doi:10.6041/j.issn.1000-1298.2017.10.025

基于 word2vec 和 LSTM 的饮食健康文本分类研究

赵　明[1]　杜会芳[1]　董翠翠[1]　陈长松[2]

(1. 中国农业大学信息与电气工程学院, 北京 100083; 2. 公安部第三研究所, 上海 200031)

摘要: 为了对饮食文本信息高效分类,建立一种基于 word2vec 和长短期记忆网络(Long-short term memory,LSTM)的分类模型。针对食物百科和饮食健康文本特点,首先利用 word2vec 实现包含语义信息的词向量表示,并解决了传统方法导致数据表示稀疏及维度灾难问题,基于 K-means + + 根据语义关系聚类以提高训练数据质量。由 word2vec 构建文本向量作为 LSTM 的初始输入,训练 LSTM 分类模型,自动提取特征,进行饮食宜、忌的文本分类。实验采用 48 000 个文档进行测试,结果显示,分类准确率为 98.08%,高于利用 tf-idf、bag-of-words 等文本数值化表示方法以及基于支持向量机(Support vector machine,SVM)和卷积神经网络(Convolutional neural network,CNN)分类算法结果。实验结果表明,利用该方法能够高质量地对饮食文本自动分类,帮助人们有效地利用健康饮食信息。

关键词: 文本分类; word2vec; 词向量; 长短期记忆网络; K-means + +

中图分类号: TP182　　**文献标识码**: A　　**文章编号**: 1000-1298(2017)10-0202-07

Diet Health Text Classification Based on word2vec and LSTM

ZHAO Ming[1]　DU Huifang[1]　DONG Cuicui[1]　CHEN Changsong[2]

(1. *College of Information and Electrical Engineering*, *China Agricultural University*, *Beijing* 100083, *China*
2. *The Third Research Institute*, *Ministry of Public Security*, *Shanghai* 200031, *China*)

Abstract: The development of Internet information age makes Internet information grow rapidly. As the main information form of the network, the texts are massive, so is texts information about diet. The diet information is closely related with people's health. It is important to make texts be auto-classified to help people make effective use of health eating information. In order to classify the food text information efficiently, a classification model was proposed based on word2vec and LSTM. According to the characteristics of food text information in encyclopedia and diet texts in health websites, word2vec realized word embedding, including semantic information which solved the problem of sparse representation and dimension disaster that the traditional method faced. Word2vec combined with K-means + + was used to cluster key words both of the proper and the avoiding to enlarge relevant words in classification dictionaries. The words were employed to work out rules to improve the quality of training data. Then document vectors were constructed based on word2vec as the initial input values of long-short term memory network (LSTM). LSTM moved input layer, hidden layers of the neural network into the memory cell to be protected. Through the "gate" structure, sigmoid function and tanh function to remove or increase the information to the cell state which enabled LSTM model the "memory" to make good use of the text context information, which was significant for text classification. Experiments were performed with 48 000 documents. The results showed that the classification accuracy was 98.08%. The result was higher than that of ways based on tf-idf and bag-of-words text vectors representation methods. Two other classification algorithms of support vector machine (SVM) and convolutional neural network (CNN) were also conducted. Both of them were based on word2vec. The results showed that the proposed model outperformed other competing methods by several percentage points. It proved that the method can automatically classify dietary texts with high quality and help people to make good use of health diet information.

Key words: text classification; word2vec; word embedding; long-short term memory network; K-means + +

收稿日期: 2017-01-13　修回日期: 2017-03-13

基金项目: 信息网络安全公安部重点实验室开放课题项目(61503386)

作者简介: 赵明(1963—),男,副教授,主要从事本体论和语义网研究,E-mail: zhaoming@cau.edu.cn

引言

网络信息时代的高速发展使互联网信息急剧增长，文本作为网络主要的信息承载形式，数据量巨大。文本自动分类技术能够将海量非结构化文本信息规范归类，帮助人们更好地管理、利用和挖掘信息[1-2]。正确的饮食信息能有效帮助人们合理饮食，保障身体健康。饮食宜、忌文本自动分类能够使人们利用有效信息，根据自身健康状况做更好的营养搭配。

