计算机研究与发展
Journal of Computer Research and Development
ISSN 1000-1239/CN 11-1777/TP
50(12): 2570-2577, 2013

基于情感特征聚类的半监督情感分类

李素科　　蒋严冰
(北京大学软件与微电子学院　北京　100871)
(lisuke@ss.pku.edu.cn)

Semi-Supervised Sentiment Classification Based on Sentiment Feature Clustering

Li Suke and Jiang Yanbing
(*School of Software and Microelectronics*, *Peking University*, *Beijing* 100871)

Abstract　Sentiment classification for text is an important aspect of opinion mining. This paper proposes a semi-supervised sentiment classification method based on sentiment feature clustering. The method only requires a small number of labeled training data instances. Firstly, the method extracts common text features and sentiment features. Common text features can be used to train the first sentiment classifier. Then the spectral clustering-based algorithm is employed to map sentiment features into extended features. The extended features and common text features are combined together to form the second sentiment classifier. The two classifiers select instances from the unlabeled dataset into the training dataset to train the final sentiment classifier. Experimental results show that the proposed method can reach higher sentiment classification accuracy than both the self-learning SVM-based method and the co-training SVM-based method.

Key words　semi-supervised learning; sentiment feature clustering; sentiment classification; opinion mining; Web mining; data mining

摘　要　情感分类是观点挖掘的一个重要的方面.提出了一种基于情感特征聚类的半监督式情感分类方法,该方法只需要对少量训练数据实例进行情感类别标注.首先从消费者评论中提取普通分类特征和情感特征,普通分类特征可以用来训练一个情感分类器.然后使用 spectral 聚类算法把这些情感特征映射成扩展特征.普通分类特征和扩展特征一起通过训练得到另一个情感分类器.2 个分类器再从未标签数据集中选择实例放入到训练集合中,并通过训练得到最终的情感分类器.实验结果表明,在同样的数据集上该方法的情感分类准确度比基于 self-learning SVM 的方法和基于 co-training SVM 的方法的情感分类准确度要高.

关键词　半监督式学习;情感特征聚类;情感分类;观点挖掘;Web 挖掘;数据挖掘

中图法分类号　TP181; TP391.4

用户在 Web 上发表的产品评论一般是有情感色彩的,这些评论可能带有褒义的情感倾向,也可能带有贬义的情感倾向.情感分类在互联网推荐系统、产品竞争情报系统和商务智能系统中有着非常重要的应用.如何自动分类消费者评论的情感倾向已经成为观点挖掘领域的重要研究内容.

收稿日期:2013-06-26;**修回日期**:2013-09-08
基金项目:国家自然科学基金项目(61170002)
通信作者:蒋严冰(jyb@ss.pku.edu.cn)

研究者常用机器学习的方法来对消费者评论进行情感分类.机器学习的方法大体上分为监督式机器学习方法、半监督式机器学习方法和非监督式机器学习方法.Pang 等人[1]是最早使用监督式机器学习方法支持向量机(support vector machine,SVM)来对电影评论进行情感分类.Pang 对比了朴素贝叶斯(naïve Bayes, NB)、最大熵分类法(maximum entropy classification, MEC)和 SVM 三种机器学习方法在情感分类中的性能[1].也有一些研究者关注非监督方式的方法.例如 Turney 使用 PMI-IR (pointwise mutual information and information retrieval)方法进行消费者评论的情感分类[2].读者可以在文献[3]找到更多关于情感分类的方法的概述.也有研究使用模糊逻辑进行情感分类[4],还有使用图进行情感分类的等等[5].基于监督式学习的情感分类方法需要标注的训练数据,这种标注工作或许需要大量的人工劳动.而半监督式的情感分类方法仅仅需要少量标注的训练数据集,因此大部分时间可以减少人工标注量.最新的情感分类研究成果也有使用半监督学习方式进行情感分类的,例如文献[6]和文献[7]等.

