	77.2
CBOW+随机森林	47.0	69.7	68.6	66.2	77.2
GloVe+随机森林	38.4	68.9	57.7	65.4	80.0
ELB+随机森林	45.4	70.1	70.5	70.0	78.7
Skip-gram+支持向量机	64.4	74.3	**77.6***	71.5	79.8
CBOW+支持向量机	66.7	73.9	76.1	70.1	82.3
GloVe+支持向量机	56.3	65.2	71.0	70.4	75.6
ELB+支持向量机	66.7	76.8	75.7	72.7	82.7
Skip-gram+MLP	64.0	74.0	75.2	73.8	73.5
CBOW+MLP	64.1	77.9	75.1	70.3	**85.2***
GloVe+MLP	23.1	51.1	39.7	56.8	57.2
ELB+MLP	**68.0***	**80.0***	66.6	**74.2***	81.5

于支持向量机，但 L2 逻辑回归、多层感知机也达到了与支持向量机相当甚至更好的水平，因此，针对具体语料选择合适的分类器具有极为现实的意义。

第三，不同语料上的最优模型不同。当固定分类器时，词表示在不同语料上的性能是不同的，如采用随机森林分类器时，五个语料上的最优词表示模型分别是：CBOW、Skip-gram、ELB、ELB和GloVE;而当固定词表示时，不同分类器的性能也不同。考虑到AIMed和BioInfer规模比其它语料大很多，本文着重考虑这两个语料上的表现。由表 1 可见，ELB 和多层感知机的组合在 AIMed、BioInfer、和 IEPA 上取得了最高的 F 值，在 LLL 上也取得了较高的 F 值，综合表现较好。

3.2.2 神经网络隐层对系统的影响

多层感知机相对其他分类器具有更多可调参数，并且在三个语料上取得了最高的 F 值，因此多层感知机具有进一步挖掘的空间。然而，如何选择参数是一项极具挑战的任务。Bengio 在 Reddit 机器学习版块的“Ask Me AnyThing”问答活动中指出，隐含层的数量应该是 1 到 3，每一层的隐含单元数目应该是 50 到 5000，给出了隐层参数的大概参数范围，但针对 PPIE 这个具体的问题，仍然需要通过实验来寻找更好的隐层参数。除了隐层参数不同之外，在本文出现的所有神经网络中，激活函数均采用 relu，优化方式均为 adam，均加入 L2 正则项，表 2 列出了在多层感知机上进一步实验结果。

通过表 2 可以得出如下结论：

表 2 多层感知机隐层参数对系统的影响

词表示+隐层参数	AIMed	BioInfer	HPRD50	IEPA	LLL
Skip-gram+MLP[]	55.7	62.6	75.1	73.3	80.7
CBOW+MLP[]	57.8	64.6	74.1	70.5	82.2
GloVe+MLP[]	24.7	39.8	52.0	64.1	75.6
ELB+MLP[]	63.0	72.0	74.8	73.3	81.5
Skip-gram+MLP[100]	64.0	74.0	**75.2***	73.8	73.5
CBOW+MLP[100]	64.1	77.9	75.1	70.3	**85.2***
GloVe+MLP[100]	23.1	51.1	39.7	56.8	57.2
ELB+MLP[100]	68.0	80.0	66.6	74.2	81.5
Skip-gram+MLP[1000]	67.1	81.0	54.3	**74.3***	65.6
CBOW+ MLP [1000]	69.0	81.9	74.8	70.1	77.0
GloVe+MLP[1000]	3.0	36.3	45.4	65.0	55.3
ELB+ MLP[1000]	**70.5***	**82.2***	60.9	73.8	78.2
Skip-gram + MLP[5000, 1000]	53.8	67.9	48.3	47.7	68.9
CBOW + MLP[5000, 1000]	52.5	78.4	53.2	68.5	78.0
GloVe + MLP[5000, 1000]	0	11.2	6.5	24.7	39.9
ELB+ MLP[5000, 1000]	43.8	79.6	29.3	42.3	65.1
Skip-gram+ MLP[1000, 50]	25.5	68.5	34.6	70.9	54.9
CBOW+ MLP[1000, 50]	57.8	69.4	73.0	61.7	80.5
GloVe+ MLP[1000, 50]	0.6	22.6	21.4	24.8	47.5
ELB+ MLP[1000, 50]	26.2	69.9	29.7	44.6	68.8
Skip-gram+ MLP[1000, 500, 50]	54.9	76.6	66.4	60.6	75.5
CBOW+ MLP[1000, 500, 50]	61.9	75.0	74.5	63.3	74.9
GloVe+ MLP[1000, 500, 50]	0	31.1	42.4	28.6	40.9
ELB+ MLP[1000, 500, 50]	50.2	78.0	51.4	64.6	69.3
Skip-gram+ MLP[1000, 500, 200, 50]	64.2	74.3	59.0	56.5	81.1
CBOW+ MLP[1000, 500, 200, 50]	62.5	76.3	73.8	69.3	79.9
GloVe+ MLP[1000, 500, 200, 50]	0	31.6	46.8	51.3	43.7
ELB+ MLP[1000, 500, 200, 50]	65.9	77.1	66.5	66.1	80.7
Skip-gram+ MLP[1000, 500, 200, 100, 50]	61.5	75.5	56.0	70.8	79.2
CBOW+ MLP[1000, 500, 200, 100, 50]	62.3	77.9	73.1	68.4	79.8
GloVe+ MLP[1000, 500, 200, 100, 50]	0	42.0	34.3	54.7	52.8
ELB+ MLP[1000, 500, 200, 100, 50]	60.2	78.1	61.2	64.2	66.5

