基于耦合关系模型的文本分类研究

孙劲光[1] 全纹敬[2]
(辽宁工程技术大学电子与信息工程学院 葫芦岛 125105)[1]
(辽宁工程技术大学研究生学院 葫芦岛 125105)[2]

摘 要 对文本的特征提取方法以及深度神经网络的分类器的搭建进行研究。首先,在全局和局部的特征提取方法的基础上,通过对文本特征内耦合关系和文本特征间耦合关系进行分析,确定用于分类的文本特征,建立文本特征的耦合关系模型;其次,将文本特征作为深度神经网络输入层进行分类;最后,通过逐层无监督的方式对网络进行训练,在顶层增加区分性结点来实现文本分类功能。

关键词 耦合关系,深度学习,深度置信网络,文本分类
中图法分类号 TP181 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.8.055

Research on Coupling Model of Text Classification

SUN Jin-guang[1] QUAN Wen-jing[2]
(School of Electronics and Information Engineering, Liaoning Technical University, Huludao 125105, China)[1]
(Institute of Graduate, Liaoning Technical University, Huludao 125105, China)[2]

Abstract In this paper, we discussed the main depth neural network classifier feature and the extraction method of text feature. First of all, the coupling relationships between text features and in text features are analyzed to build textual features for classification and establish the coupling relation model of text feature. Secondly, the text feature is classified as the depth of the neural network input layer. Finally, the network is trained without supervision, increasing distinct nodes on the top floor to achieve the function of text classification.

Keywords Coupling relationship, Deep learning, Deep belief nets, Text classification

1 引言

当今网络平台的迅猛发展而产生的大量非结构化信息正以指数级增长。大量数据的存在形式繁多,互联网及各行业领域的信息以文本形式展现。如何有效地对这些海量文档加以有序的管理,提高数据的利用率并使其产生价值,是文本分类方法研究的主要目的之一。文本分类技术作为处理海量文本数据信息的有效手段,可以对其进行较为精准的管理与定位,节约大量的人力及物力,已在数据信息过滤、信息组织管理及定位、网页分类和数字图书馆等领域得到广泛应用。因此,开展文本分类技术的研究具有非常重要的现实意义。

20世纪90年代开始,以机器学习为主的文本自动分类方法逐渐占据主要地位。基于机器学习的自动文本分类方法由于不需要知识工程师和相关领域专家的参与而拥有较好的移植性,开放速度更快,推广性能更好。文本分类领域中逐渐引入其他重要的、关键的机器学习方法,例如K最近邻[2]、决策树[3]、贝叶斯[4]、线性分类器[5]、神经网络[6]、最小二乘拟、回归模型[7]以及支持向量机等。其中,支持向量机方法由Vapnik[8]提出,它运用最小化结构风险的原则,实现了有限样本情况下较好的分类器设计。支持向量机通过构造一个最佳的分类超平面被应用到文本分类中[9],并取得了很好的分类性能。2008年,NEC Labs America研究员Collobert和Weston将embedding和多层一维卷积的结构用于词性标注、语句分块、命名实体识别以及语义角色标注4个典型NLP问题[10];他们将同一个模型用于不同任务,都取得了与经典算法相当的准确率。2009年,Salakhutdinov等提出的Semantic Hashing方法将深度学习应用于文本理解和分析领域[11];提出通过DBN模型将基于词频统计的文本特征向量映射到新的低维语义空间中,在语义特征空间中距离相近的点所对应的文本在内容上更加相关,并在大规模的文本重建训练过程中模型隐式地获得了词语之间的语义联系,使特征的映射更趋于合理。2011年,斯坦福大学教授Manning等人将深度学习用于自然语言和自然风景图片的融合[12]。在文本挖掘领域中,哈尔滨工业大学的王晓龙教授研究组是国内较早开展深度学习研究的,他们巧妙地将深度置信网络应用于网络社区问答对的语义挖掘中[13]。