目前，国内外对文本自动分类的研究十分关注，文本表示以及分类器的选择一直是文本分类的两大技术难点及热点。ZHANG 等[3]利用独热表示方法（One-hot representation）把文本表示为向量，然后将支持向量机（SVM）和 BP 神经网络结合对文本进行分类。PACCANARO 等[4]提出 Distributed representation 概念，通称为 Word embedding，即词向量。龚静等[5]利用改进的 tf－idf 算法提取文本特征，并利用朴素贝叶斯分类器进行文本分类。豆孟寰[6]基于 N－gram 统计语言模型对越南语文本进行分类，N－gram 模型根据每个词出现在其前面 n 个词的概率来表示文本，但是 N－gram 模型无法对更远的关系建模。BENGIO 等[7]提出用神经网络来构建语言模型，一定程度上解决了 N－gram 模型的问题。以上方法中对文本进行数值化表示面临数据稀疏以及建模词之间语义相似度大等问题，且限于对词汇特征、句法特征的发现。MIKOLOV 等[8]指出使用工具 word2vec 训练得到的向量低维、连续，同时通过计算这些向量间余弦距离可以判断词语之间的语义相似度[9]。LILLEBERG 等[10]利用 word2vec 提取语义特征并基于 SVM 进行文本分类，然而当样本数量较大时，SVM 的训练速度较慢。

对于序列化输入，循环神经网络（Recurrent neural network，RNN）能够把邻近位置信息进行有效整合[11-12]，用于自然语言处理的各项任务。RNN 的子类长短期记忆网络模型 LSTM[13-14] 能避免 RNN 的梯度消失问题，具有更强的“记忆能力”，能够很好地利用上下文特征信息，并保留文本的顺序信息，自动选择特征，进行分类。

本文利用 word2vec 和 LSTM 进行饮食健康文本分类。首先基于饮食健康文本语料库，利用 word2vec 训练得到具有语义信息的词向量，然后采用 K-means＋＋聚类饮食文本宜、忌类词语提高数据质量，最后训练 LSTM 模型捕获文本的完整语义并进行文本分类。

1　材料与方法

1.1　获取语料

通过 python 库 Beautiful soup 和 Request，爬取食物百度百科、互动百科、饮食健康类网站等关于食物营养价值或者饮食宜、忌中文文本语料。

1.1.1　语料预处理

中文与英文不同，中文以字为基本单文，单独的字大多数不能独立表达意思，因此需要对中文文本进行分词处理。采用结巴分词系统，以精确模式来进行分词。结巴分词是基于 Trie 树结构的高效词图扫描，生成句子中汉字所有可能成词情况所构成的有向无环图，采用动态规划查找最大概率路径，找出基于词频的最大切分组合，对于未登录词，采用基于汉字成词能力的 HMM 模型和 Viterbi 算法。

停用词通常没有实际含义，针对饮食健康宜、忌文本词条的特点，将文本内容中出现频率非常高或者一些介词、代词、虚词等停用词以及特殊符号去除，比如“而言”、“根据”、“人们”、“￥”等。同时本文通过添加饮食相关词汇词典来提高分词的正确率。预处理后的语料如图 1 所示，词间以空格作为分隔。

高糖 食物　　高糖 食物 很 容易 造成 肌肤 长 粉刺 的 惨状。所以，如果 你 恰巧 喜欢 喝 含糖 的 饮料 或是 加糖 的 咖啡，是 时候 改掉 这个 不良习惯 了。除了 喜欢 含糖 的 饮料，如果 你 还 喜欢 吃 巧克力、蛋糕 等，更要 控制 好 自己，不要 对 高糖 食物 摄入 太多。

乳制品　　单纯 从 营养学 的 角度 来看，乳制品 是 很棒 的 食物。其 营养价值 丰富，能为 身体 提供 能量，还 可以 补钙……但是，如果 过量 进食 乳制品，如 奶酪、黄油、牛奶 等，很容易 刺激 体内 激素 分泌，从而 影响 肌肤。甚至 有人 认为：过量 摄入 乳制品 是 导致 女性 内分泌 紊乱、出现 皱纹 的 元凶。无论 这种 观点 是 真是假，适量 摄入 乳制品 才 是 兼顾 营养 与 美肤 的 好 办法。

加工 食品　　生活 越来越 方便，加工 食品、半成品 在 我们 的 日常 饮食 中 逐渐 占据 一席之地。可 你 知道 吗？含有 大量 食品 添加剂 的 加工 食品 是 导致 毛孔 堵塞、皮肤 老化 的 主要 原因 之一。所以，若想 美肤 还是 远离 这些 加工 食品 吧，多 吃 新鲜 蔬菜、水果 才是 王道。

咖啡因　　咖啡因 可能 会 导致 女性 皮肤 松弛，美容 专家 建议 喜欢 喝咖啡 的 女性 最好 选择 不含 咖啡因 的 咖啡。而 红茶、绿茶 中 也 含有 少量 咖啡因，因此 女性朋友 们 每日 的 饮茶 量 应该 有所 控制。若想 追求 白皙 光滑 的 肌肤，白开水 才是 最好 的 选择。

甲壳类 食物　　无论是 虾、蟹 还是 龙虾，海产品 中 通常 富含 碘，而 碘 的 代谢 并非 通过 消化道 排出 体外，而是 通过 皮脂腺 排出 体外。因此，碘 元素 摄入 过多 很 容易 导致 毛孔 堵塞，引发 粉刺、黑头 等 肌肤 状况。