本文提出一种基于特征聚类的半监督式情感分类方法.这种方法根据情感特征的共现关系构建共现矩阵,然后利用 spectral 聚类方法生成分类用的扩展特征,结合原有训练域内的分类特征来训练新的情感分类器;原有训练域的分类特征形成另一个分类器;2 个分类器共同完成最后的情感分类工作.本文提出的方法基于一个基本的假设:具有相似情感倾向的情感词有较高的概率出现在同样情感倾向的消费者评论中;那么这些"相似"的情感特征就可以通过其出现在评论中的共现关系进行聚类.

实验结果表明,本文提出的基于情感特征聚类的情感分类方法性能比基于 self-learning SVM 的方法和基于 co-training SVM 的情感分类方法的性能要好.

本文工作是作者已经发表的会议论文[8]的深化和扩展,但是文献[8]给出的实验相对薄弱,并没有与重要的基于 co-training SVM 的半监督方法进行对比实验,本文在文献[8]的基础上加入了与基于 co-training SVM 方法的对比实验.

1 基于情感特征聚类的半监督情感分类

传统机器学习的情感分类方法需要人工标注训练数据,也许需要大量的人力和时间.而基于半监督式的机器学习的情感分类方法只需要人工标注少量的训练数据,然后利用大量的没有标注情感类别的数据来训练学习器,从而分类未知情感类别的消费者评论.因为半监督的学习方法的训练数据集只有少量的训练数据,因此可能出现这样的情况:在测试数据集中的特征并没有出现在训练数据集中,这样会导致分类的结果并不一定理想.也就是说,这种特征分布的稀疏性会影响情感分类的准确度.但如能利用未标注情感类别数据中的情感特征,把情感特征进行聚类,可获得的聚类的个数是有限的.那么,这时把情感特征是否出现在某个聚类中这个二值数值表示为一个特征扩展到原有的训练特征集合中,可以缓解特征稀疏的问题.

本文基于特征聚类的半监督情感分类方法是根据一个观察:具有相同情感倾向的词汇或短语更容易出现在同样的产品评论中.例如,一般情况下"非常好"和"完美"出现在同样的具有褒义情感倾向的评论中的概率要大于"非常差"和"完美"共同出现在一个评论中的概率.根据这样的观察,本文认为如果能对这些具有共现关系的情感特征进行聚类将有助于对产品评论的分类.

为了便于对本问题提出的方法的进一步说明,本文先给出 2 个定义.

定义 1. 情感特征.情感特征指的是具有情感类别表征的特征.

定义 2. 名词单元.名词单元是在子句中包括最大数量连续名词的序列.

一般情况下,形容词和动词是有情感色彩的,例如,中文的"好"、"酷"和"喜欢"等词,英文中的"lovely","friendly"和"love"等词,这些形容词和动词就是情感特征.还有一些情感特征是出现在否定上下文中的,也就是有一些否定指示词与这些形容词和动词共同构成情感特征.例如,"not good"中的"good"表示褒义情感特征,而"not"是否定指示词.本文把带有否定指示词的形容词和动词也当成一个情感特征.在本文中,抽取词性标注(part-of-speech, POS)标签是 JJ,JJR,JJS,VB,VBN,VBG,VBZ,VBP 的形容词和动词,且这些词在上下文窗口[-3,3]之内有标签为 NN 或者 NNS 的词,那么这些形容词和动词为情感特征.一个词的上下文窗口[-3,3]表示在一个子句中的这个词的左右各 3 个单词覆盖的范围.与此同时,如果在情感的上下文窗口[-3,0]内有否定指示符,那么否定指示符和这

个情感词共同构成情感特征.这些否定指示词包括“not”,“no”,“donot”,“don't”,“didn't”,“didnt”,“wasn't”,“wasnt”,“isn't”,“isnt”“weren't”,“werent”,“doesn't”“doesnt”,“hardly”,“never”,“neither”和“nor”等.

本文中规定:2 个情感特征共现表示这 2 个情感特征的上下文窗口有同样的名词单元出现,否则不能称为共现.例如,2 个情感特征虽然出现在同样的一个消费者评论中,但是由于这 2 个情感特征的上下文窗口并没有同样的名词单元,那么这 2 个情感特征的共现频率就是 0.同样,共现频率指的是在整个训练及未标签数据集上的共现频率,而不是说仅仅在某一个消费者评论中出现的频率.图 1 说明了本文算法的基本思想和步骤.在图 1 中,首先该算法先标注少量的褒义和贬义的消费者评论,根据这些评论得到一般的分类训练特征(本文中用的是 uni-gram),再用这些特征使用 SVM 训练得到情感分类器 f_1.