本文的主要贡献有:(1)通过对文本特征内耦合关系和特征间耦合关系进行分析,确定用于分类的文本特征,建立文本特征的耦合关系模型;(2)采用基于耦合关系模型的特征权值计算方法得到文本表示,并将其作为深度神经网络输入层进

到稿日期:2015-04-24 返修日期:2015-08-02 本文受国家科技支撑项目(2013BAH12F02)资助。
孙劲光(1962—),女,博士,教授,主要研究方向为图形理论与技术、图像工程、知识工程,E-mail:sunjinguang@lntu.edu.cn;**全纹敬**(1990—),女,硕士,主要研究方向为数据挖掘,E-mail:297404299@qq.com(通信作者)。

行分类;(3)通过逐层无监督的方式对网络进行训练,在顶层增加区分性结点以实现文本分类功能。

深度置信网络文本分类过程如图 1 所示。

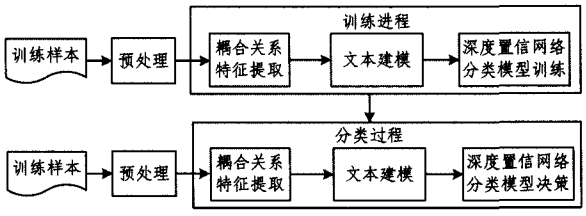

图 1 深度置信网络文本分类过程

2 文本特征耦合关系模型

文本特征提取方法中普遍采用整体或局部频率的特征提取方法[1]。首先,去除语料库中普遍存在的、对文本类别判定权重较低的噪声特征;其次,去除出现频率过高与过低的、对文本类别判定发挥较低作用的特征。在整个过程中,除了词频 TF 和文档频率 DF,还需要计算用于全局特征提取的整体频率 OF 和用于局部特征提取的类别频率 CF。这些提取方法往往很少,甚至没有充分考虑文本特征之间的相互关联关系,从而造成了分类器性能低等问题。对于特征之间存在的非线性关系,本文采用耦合相似度度量方法,通过建立文本特征耦合关系模型,更加全面地捕获文本间的关联关系[14,15]。

2.1 文本特征的整体及局部频率

文本特征的整体频率:在整个训练语料库中(Training Corpus,TC),特征 t 的频率被称作整体频率(Overall Frequency,OF),定义为

$$OF(t)=\frac{n(t,TC)}{N_{TC}} \tag{1}$$

其中,$n(t,TC)$表示 t 出现在 TC 中的次数,N_{TC}表示 TC 中所有词的个数。

文本特征的局部频率:在整体训练语料库中,类别集合 $C=(C_1,C_2,\cdots,C_m)$,m 是类别的个数,TC 中每个文本有一个类别信息,特征 t 出现在类别 $C_i\in C$ 中的频率称作特征 t 的类别频率,又称为局部频率,定义为

$$CF(t,C_i)=\frac{n(t,C_i)}{N_{C_i}} \tag{2}$$

其中,$n(t,C_i)$表示特征 t 出现在类别 C_i 中的次数,N_{C_i}表示类别 C_i 中包含的词的个数。

2.2 文本特征内耦合相似度

对于给定的特征集合 T,t_j 为文本的某一特征,特征 t_j 的内耦合相似度(IaAVS)是指该特征在两个文本之间出现频率的关联程度值[14,15]。特征内耦合相似度定义为

$$\delta_j^{ia}(x,y)=\frac{|f_j(x)|\cdot|f_j(y)|}{|f_j(x)|+|f_j(y)|+|f_j(x)|\cdot|f_j(y)|} \tag{3}$$

其中,x,y 表示在语料库的文本集合中的属性分别为 x 和 y 的两个不同文本;$f_j(x)$,$f_j(y)$表示文本集合 D 中特征 t_j 的属性值为 x,y 的子集合;$|f_j(x)|$,$|f_j(y)|$表示对应子集合的大小,即文本集合中特征 t_j 出现在文本 x 和文本 y 的频率。