图 1　饮食文本预处理结果

Fig. 1　Pretreatment result of diet text data

1.1.2　基于 word2vec 训练词向量

word2vec 有连续词袋模型（Continuous bag-of-words，CBOW）和 Skip－Gram 两种模型。word2vec 能够将文本词语转化为向量空间中的向量，而向量的相似度可以表示文本语义的相似度。

本文采用基于 Hierarchical Softmax 算法的 Skip－Gram 模型，词向量维度设置为 200，训练窗口设置为 5。Skip－Gram 模型以当前词来预测上下文的词，即预测 $P(w_m \mid w_t)$，其中 w_t 为当前词，$t-c \leqslant m \leqslant t+c$ 且 $m \neq t$，c 是窗口尺寸。输入层是当前词的词向量，然后是特征映射层，输出层是一棵 Huffman 树[15-16]。此 Huffman 树以语料库中出现的词作为叶子结点，

以各词在语料库中出现的次数为权值。利用 Hierarchical Softmax 算法结合 Huffman 编码,一般左子树编码为 1,右子树编码为 0,每条边上都有相应的权重,语料库中的每个词可以从根节点沿着唯一路径被访问到,路径即形成了其编码,目标是使预测词的二进制编码概率最大。利用针对 $w_1, w_2, \cdots, w_t$ 的词组序列,Skip - Gram 的优化目标函数为

$$J = \frac{1}{L}\sum_{t=1}^{L}\sum_{-c\leqslant j\leqslant c,\ j\neq 0}\lg p(w_{t+j}\mid w_t) \tag{1}$$

式中 p——概率函数

窗口 $c>0$,并利用梯度下降法对其进行优化。

由 word2vec 训练得到的词向量可以余弦距离来判断语义相似程度。余弦值越大,语义越相近;反之,语义相差较远,如图 2 所示。如图 3 所示,在二维空间中展示词向量之间的语义距离。

```
Enter word or sentence (EXIT to break): 番茄

Word: 番茄  Position in vocabulary: 412

                                  Word       Cosine distance
------------------------------------------------------------
                                番茄酱       0.767114
                               番茄红素       0.711699
                                  红素                0.645804
                                番茄汁       0.612789
                                  茄汁                0.577004
                                氧化剂       0.529488
                               抗氧化剂       0.526189
                                  制品                0.510170
                                  美国                0.509126
                                坏血酸       0.498949
                               抗坏血酸       0.494858
                               蔬菜水果       0.494290
                               氧化作用       0.492796
                                  化剂                0.492097
                                发生率       0.484277
                                抗氧化       0.481484
                                  摄入                0.476417
                                  坏血                0.471038
```

图 2 词向量语义相似度

Fig. 2 Semantic similarity of word embeddings

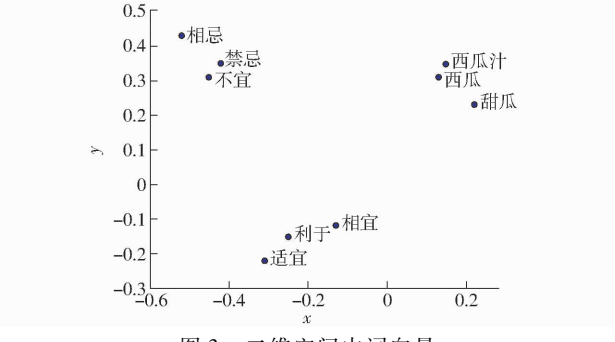


图 3 二维空间中词向量

Fig. 3 Word embeddings in two dimension

1.1.3 获得训练数据

食物类百科和各饮食健康网上对饮食的描述文档中,往往会使用不同的词(带有下划线的词语)来表达饮食宜或忌的情况。比如描述忌食的词条:

“贫血者忌食辛辣、生冷不易消化的食物,忌摄入过多牛奶、大蒜、蚕豆、脂肪、糖和盐”。

“啤酒中嘌呤含量较高,配合肉类海鲜一起吃,易引发痛风”。

描述饮食适宜的词条:

“香橙和牛奶同食营养更加丰富,具有清凉解渴、抗癌防癌的功效”。

“毛豆和豆腐、豆浆等豆制品含有大量的植物化合物异黄酮,对皮肤胶原具有保护作用”。

饮食宜、忌分类词典部分关键词如表 1 所示。与各网络交流平台信息文本不同,饮食类文本词义较规范,而网络流行词容易造成一词多义和歧义。此特点为本文利用 word2vec 训练词向量并采用 K-means + +[17] 聚类语义相近的词向量提供了良好的充分性。本文利用 K-means + + 基于余弦距离对词向量进一步聚类,得到表达饮食适宜语义相近的词向量聚类结果以及表达饮食禁忌语义相近的词向量聚类结果,根据语义关系扩展相应的词典。根据词典以及句子模型制定正则表达式来自动提取饮食宜、忌的文本:提取含有饮食适宜类字典中词语,但不含忌类别词典中词的句子归为饮食适宜的类别;否则,归为忌的类别。由此可知,饮食宜、忌类别词典中的词越多且精确,训练语料的质量就会越好。利用 K-means + + 扩展词语有利于提高训练语料的数据质量,并为训练良好的分类模型打下基础。