首先，隐层有可能有助于提升系统性能，也有可能起到相反的作用。以 AIMed 语料和 Skip-gram 词表示模型为例，当没有隐层时 F 值为 55.7%，当采用 1 个、4 个、5 个隐层时，能取得更高的 F 值，而当采用 2 个、3 个隐层时，F 值反而会下降。整个神经网络可视作对数据的非线性变换，通过改变隐层数可以改变神经网络的非线性变换效果，但能否对系统带来提升是难以通过理论分析预知的，需要通过实验才能验证。

其次，隐层的个数为 1 时系统性能较好。尽管深度学习的初衷是通过逐层的抽象和优化来得到更好的表示，但从实际的实验效果上来看，有时候更多的隐层效果未必更好，例如 ELB+[1000]的 F 值要好过 ELB+[5000, 1000]、ELB+[1000, 50]，ELB+[1000, 500, 50]、ELB+[1000, 500, 200, 50]和[1000, 500, 200, 100, 50]。尽管从[1000, 50]到[1000, 500, 200, 50]，

F 值呈递增的趋势，但[1000, 500, 200, 100, 50]的 F 值比[1000, 500, 200, 50]又有所下降。因

表 3 特征优化对系统的影响

特征优化+分类器	AIMed	BioInfer	HPRD50	IEPA	LLL
朴素贝叶斯	38.2	52.5	67.7	65.8	70.5
主成份分析+朴素贝叶斯	2.2	48.4	21.5	0	1.1
L1 逻辑回归+朴素贝叶斯	33.9	50.5	64.5	65.2	68.7
L2 逻辑回归+朴素贝叶斯	37.0	53.1	65.7	67.7	75.1
随机森林+朴素贝叶斯	39.1	53.5	67.2	65.5	73.3
支持向量机+朴素贝叶斯	36.8	53.2	66.3	67.2	75.8
L1 逻辑回归	63.2	75.0	68.7	69.7	77.3
主成份分析+ L1 逻辑回归	63.3	75.0	76.0	72.3	80.5
L1 逻辑回归+L1 逻辑回归	62.9	74.9	68.6	69.8	77.9
L2 逻辑回归+L1 逻辑回归	62.9	74.8	68.6	69.3	77.6
随机森林+L1 逻辑回归	57.5	69.9	67.5	70.8	78.7
支持向量机+L1 逻辑回归	62.9	74.9	69.5	69.4	77.4
L2 逻辑回归	66.9	76.6	79.1	72.8	83.7
主成份分析+ L2 逻辑回归	67.2	76.6	**80.0***	72.7	83.4
L1 逻辑回归+L2 逻辑回归	62.7	75.2	70.0	66.6	77.2
L2 逻辑回归+ L2 逻辑回归	66.1	76.3	80.0	71.5	82.6
随机森林+L2 逻辑回归	61.4	70.8	69.9	69.2	79.8
支持向量机+L2 逻辑回归	67.1	76.4	**80.0***	71.8	83.0
随机森林	45.4	70.1	70.5	70.0	78.7
主成份分析+随机森林	31.1	38.9	47.2	29.1	30.3
L1 逻辑回归+随机森林	41.6	69.6	67.6	64.9	76.2
L2 逻辑回归+随机森林	45.0	70.4	65.2	64.7	83.8
随机森林+随机森林	42.9	68.7	66.5	66.6	75.2
支持向量机+随机森林	42.8	69.7	69.0	65.3	77.8
支持向量机	66.7	76.8	75.7	72.7	82.7
主成份分析+支持向量机	66.7	76.8	75.7	72.7	82.7
L1 逻辑回归+支持向量机	64.5	74.9	70.1	67.9	79.2
L2 逻辑回归+支持向量机	66.2	76.9	77.7	73.6	**84.1***
随机森林+支持向量机	63.4	71.9	71.3	70.1	79.9
支持向量机+支持向量机	67.3	76.4	76.1	72.4	82.9
MLP[1000]	**70.5***	**82.2***	60.9	73.8	78.2
主成份分析+MLP[1000]	67.5	79.7	79.7	**75.5***	82.0
L1 逻辑回归+MLP[1000]	66.7	79.0	72.6	69.2	77.2
L2 逻辑回归+MLP[1000]	68.5	81.0	61.5	73.7	74.7
随机森林+MLP[1000]	63.5	81.4	70.8	68.6	79.6
支持向量机+MLP[1000]	68.4	81.1	67.4	72.1	80.5