2.3 文本特征间耦合相似度

由于文本特征的内耦合相似度仅考虑了同一特征在不同文本之间分布上的相关关系,文本特征的间耦合相似度考虑了不同特征之间的关联程度,从而挖掘其中的隐含关联关系。对于任意两个特征 t_j 和 t_k,在不同文本 x 和 y 下,特征间耦合相似度(IeAVS)定义如下:

$$\delta_j^{ie}(x,y)=\sum_{k=1,k\neq j}^{N}\gamma_k\delta_{j|k}(x,y) \tag{4}$$

其中,γ_k 代表特征值 t_k 的权重参数,$\gamma_k\in[0,1]$,$\sum_{k=1,k\neq j}^{N}\gamma_k=1$;$\delta_{j|k}(x,y)$表示在文本 x 和文本 y 中特征 t_j 和特征 t_k 的特征间耦合相似度。$\delta_{j|k}(x,y)$的定义如下:

$$\delta_{j|k}(x,y)=\sum_{w\in\cap}\min\{P_{k|j}(w|x),P_{k|j}(w|y)\}$$

其中,$\cap$表示在特征 t_j 分别在文本 x,y 中的情况下,特征 t_k 均为 W 的交集;$P_{k|j}(w|x)$是特征 t_j 在文本 x 中的条件下特征 t_k 取值为 w 的条件概率,其定义如下:

$$P_{k|j}(w|x)=\frac{|f_k(w)\cap f_j(x)|}{|f_j(x)|} \tag{5}$$

2.4 文本特征耦合特征模型

通过对特征内耦合相似度(IaAVS)和特征间耦合相似度(IeAVS)的分析,特征值在 x,y 之间的耦合相似度(CAVS)为

$$\delta_j^{T}(x,y)=\delta_j^{ia}(x,y)*\delta_j^{ie}(x,y) \tag{6}$$

特征 t_i 和 t_j 之间的耦合相似度(CIS)由特征内耦合相似度和特征间耦合相似度的乘积得出,即

$$CIS(t_i,t_j)=\sum_{j=1}^{N}\delta_j^{T}(x,y) \tag{7}$$

2.5 基于耦合关系模型的文本特征提取算法

设定一个耦合相似度的阈值,当计算出的特征耦合相似度高于这个阈值时,说明当一个特征 x 出现时,与其耦合的特征 y 出现的概率比较大,因此在特征提取时,去掉特征 x 时也应该去掉特征 y。

输入:训练语料库 TC,耦合特征相似度阈值 cis-rate,文本特征的整体频率阈值 of_low 和 of_high,文本特征的局部频率阈值 cf_low 和 cf_high;

输出:有效特征集 k;

Step1:计算文本特征的整体频率 z_of,保留 of_low 与 of_high 之间的特征;

Step2:计算文本特征的局部频率 z_cf,保留 cf_low 与 cf_high 之间的特征;

Step3:计算 TC 中每个特征集的内耦合相似度、间耦合相似度和耦合相似度 cis;

Step4:求得特征文本 OF,CF,TF,DF;

Step5:去除向量空间中耦合相似度 cis>耦合相似度阈值 cis-rate 的特征,获取最终的特征集 k。

3 基于耦合关系模型的文本分类方法研究

基于耦合关系模型的文本分类方法研究,以文本耦合特征相似度的提取算法训练得到的文本耦合特征作为深度置信网络的输入层;隐层采用受限玻尔兹曼机;输出层根据文本分类需求,构建区分性结点。

3.1 深度置信网络分类模型

深度置信网络(Deep Belief Nets,DBN)是一个概率生成模型,由多层随机的、隐含的变量组成。规定网络有一个可视层和一个隐层,层间存在连接,层内单元间不存在连接。图 2 描述了深度置信网络的结构和受限玻尔兹曼机之间的关系。最底层的单元代表数据向量。顶端两层拥有无向的、对称的