表 1 饮食宜、忌分类关键词库

Tab. 1 Keywords of proper and avoiding about diet

宜类词汇	忌类词汇
宜食、为宜、保护、良药、防癌、抗癌、美容、润肤、明目、补钙、补锌、保健、解暑、健脾、健胃、消食	慎食、忌食、不宜、忌、禁忌、大忌、导致、引发、胀痛、少食、致癌物、积食、恶心、危及、破坏、相克

K-means + + 是针对 K-means[18] 聚类方法随机选择初始化中心的不足而改进的方法,K-means + + 是以正比于每个数据点到其最近中心点距离的概率来选择中心点。算法步骤如下:

(1)开始时,初始化中心点集合为空。

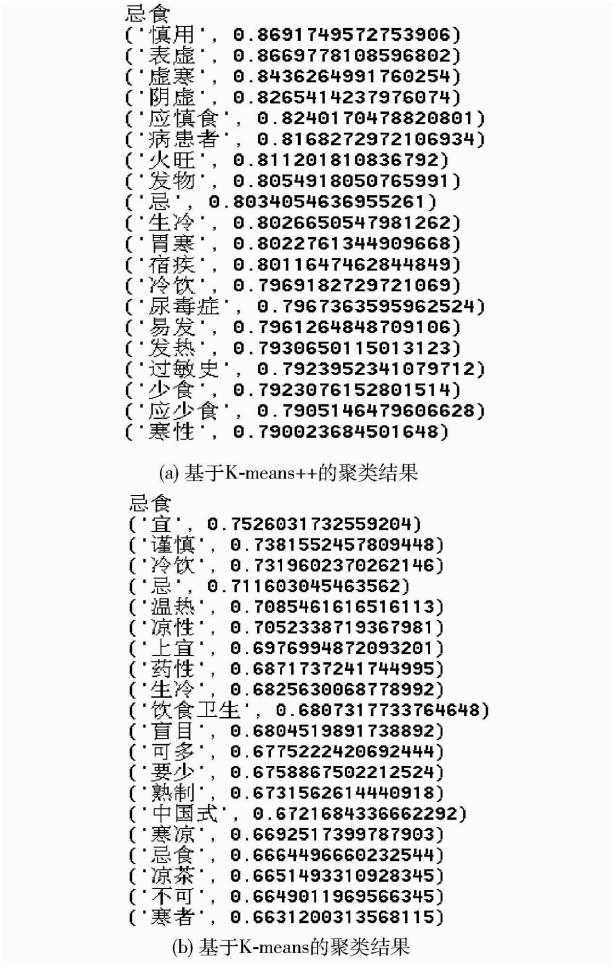
(2)从数据中随机选择第 1 个中心点,然后重复以下步骤,直到选出 k 个初始中心点为止。

(3)计算每个数据点到最近中心点的距离 D,以正比于 D 的概率,随机选择一个数据点作为新中心点加入到中心点集合中。

(4)重复步骤(3)。

图 4 为基于 word2vec 训练的词向量并分别利用 K-means + + 和 K-means 聚类,与“忌食”同一类余弦距离最近的前 20 个词。由于聚类效果受初始中心选取的影响,K-means 初始化中心点的随机性

有可能导致选择的中心点很差。利用 K-means 和 K-means + + 两种聚类方法，表 2 列出了用于扩充饮食宜、忌类词典的词所属于的簇聚类效果，由表 2 可知，K-means + + 算法效果更好，比利用 K-means 聚类方法 F 高 4 ~ 9 个百分点。



(a) 基于K-means++的聚类结果

(b) 基于K-means的聚类结果

图 4　饮食禁忌词聚类结果

Fig. 4　Cluster results of avoiding diet words

1.2　计算文档向量

语料库中文档长度为 15 ~ 130 个词，由 word2vec 训练得到文档中每个词的词向量，将词向量对应相加，并平均处理，以此得到文档的空间向量。同时采用 tf-idf[19]、bag-of-words[20] 模型分别计算饮食宜、忌文档向量。对3种情况下得到的文档向量进行二维可视化对比展示如图 5 所示。红圈代表饮食禁忌类文档向量，蓝圈代表饮食适宜文档向量。