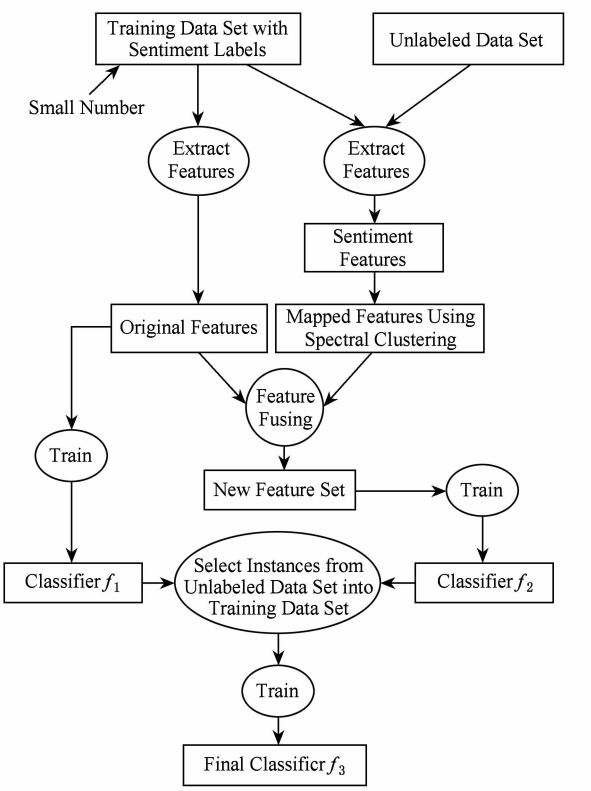

Fig. 1 A framework of semi-supervised sentiment classification based on sentiment feature clustering.

图 1 基于情感特征聚类的半监督情感分类方法框架

从已经标注情感类别的评论和未标注情感类别的评论中抽取情感特征,这些特征使用 Spectral 聚类算法进行特征映射得到新的特征,把这些新的特征与原来的训练特征(uni-gram)进行特征融合,用融合后特征使用 SVM 进行训练得到情感分类器 f_2.最后,由分类器 f_1 和分类器 f_2 从未标签情感类别的数据集合中选择 f_1 和 f_2 的分类结果一致的未标签评论到训练集合中,用这个新的训练集合得到最终的分类器 f_3.这时,就可以使用情感分类器 f_3 来分类新评论的情感类别.

本文的方法实际上利用 2 个分类器 f_1 和 f_2,假设分类器 f_1 的分类准确度为 p,f_2 的分类准确度为 q,那么对于这 2 个分类器而言,把 1 个消费者评论 r 都分类正确的概率为 $x_1=pq$.2 个分类器把 1 个消费者评论 r 都分类错误的概率是 $x_2=(1-p)(1-q)$.在本文的方法中实际上需要的是分类器 f_1 和分类器 f_2 分类结果一致的结果.就某一个消费者评论,如果 2 个分类器的分类结果是一样的,那么要么 2 个分类器都分类为正确的情感类别,要么 2 个分类器都分类为错误的情感类别.因此,由 2 个分类器分类结果一致的概率是 $A=pq+(1-p)(1-q)=1-p-q+2pq$.同样,分类器 f_1 把评论分类正确,而分类器 f_2 把结果分类错误的概率是 $y_1=p(1-q)$.分类器 f_2 把评论分类正确,而分类器 f_1 把结果分类错误的概率是 $y_2=(1-p)q$,因此,结果分类不一致的概率是 $B=y_1+y_2=p(1-q)+(1-p)q=p+q-2pq$ 因此,可以得到 A 与 B 的比值:$l(p,q)=A/B=[1-(p+q-2pq)]/(p+q-2pq)$.如果可以把分类器的分类准确度看作连续的变量,分类器 f_1 的准确度对应的变量是 p,分类器 f_2 对应的变量是 q.如果 2 个分类器的准确度的下限是 a 和 b,也就是 $a<p<1$ 且 $b<q<1$ 时,$\rho(x,y)=1$;其他情况下 $\rho(x,y)=0$.假设 p 和 q 满足均匀分布,那么,$X=pq$ 的数学期望 $\boldsymbol{E}(X)$ 为