链接，建立了联合记忆。DBN 的重要属性是拥有有效的、逐层学习自底向上生成权值的过程，每一层从前一层的隐含单元捕获更高程度的关联。

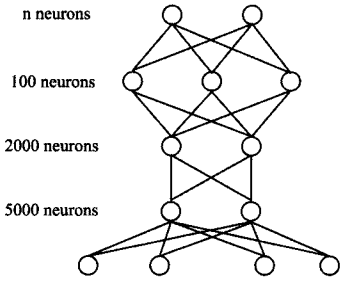

图 2　多个 RBM 和对应深度置信网络

受限玻尔兹曼机在深度学习领域中占据核心位置并且其自身拥有良好的性质。深度置信网络的相邻两层也可解释为一个 RBM。由图 2 可知，底层 RBM 的输出作为顶层 RBM 的输入，自底向上逐次传递。

玻尔兹曼机是一种特殊形式的对数线性的马尔科夫随机场(Markov Randow Field，MRF)，即能量函数是自由变量的线性函数。通过引入隐含单元，可以提高模型的表达能力，以表示非常复杂的概率分布。受限玻尔兹曼机进一步添加一些约束，在 RBM 中可见单元与可见单元之间无链接，隐含单元和隐含单元之间无链接。RBM 的能量函数 $E(v,h)$ 定义为

$$E(v,h)=-b^{\mathrm{T}}v-a^{\mathrm{T}}h-v^{\mathrm{T}}Wh \tag{8}$$

其中，模型参数 a,b 分别是可见层和隐含层的偏置量，W 是可见层和隐含层的连接权重。根据能量模型得到能量函数，就可以定义可视节点和隐藏节点的联合概率

$$p(v,h)=\frac{e^{-E(v,h)}}{\sum_{v,h}e^{-E(v,h)}} \tag{9}$$

进一步，由联合概率可以得到变量的条件概率表达形式：

$$p(v)=\frac{\sum_{h}e^{-E(v,h)}}{\sum_{v,h}e^{-E(v,h)}},p(h)=\frac{\sum_{v}e^{-E(v,h)}}{\sum_{v,h}e^{-E(v,h)}} \tag{10}$$

$$p(v|h)=\frac{e^{-E(v,h)}}{\sum_{v}e^{-E(v,h)}},p(h|v)=\frac{e^{-E(v,h)}}{\sum_{h}e^{-E(v,h)}} \tag{11}$$

对于常用的神经网络模型，一般只包含输入层、输出层和一个隐藏层，理论上来说，隐藏层越多，模型的表达能力应该越强。但是，当隐藏层数大于 1 时，如果使用随机值来初始化权重且使用梯度下降来优化参数，就会出现许多问题[16]。

(1)无监督学习的微调过程

无监督学习训练过程使用无标注数据，减少了使用大量标签数据的负担。深度置信网络结构进行训练时，采用了逐层无监督的方法来学习参数[17]。首先把输入数据向量 x(即基于耦合相似度的文本特征)和第一层隐含层作为一个 RBM，使用受限玻尔兹曼机学习方法来求得 RBM 网络的参数(链接权重和各个节点的偏置量)；然后固定这个 RBM 的参数，把隐含层 h_1 视作可见向量，把下一层视作隐藏向量，训练第二个 RBM，得到其参数；然后固定这些参数，训练 h_2 和 h_3 构成的 RBM，自底向上如此迭代，最后得到网络的生成模型：

$$p(v,h_1,h_2,\cdots,h_l)=p(v|h_1)(\prod_{k=1}^{l-2}p(h_k|h_{k+1}))p(h_{l-1},h_1) \tag{12}$$

(2)有监督学习的微调过程

在使用上述逐层无监督方法得到节点之间的权重值以及节点的偏置量之后(亦即初始化预训练过程)，由于初始化步骤使用无监督学习方法，网络参数比较接近最优值但不是最优值，因此就需要有监督训练的微调，使网络获得更好的性能。