表 2　基于不同聚类方法的聚类结果

Tab. 2　Cluster result based on different methods　%

聚类方法与指标		忌食	发病	致命	不适	相宜	功效	养颜
K-means	准确率	73	71	72	75	80	71	68
	召回率	71	69	70	73	79	72	71
	F	72	70	71	74	79	71	69
K-means + +	准确率	79	76	76	81	84	81	79
	召回率	80	79	78	83	83	77	78
	F	79	77	77	82	83	79	78

bag-of-words 模型是基于字典根据文档中的词出现的次数来表示文档向量的，未在字典中出现过的词表示为 0。假如有字典为：{“牛奶”:1，“草莓”:2，“丰富”:3，“清凉”:4，“解渴”:5，“增加”:6，“营养”:7，“生津”:8}，则文档“牛奶营养丰富，牛奶苹果宜同食”用 bag-of-words 方法可以表示为 [2,0,1,0,0,0,1,0]。而 tf-idf 是在 bag-of-words 表示基础上对文档中的词进行加权来表示文本。tf 指某词 t 在文档中出现的次数，逆文档频率为

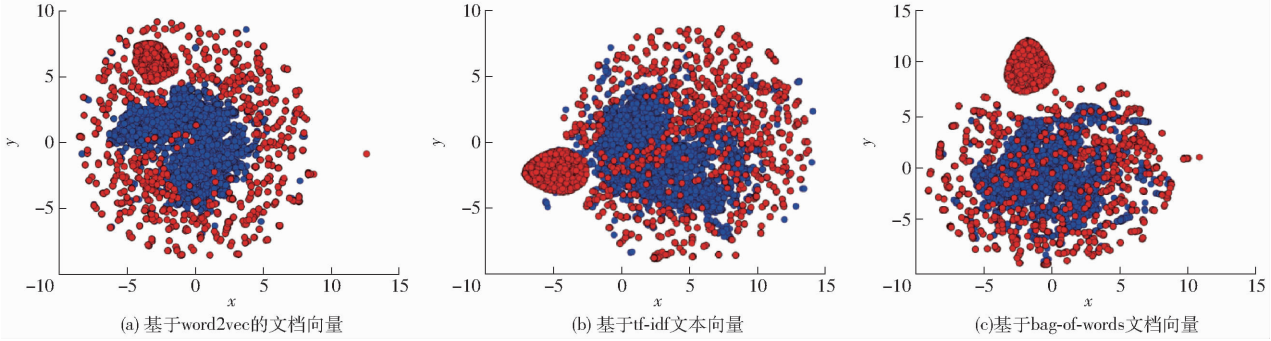
$$i_{\mathrm{df}} = \lg\left(1 + \frac{N}{N_t}\right) \tag{2}$$

式中　N——所有文档数

N_t——含有词 t 的文档数

tf-idf 用于评估一个词在语料库中的重要程度。然而，tf-idf 和 bag-of-words 方法在表示空间向量时都有一个缺点：忽略了文本中词语间的语义信息。比如对于“牛奶和草莓相宜”与“牛奶和大枣同食为宜”，利用 tf-idf 和 bag-of-words 模型表示“相宜”和“为宜”，在空间向量中距离则较远，但两者是具有相似的语义信息的。

由图 5 可知，word2vec、tf-idf 和 bag-of-words 方法都可以将文本进行向量化。根据基于 word2vec 得到的饮食宜、忌类文档向量在向量空间中界限明显，利用 tf-idf 方法得到饮食相宜的文档向量和饮食



(a) 基于word2vec的文档向量　(b) 基于tf-idf文本向量　(c)基于bag-of-words文档向量

图 5　饮食宜、忌文档向量表示

Fig. 5　Document vectors of proper and avoiding diet

禁忌的文档向量在向量空间中有少部分重叠现象，而利用 bag-of-words 方法表示的两类文档向量界限不明显。

本文采用的是 word2vec 模型，将其得到的文档空间向量作为 LSTM 神经网络的初始输入。

1.3 LSTM 分类算法

LSTM 的隐含层之间形成闭环。LSTM 隐藏层到隐藏层的权重是网络的记忆控制者，负责调度记忆，而隐藏层的状态作为某个时刻记忆状态将参与下一次的预测。

LSTM 将 RNN 的输入层、隐层移入记忆单元(Memory cell)加以保护[21]，并通过“门”结构来去除或增加信息到细胞状态，如图 6 所示。

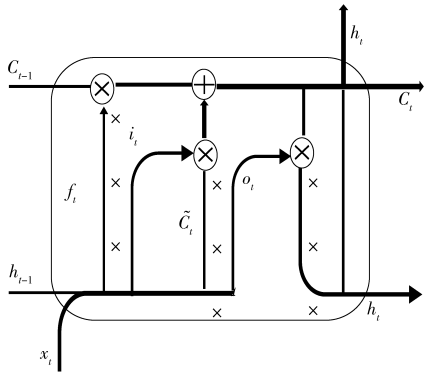


图 6 LSTM 门结构

Fig. 6 LSTM gate architecture

LSTM 解决了标准 RNN 的梯度消失和梯度爆炸问题[22]。x 是输入数据，h 为 LSTM 单元的输出，C 为记忆单元的值。在 LSTM 动态门结构中，遗忘门决定要忘记什么信息，该门读取 h_{t-1} 和 x_t，输出一个在 0 到 1 之间的数值，f_t 表示要舍弃信息的百分值，0 代表完全舍弃，1 代表完全保留。f_t 的计算公式为