此，隐层个数具体取多少最好，很难一概而论，需要通过实验验证才能得出结论，而在 PPIE 的实验上发现，当隐层为 1 时多层感知机在五个语料上取得了较好的表现。

第三，隐层节点数对性能有一定影响。例如，ELB+[1000]在 5 个 PPIE 语料上分别取得了 70.5%、82.2%、60.9%、73.8%和 78.2%的 F 值，比 ELB+[100]分别高出了 2.5、2.2、-5.7、-0.4、-3.3 个百分点，因此，在不同的语料上，最优隐层节点个数是不同的。即便采用同

样的预处理、特征选取、向量组合方法，同样的隐层节点个数对不同的语料会产生不同的影响，在其原因可能在于语料的分布不同。

第四，隐层个数比隐层节点数的影响可能更大。从表 2 的实验结果看出，在 5 个语料上性能最好的神经网络均只含一个隐层，而不论节点数为 100 还是 1000，都取得了比两个及以上隐层更好的效果，因此，隐层个数对系统的影响可能更大。

3.2.3PCA、特征选择表示优化能力对比

在传统的机器学习方法中，PCA 和特征选择均属于特征优化的方法，因此，本文通过在朴素贝叶斯、L1 逻辑回归、L2 逻辑回归、随机森林、支持向量机和多层感知机之前，引入

特征优化模块，以对比 PCA 和特征选择的特征优化能力。

表 3 对比了特征优化对系统性能的影响，选用的词向量由 ELB 模型训练得到，从中可以得出如下结论：

首先，神经网络的表示能力在 AIMed 和 BioInfer 两个语料上最为明显。PCA、4 种特征选择方法与 6 种分类器共计 30 种组合，在这两个语料上的 F 值都没有超过单隐层多层感知机，这说明了神经网络的表示优化能力在某些语料上可以与传统的基于 PCA 和特征选择相抗衡。

其次，在 HPRD50、IEPA 和 LLL 这三个较小的语料上，PCA 和特征选择能对多层感知机起到增强作用。例如，在单隐层的基础上引入 PCA 后，在这 3 个语料上的 F 值均有进一步提升，分别从 60.9%、73.8%和 78.2%提升到了 79.7%、75.5%和 82.0%，其他特征选择算法也起到了一定的提升作用。

第三，PCA 和特征选择有时候会对性能起提升作用，但有时候也会起相反的作用。这样的证据很容易从表 3 中找到，例如加入 PCA 的 L2 逻辑回归，在 AIMed 上从 66.9 提升到了 67.2，但在 LLL 上从 83.7 下降到了 83.4。

3.2.4 与其他方法的比较

表 4 对比了本文与其它实例级十倍交叉验证 PPIE 方法的性能。从中分析发现：

首先，表 4 列出的大多数方法仅在 AIMed 语料上进行了验证，可以看出，本文的方法

表 4 与其他方法对比

实验	AIMed	BioInfer	HPRD50	IEPA	LLL
Yakushiji[1]	57.3	-	-	-	-
Giuliano[15]	63.9	-	-	-	-
Mitsumori[16]	54.3	-	-	-	-
Erkan[17]	60.0	-	-	-	-
Katrenko[18]	54.3	-	-	-	-
Fundel[2]	-	-	-	-	82
Sætre[19]	69.5	-	-	-	-
Landeghem[7]	62	-	-	-	82
Fayruzov[20]	45	-	-	-	78
Yu[21]	69.5	-	-	-	-
Li[22]	63.2	-	-	-	-
Li[23]	69.5	-	-	-	-
Li[5]	69.7	74.0	78.0	**76.5***	**87.3***
ELB+ MLP [1000]	**70.5***	**82.2***	60.9	73.8	78.2
支持向量机+L2 逻辑回归	67.1	76.4	**80.0***	71.8	83.0
主成份分析+MLP[1000]	67.5	79.7	79.7	75.5	82.0