3.2　基于耦合特征的深度学习算法

DBN 分类方法的原理是第一个 RBM 从变量接受输入，并且在隐含层上对其建模；然后隐含层的输入传递给第二层可视节点。这个建模和传递方法延续到只包含一个可视节点的最后一层。同时，网络最后一层只取一个节点，对输出的结果进行区间编号，并将区间编号与类别匹配，从而实现分类功能。

输入层：通过耦合相似度分析得到的特征向量。

Step1：初始化吉布斯方法执行次数 $gn=1$，初始化学习效率 ε；

Step2：获得具有区分性的特征，并使用数值型[0,1]区间这样的数据对其进行表示；

Step3：初始化 RBMs 的数量为 3，初始化 RBM 隐含层单元的个数；

Step4：随机初始化 DBN 网络的权值数组化 $W=[W_1,W_2,W_3]$；

Step5：定义 DBN 训练权值 $W'=[W_1',W_2',W_3']$；

Step6：调用 DBN 训练算法 DBN_train(n,v,gn)；

Step7：聚类输出，将类别标签赋值给每个聚类簇；

Step8：用训练过的 DBN 运行测试数据集中的样本，根据 DBN 层输出的聚类簇的范围，定义样本的类别。

4　实验结果及分析

4.1　数据集

数据集划分为训练集和测试集两个部分，训练集用于分类模型的学习。使用 Iris 数据集对深度置信网络分类模型的有效性进行验证。Iris 数据集如表 1 所列。

表 1　Iris 数据集描述

DataSet	Object	Attribute	Class
Iris	150	4	3

4.2　耦合相似度阈值确定实验验证方案

耦合相似度的阈值是实验中的重要判定部分，当计算得出的耦合相似度大于该阈值时，说明两个特征间存在耦合关系，因此当其中一个特征值被舍去时，另一个特征也应该被去掉。通过多次实验验证，该阈值设定在 0.6～0.8 之间时得出的分类效果比较准确，因此在该区域间做了细微的调整并最终选定该阈值为 0.73。阈值确定如表 2 所列。

表 2　阈值确定

阈值选定	0.6	0.65	0.7	0.73	0.75	0.8
分类效果匹配度(%)	90.32	91.97	93.86	94.98	94.11	90.75

4.3　分类方法实验验证方案

以分类方法应用到 UCI 数据库中著名的 Iris 数据集为例，该数据集包含 150 个样本和 3 个类。这里将数据集进行

归一化处理，分别使用 90%的样本进行训练，10%的样本进行测试。网络的结构将 DBN 网络的输出分成 3 个不同的区间，分别是区间 1:[0,0.255]，区间 2:[0.256,0.795658]，区间 3:[0.796,1]。样本的标签显示类 2,1,0 分别对应区间 1,2,3。将数据集的 10%应用到训练过的 DBN 并且找到每个样本的输出所在的范围。根据划分的区间，决定每个样本的类别，然后将结果和已知样本相关联的类标签进行比较。

实验结果发现，分类的性能不通过增加 RBM 层数或者吉布斯方法的数量来改变。用深度置信网络分类方法时，Iris 数据上的分类准确性达到 99.5%，这样的性能是令人满意的。

下面描述 DBN，SVM 以及 Coupling-DBN 训练的细节以及在 3 个数据集上分类的性能表现。为了在文本分类中比较 DBN，Coupling DBN 和 SVM 的性能，这里只记录每种分类法在每个数据集上的最好表现，表 3－表 5 分别列出了 LingSpam，SpamAssassin 和 Enronl 中性能评价指标的比较结果。

表 3　LingSpam 中分类方法性能表现对比(%)

Performance	DBN	SVMC-1	Coupling DBN
Accuracy	99.45	99.24	99.51
Spam Recall	98.54	96.67	98.40
Spam Precision	98.52	98.74	98.64
Ham Recall	99.64	99.75	99.82
Ham Precision	99.71	99.35	99.68

表 4　SpamAssassin 中分类方法性能表现对比(%)

Performance	DBN	SVMC-1	Coupling DBN
Accuracy	97.50	97.32	98.11
Spam Recall	95.51	95.24	95.89
Spam Precision	96.40	96.14	96.32
Ham Recall	98.39	98.24	98.35
Ham Precision	98.02	97.89	98.28