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \tag{3}$$

式中 σ——sigmoid 函数 W_f——遗忘门权重

b_f——遗忘门偏置

更新的值为 i_t，用于控制当前数据输入对记忆单元状态值的影响。然后，一个 tanh 层创建一个新的候选值向量，会被加入到状态中。

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \tag{4}$$

$$\tilde{C}_t = \tanh(W_c[h_{t-1}, x_t] + b_c) \tag{5}$$

式中 W_i——更新门权重

b_i——更新门偏置

tanh——双曲正切函数

W_c——更新候选值

b_c——更新候选值偏置

$\tilde{C}_t$——候选值

之后，把旧状态与 f_t 相乘，丢弃掉确定需要丢弃的信息，根据决定更新每个状态的程度进行变化。

$$C_t = f_t C_{t-1} + i_t \tilde{C}_t \tag{6}$$

式中 C_t——新的状态值

输出门值 o_t 控制记忆单元状态值的输出，计算公式为

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \tag{7}$$

$$h_t = o_t \tanh C_t \tag{8}$$

式中 W_o——更新输出值的权重

b_o——更新输出值偏置

h_t——最终确定输出的那部分

LSTM 采用梯度下降法更新各层权重，使得代价函数值最小。

利用基于 word2vec 得到的文档向量训练集来训练 LSTM 模型，采用一个 LSTM 层和全连接 softmax 层，对测试文档进行分类。

2 实验结果与分析

利用网络爬虫技术爬取食物类百度百科、互动百科以及有关饮食健康类网站的文本数据，经过处理后得到 24 000 个饮食相宜类的文档和 24 000 个饮食禁忌类的文档。其中训练集、交叉验证集、测试集比例为 6∶2∶2。本文分别基于 word2vec 和 LSTM 分类方法、tf-idf 和 LSTM 分类方法、bag-of-words 和 LSTM 分类方法进行实验，分类结果如表 3 所示。评估文本分类的主要指标有精确率、召回率、F_1 值(精确率和召回率加权调和平均值)及正确率。

表 3 基于不同文本表示方法的分类结果

Tab. 3 Classification results based on different text representation methods %

方法	类别	精确率	召回率	F_1	正确率
bag-of-words + LSTM	忌	89.98	96.91	93.32	90.93
	宜	93.10	80.03	86.07	
tf-idf + LSTM	忌	96.03	95.59	95.81	94.71
	宜	93.29	93.75	93.52	
word2vec + LSTM	忌	98.63	98.39	98.51	98.08
	宜	97.08	97.52	97.30	

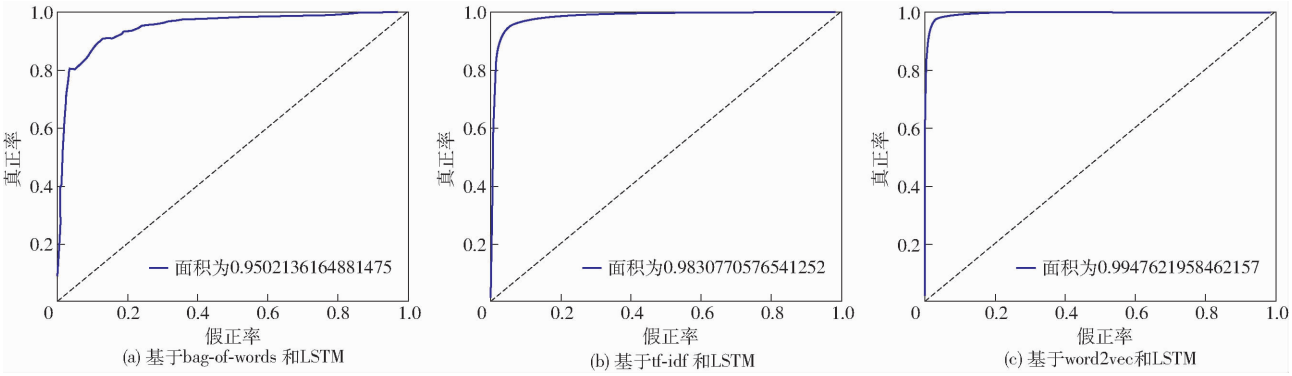
由表 3 可知，在饮食适宜、禁忌文本分类中，基于 word2vec 和 LSTM 方法的精确率、召回率、F_1 均高于基于 tf-idf 和 LSTM 方法与基于 bag-of-words 和 LSTM 方法。正确率高于基于 tf-idf 和 LSTM 分类方法 3.37 个百分点，高于基于 bag-of-words 和 LSTM 分类方法 7.51 个百分点。实验证明利用 word2vec 训练能够表示词间语义关系的词向量对提高文本分类精度的有效性。