CBOW+ MLP [100]	64.1	77.9	75.1	70.3	85.2

70.5%的 F 值是所有方法中最高的，从而证实了通过发挥词向量的语义表示能力和神经网络的表示优化能力，可以取得比传统的基于特征、核函数方法更好的效果。

其次，在表 4 中，Li 等[5]通过融合 Skip-gram 词表示、布朗聚类和树核构建组合核 SVM 分类器，已经在 5 个 PPIE 语料上取得了很好的效果，比其他过去的方法具有更高的 F 值。本文所采用的基于多层感知机的方法在 AIMed 和 BioInfer 这两个相对较大的语料上比 Li 等[5]的方法有所提升，分别提高了 0.8 和 8.2 个百分点，当采用支持向量机作特征优化、以 L2 逻辑回归为分类器时，在 HPRD50 语料上也取得了较好的结果，但在 IEPA 和 LLL 这两个语料上的性能稍低 Li 等[5]的方法。这可能是因为神经网络的表示优化能力受到语料规模的影响，在规模较大的语料上经过充分训练后才能取得较好的效果，而树核、布朗聚类和词表示的组合核函数属于人工表示，不依赖于训练数据规模，因此在小语料上 Li 等[5]的方法具有更好的表现。

4. 结论

本文针对蛋白质间关系抽取问题，提出了一种实例表示模型，并对该模型进行了验证，充分考虑了 4 种词表示模型（Skip-gram、CBOW、GloVe 和 BLE）、PCA、4 种特征选择算法（L1 逻辑回归、L2 逻辑回归、随机森林和支持向量机）和 6 种分类器（朴素贝叶斯、L1 逻辑回归、L2 逻辑回归、随机森林、支持向量机和多层感知机）在不同组合下的性能，从中得出的主要结论包括：

（1）通过实验验证了借助词表示和深层神经网络的表达能力可以取得比传统的基于特征或核函数的方法更好的效果。尽管本文针对词表示和多层感知机做了较为充分的实验，但词表示和深度学习仍有广阔的研究空间，例如卷积神经网络、循环神经网络等深度学习框架已经在相关领域取得了一定成就，因此，本文为将来更广泛的基于深度学习和词表示的关系抽取研究提供了可行性依据。

（2）深层神经网络对隐层较为敏感。当隐层设置得当时，神经网络比其他机器学习具有较为明显的优势；但当隐层设置不合理时，反而会导致性能的急剧下降。在蛋白质关系抽取任务上，隐层的层数并非越多越好，更多的层数不意味着更好的性能；但相比隐层节点个数，隐层个数的选择相对更为重要。

（3）特征优化对 PPI 抽取性能不总是起提升作用，有时候可能会起相反的作用，并且神经网络的表示优化能力在某些语料上可以与传统的基于 PCA 和特征选择相抗衡。

（4）针对不同的语料，其最佳模型是不同的，应当针对具体的语料来设计和定制模型。即便同样是蛋白质关系抽取任务，很难找到能够同时提升 5 个语料的抽取性能的模型。

本文提出并验证了实例表示模型，在蛋白质关系抽取任务上取得了较好的效果。正如上文第四点所述，想要找到一款泛化性很强的模型，能同时运用于多个语料或任务是非常困难的，针对具体任务设计模型以提升抽取性能是较为现实的途径。

参考文献：

[1] Yakushiji A, Miyao Y, Tateisi Y, et al. Biomedical information extraction with predicate-argument structure patterns[C]//Proceedings of the first International Symposium on Semantic Mining in Biomedicine. 2005: 60-69.

[2] Fundel K, Küffner R, Zimmer R. RelEx—Relation extraction using dependency parse trees[J]. Bioinformatics, 2007, 23(3): 365-371.

[3] Miwa M, Sætre R, Miyao Y, et al. A rich feature vector for protein-protein interaction extraction from multiple corpora[C]//Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2009: 121-130.

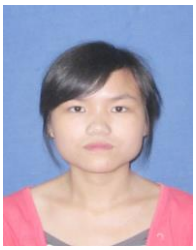
[4] Tikk D, Solt I, Thomas P, et al. A detailed error analysis of 13 kernel methods for protein–protein interaction extraction[J]. BMC bioinformatics, 2013, 14(1): 1-20.