表 5　Enronl 中分类方法性能表现对比(%)

Performance	DBN	SVMC-1	Coupling DBN
Accuracy	97.43	96.92	97.67
Spam Recall	96.77	97.27	98.81
Spam Precision	94.94	92.74	94.82
Ham Recall	97.83	96.78	97.98
Ham Precision	98.53	98.84	99.10

结束语　本文旨在解决当前文本分类中分类器分类性能低的问题。由于特征的耦合关系能改善分类性能从而提高分类的准确性，本文采用结合耦合相似度度量方法与深度置信网络的分类方法来优化分类器性能，以便更加精准地对文本进行分类。分析和实验结果表明耦合相似关系为文本分类中特征提取提供了更加准确的分类信息。在接下来的工作中，将收集更多带有特征信息的相关数据集，进一步改进算法，争取在此项研究结果的基础上取得进一步成果。

参 考 文 献

[1] Jiang J Y. Text classification based on deep belief networks research[J]. Liaoning Technical University, 2013(in Chinese)
蒋金叶. 基于深度置信网络的文本分类研究[D]. 辽宁工程技术大学，2013

[2] Masand B, Linoff G, Waltz D. Classifying news stories using memory based reasoning [C]// Proceeding of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 1992:59-65

[3] Lewis D D, Ringuette M. A comparison of two learning algorithms for text categorizationg[C]// Third Annual Symposium on Document Analysis Information Retrieval. 1994,33:81-93

[4] Langley P, Iba W, Thompson K. An analysis of Bayesian classifiers [C]// Proceedings of the 10th National Conference on Aritificial Intelligence. 1998:233-228

[5] Lewis D D, Schapire R E, Callan J P, et al. Training algorithms for linear text classifiers [C]// Proceesings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 1996:298-306

[6] Wiener E, Pedersen J O, Weigend A S. A neural network approach to topic spotting [C]// Proceesdings of 4th Annual Symposium on Document Analysis and Information Retrieval. 1995:317-332

[7] Yang Y, Chute C G. An example-based mapping method for text categorization and retrieval[J]. ACM Transactions on Information Systems(TOIS), 1994, 12(3):252-277

[8] Vapnik V. The nature of statistical learning theory [M]. Springer, 2000

[9] Joachims T. Text categorization with support vector machines: Learning with many relevant features[M]. Springs Berlin Heidelberg, 1998

[10] Colloberr R, Weston J, Bottou L, et al. Natural language processing(Almost)from scratch [J]. Joumal of Machine Learning Research, 2011, 12:2493-2537

[11] Salakhutdinov R, Hinton G. Semantic Hashing[J]. International Journal Approximate Reasoning, 2009, 50(7):969-978

[12] Socher R, Lin C, Ng A. Parsing Natural Scenes and Natural Language with Recursive Neural Networks[C]// Proceeding of the 28th Int Conf on Machine Learning. Garrnany: International Machine Learning Society, 2011

[13] Wang B X. Research on the semantic mining of question-answer pairs in web communities[D]. Harbin: Harbin Institute of Technology, 2013(in Chinese)
王宝勋. 面向网络社区问答对的语义挖掘研究 [D]. 哈尔滨：哈尔滨工业大学，2013

[14] Cao L B, Ou Y, Yu P S. Couple behavior analysis with applications[J]. IEEE Transactions on, Knowledge and Data Engineering, 2012, 24(8):1378-1392

[15] Song Y, Cao L B, Wu X, et al. Coupled behavior analysis for capturing coupling relationships in group-based market manipulations[C]// Proceedings of the 18th International Conference on Knowledge Discovery and Data Mining. 2012:976-984

[16] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural network[J]. Science, 2006, 313(5786):504-507

[17] Hinton G E, Osidero S, The Y W. A fast learning algorithm for deep belief nets[J]. Neural Computation, 2006, 18(7):1527-1554