ROC 曲线下方的面积 AUC(Area under the ROC

curve)提供了评价模型平均性能的另一种方法。如果分类模型较好,曲线靠近左上角,且 AUC 接近于 1,即 ROC 曲线下的面积(AUC)越大,表示分类效果越好。

绘制以上 3 种方法相应的 ROC 曲线如图 7 所示,由图 7 可知,基于 word2vec 和 LSTM 方法的分类效果最好。

同时,采用 SVM、CNN 分类算法分别进行实验。



(a) 基于bag-of-words 和LSTM　(b) 基于tf-idf 和LSTM　(c) 基于word2vec和LSTM

图 7　ROC 曲线

Fig. 7　ROC curves

SVM 寻求结构风险最小化,求解化为一个线性约束的凸二次规划问题;实验采用线性核函数构造判别函数以及利用梯度下降法来选取 SVM 模型的参数。CNN 具有局部感知、权值共享等特征,实验中采用一层有 128 个神经元的卷积层、一层有 128 个神经元的池化层和一层含有 2 个神经元的全连接 softmax 层对饮食宜、忌文本进行分类。同样采用一层有 128 个神经元的 LSTM 层和一层有 2 个神经元的全连接 softmax 层进行本文分类,结果如表 4 所示。

表 4　基于不同分类算法的分类结果

Tab. 4　Classification results based on different classification methods　%

方法	类别	精确率	召回率	F_1	正确率
word2vec + SVM	忌	87.35	91.80	89.52	89.25
	宜	91.36	86.70	88.94	
word2vec + CNN	忌	93.03	93.59	93.31	93.43
	宜	94.14	93.82	93.98	
word2vec + LSTM	忌	98.63	98.39	98.51	98.08
	宜	97.08	97.52	97.30	

由表 4 可知,基于 word2vec 和 LSTM 的分类结果最好。相对于 SVM 模型,深度神经网络模型不需要手动提取特征,自动学习复杂特征的能力强大,并且效率较高。CNN 是在图像识别领域比较成熟的技术,注重全局模糊感知,LSTM 侧重相邻位置的信息重构。由此可见,对于序列化的自然语言处理任务,LSTM 更具有说服力,表 4 也验证了 LSTM 在饮食文本分类中的有效性。

3　结论

(1)针对食物百科和饮食健康网站文本上下文较长、语义表征联系紧密等特点,利用 word2vec 对词进行空间向量表示,一定程度上解决了文本表示面临的数据稀疏和词间语义关系建模困难等问题。采用对处理序列化数据具有优势的 LSTM 模型获取整个文本语义特征并进行分类,有利于分类精度的提高。

(2)基于食物百科和饮食健康网站文本描述规范的特点,进一步利用 K-means + + 方法基于词向量之间的余弦距离将语义相近的词聚类,这能够在深层语义关系上全面扩充饮食宜、忌类别关键词词典,提高训练数据质量。

(3)利用基于 word2vec 和 LSTM 的分类系统对饮食文本进行宜、忌分类效果较好。

参　考　文　献

1　魏芳芳,段青玲,肖晓琰,等. 基于支持向量机的中文农业文本分类技术研究[J/OL]. 农业机械学报,2015,46(增刊):174 – 179. http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx ? file_no = 2015S029& flag = 1&journal_id = jcsam. DOI: 10.6041/j.issn.1000-1298.2015.S0.029.

WEI Fangfang, DUAN Qingling, XIAO Xiaoyan, et al. Classification technique of Chinese agricultural text information based on SVM[J/OL]. Transactions of the Chinese Society for Agricultural Machinery, 2015,46(Supp.): 174 – 179. (in Chinese)

2　段青玲, 魏芳芳, 张磊,等. 基于 Web 数据的农业网络信息自动采集与分类系统[J]. 农业工程学报, 2016, 32(12): 172 – 178.

DUAN Qingling, WEI Fangfang, ZHANG Lei, et al. Automatic acquisition and classification system for agricultural network

information based on web data[J]. Transactions of the CSAE, 2016, 32(12):172 - 178. (in Chinese)

3 ZHANG W, TANG X, YOSHIDA T. Text classification with support vector machine and back propagation neural network[C]// International Conference on Computational Science-ICCS 2007, Part IV, LNCS 4490, 2007:150 - 157.