[5] Li L, Guo R, Jiang Z, et al. Improving Kernel-based protein-protein interaction extraction by unsupervised word representation[C]//2014 IEEE International Conference on Bioinformatics and Biomedicine. 2014: 379-384.
[6] Zhao Z, Yang Z, Luo L, et al. Deep neural network based protein-protein interaction extraction from biomedical literature[C]//2015 IEEE International Conference on Bioinformatics and Biomedicine. 2015: 1156-1156.
[7] Van Landeghem S, Saeys Y, De Baets B, et al. Extracting protein-protein interactions from text using rich feature vectors and feature selection[C]//3rd International symposium on Semantic Mining in Biomedicine. Turku Centre for Computer Sciences, 2008: 77-84.
[8] Li Y, Lin H, Yang Z. Applying feature coupling generalization for protein-protein interaction extraction[C]//2009 IEEE International Conference on Bioinformatics and Biomedicine.2009: 396-400.
[9] Li L, Jiang Z, Huang D. A general instance representation architecture for protein-protein interaction extraction[C]//2014 IEEE International Conference on Bioinformatics and Biomedicine. 2014: 497-500.
[10] Jiang Z, Li L, Huang D, et al. Training word embeddings for deep learning in biomedical text mining tasks[C]//2015 IEEE International Conference on Bioinformatics and Biomedicine. 2015: 625-628.
[11] Bunescu R, Ge R, Kate R J, et al. Comparative experiments on learning information extractors for proteins and their interactions[J]. Artificial intelligence in medicine, 2005, 33(2): 139-155.
[12] Pyysalo S, Ginter F, Heimonen J, et al. BioInfer: a corpus for information extraction in the biomedical domain[J]. BMC bioinformatics, 2007, 8(1): 50.
[13] Ding J, Berleant D, Nettleton D, et al. Mining MEDLINE: abstracts, sentences, or phrases[C]//Proceedings of the pacific symposium on biocomputing. 2002: 326-337.
[14] Nédellec C. Learning language in logic-genic interaction extraction challenge[C]// Proceedings of the 4th Learning Language in Logic Workshop. 2005: 1-7.
[15] Giuliano C, Lavelli A, Romano L. Exploiting shallow linguistic information for relation extraction from biomedical literature[C]//The 11th Conference of the European Chapter of the Association for Computational Linguistics. 2006: 401-408.
[16] Mitsumori T, Murata M, Fukuda Y, et al. Extracting protein-protein interaction information from biomedical text with SVM[J]. IEICE Transactions on Information and Systems, 2006, 89(8): 2464-2466.
[17] Erkan G, Ozgur A, Radev D R. Extracting interacting protein pairs and evidence sentences by using dependency parsing and machine learning techniques[C]//Proceedings of the Second BioCreative Challenge Workshop. 2007: 2-8.
[18] Katrenko S, Adriaans P. Learning relations from biomedical corpora using dependency trees[C]// Knowledge Discovery and Emergent Complexity in Bioinformatics first International Workshop. 2006:61-80.
[19] Sætre R, Sagae K, Tsujii J I. Syntactic features for protein-protein interaction extraction[C]// Proceedings of the International Symposium on Languages in Biology and Medicine. 2007:1-15.
[20] Fayruzov T, De Cock M, Cornelis C, et al. DEEPER: a full parsing based approach to protein relation extraction[C]//European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics. 2008: 36-47.
[21] Yu H, Qian L, Zhou G, et al. Extracting protein-protein interaction from biomedical text using additional shallow parsing information[C]//2009 IEEE International Conference on Bioinformatics and Biomedicine. 2009: 1-5.

[22] 李丽双，刘洋，黄德根．基于组合核的蛋白质交互关系抽取[J]．中文信息学报，2013，27(1)：86-93.
[23] Li L, Zhang P, Zheng T, et al. Integrating semantic information into multiple kernels for protein-protein interaction extraction from biomedical literatures[J]. PloS one, 2014, 9(3): 28-47.

李丽双（1967 年--），女，教授，博士生导师，主要研究方向为自然语言处理、信息抽取与机器翻译。Email：lils@dlut.edu.cn，通讯作者

蒋振超，（1988 年--），男，博士研究生，主要研究方向为自然语言处理。Email：jzc_nlp@163.com



万佳，（1992 年--），女，硕士研究生，主要研究方向为自然语言处理。Email：1725799902@qq.com

黄德根，（1965 年--），男，教授，博士生导师，主要研究方向为自然语言理解与机器翻译。Email：huangdg@dlut.edu.cn