4 PACCANARO A, HINTON G E. Learning distributed representations of concepts using linear relativnal embedding[J]. IEEE Transactions on Knowledgs & Data Engineering, 2002, 13(2):232 - 244.

5 龚静, 胡平霞, 胡灿. 用于文本分类的特征项权重算法改进[J]. 计算机技术与发展, 2014(9):128 - 132.
GONG Jing, HU Pingxia, HU Can. Improvement of algorithm for weight of characteristic item in text classification[J]. Computer Technology and Development, 2014(9):128 - 132. (in Chinese)

6 豆孟寰. 基于词袋和 N - Gram 统计语言模型的越南语文本分类研究[D]. 武汉:武汉理工大学, 2015.
DOU Menghuan. Vietnamese text classification based on bag-of-words and statistical n-gram language modeling[D]. Wuhan: Wuhan University of Technology, 2015. (in Chinese)

7 BENGIO Y, SCHWENK H, SENECAL J, et al. Neural probabilistic language models[J]. Journal of Machine Learning Research, 2003, 3(6):1137 - 1155.

8 MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[C]// Computer Science 2013, 2013:1 - 12.

9 赵明, 杜亚茹, 杜会芳, 等. 植物领域知识图谱构建中本体非分类关系提取方法[J/OL]. 农业机械学报, 2016, 47(9): 278 - 284. http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx? file_no = 20160938&flag = 1& journal_id = jcsam. DOI: 10.6041/j.issn.1000-1298.2016.09.038.
ZHAO Ming, DU Yaru, DU Huifang, et al. Research on ontology non-taxonomic relations extraction in plant domain knowledge graph construction[J/OL]. Transations of the Chinese Society for Agricultural Machinery, 2016, 47(9):278 - 284. (in Chinese)

10 LILLEBERG J, ZHU Y, ZHANG Y. Support vector machines and word2vec for text classification with semantic features[C]// IEEE International Conference on Cognitive Informatics & Cognitive Computing, 2015:136 - 140.

11 CHO K, MERRIENBOER B V, GULCEHRE C, et al. Learning phrase representations using RNN Encoder-Decoder for statistical machine translation[C]// Computer Science 2014, 2014:1 - 12.

12 EBRAHIMI J, DOU D. Chain based RNN for relation classification[C]// Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2015:1244 - 1249.

13 HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8):1735 - 1780.

14 GRAVES A. Supervised sequence labelling with recurrent neural networks[M]. Berlin Heidelberg: Springer, 2012.

15 XIONG F, DENG Y, TANG X. The architecture of word2vec and its applications[J]. Journal of Nanjing Normal University, 2015.

16 KABIR S, AZAD T, ASHRAFUL ALAM A S M, et al. Effects of unequal bit costs on classical huffman codes[C]// International Conference on Computer and Information Technology. IEEE, 2014:96 - 101.

17 ARTHUR, DAVID, VASSILVITSKII, et al. K-means + + : the advantages of careful seeding[C]// 8th Acm-Siam Symposium on Discrete Algorithms, SODA 2007, 2007:1027 - 1035.

18 霍迎秋, 秦仁波, 邢彩燕, 等. 基于 CUDA 的并行 K-means 聚类图像分割算法优化[J/OL]. 农业机械学报, 2014, 45(11): 47 - 53. http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx? flag = 1&file_no = 20141108&journal_id = jcsam. DOI: 10.6041/j.issn.1000-1298.2014.11.008.
HUO Yingqiu, QIN Renbo, XING Caiyan, et al. CUDA-based parallel K-means clustering algorithm[J/OL]. Transactions of the Chinese Society for Agricultural Machinery, 2014, 45(11):47 - 53. (in Chinese)

19 YOU E S, CHOI G H, KIM S H. Study on extraction of keywords using TF - IDF and text structure of novels[J]. Hermeneus, 2015, 20(2):121 - 129.

20 WU L, HOI S C, YU N. Semantics-preserving bag-of-words models and applications[J]. IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society, 2010, 19(7):1908 - 1920.

21 GERS F A, SCHMIDHUBER J, CUMMINS F, et al. Learning to forget: continual prediction with LSTM[C]// International Conference on Artificial Neural Networks. IET, 1999:850 - 855.

22 梁军, 柴玉梅, 原慧斌, 等. 基于极性转移和 LSTM 递归网络的情感分析[J]. 中文信息学报, 2015, 29(5):152 - 159.
LIANG Jun, CHAI Yumei, YUAN Huibin, et al. Polarity shifting and LSTM based recursive networks for sentiment analysis[J]. Journal of Chinese Information Processing, 2015, 29(5):152 - 159. (in Chinese)