$$\boldsymbol{E}(X)=\int_a^1\int_b^1 pq\rho(p,q)\mathrm{d}p\mathrm{d}q=\frac{(1-a^2)(1-b^2)}{4}. \tag{1}$$

同样得到 $Y=(1-p)(1-q)$ 的数学期望 $\boldsymbol{E}(Y)$ 为

$$\boldsymbol{E}(Y)=\int_a^1\int_b^1(1-p)(1-q)\rho(p,q)\mathrm{d}p\mathrm{d}q=\left[(1-a)-\frac{(1-a^2)}{2}\right]\left[(1-b)-\frac{(1-b^2)}{2}\right]. \tag{2}$$

但如果假设分类器的分类准确度均大于 1/2,也就是当 $1/2<p<1$ 且 $1/2<q<1$ 时,$\rho(x,y)=1$;其他情况下 $\rho(x,y)=0$.那么:

$$E(X)=\int_{\frac{1}{2}}^1\int_{\frac{1}{2}}^1 pq\rho(p,q)\mathrm{d}p\mathrm{d}q=9/64. \tag{3}$$

同样,求 $Y=(1-p)(1-q)$ 的数学期望为

$$\boldsymbol{E}(Y)=\int_{\frac{1}{2}}^{1}\int_{\frac{1}{2}}^{1}(1-p)(1-q)\rho(p,q)\mathrm{d}p\mathrm{d}q=1/64. \tag{4}$$

这时可以得到 $\boldsymbol{E}(X)$与 $\boldsymbol{E}(Y)$的比值 T 为

$$T=\frac{\boldsymbol{E}(X)}{\boldsymbol{E}(Y)}=9. \tag{5}$$

因此,如果 2 个分类器的分类准确度都大于 0.5,那么这 2 个分类器把一个未标注类别的评论都分类为正确的概率的数学期望是都分类为错误的数学期望的 9 倍.

本文算法是基于 spectral 聚类方法[9-11]的. spectral 聚类算法被广泛应用在文本聚类问题中. 本 spectral 聚类算法有很多变种,但本文使用了相对简单的一种,对 spectral 聚类方法的描述如算法 1 所示. 在算法 1 中,相似矩阵 $\boldsymbol{S}$ 在实数域 $i^{n\times n}$ 中. spectral 聚类算法需要从相似矩阵中计算 Laplacian 矩阵 $\boldsymbol{L}$,并从这个矩阵得到特征向量,然后对这些特征列向量进行聚类,最终得到聚类的结果. spectral 算法容易实现且聚类性能较好.

算法 1. Spectral 聚类算法[9-11].

输入:相似矩阵 $\boldsymbol{S}\in i^{n\times n}$、最终得到的聚类个数 k;

输出:k 个类.

开始

① 计算对角矩阵 $\boldsymbol{D}$,其中 $\boldsymbol{D}_i=\sum_{1\leqslant j\leqslant n}\boldsymbol{S}_{ij}$;

② 计算 Laplacian 矩阵 $\boldsymbol{L}$,$\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{S}$;

③ 计算矩阵 $\boldsymbol{L}$ 的前 k 个特征向量 $\boldsymbol{u}_1,\boldsymbol{u}_2,\cdots,\boldsymbol{u}_k$,$\boldsymbol{u}_i$ 是列向量,这些特征向量形成矩阵 $\boldsymbol{V}$,且 $\boldsymbol{V}\in i^{n\times n}$,$\boldsymbol{y}_i$ 表示矩阵 $\boldsymbol{V}$ 的第 i 个行向量;

④ 使用 k-means 对$(\boldsymbol{y}_i)_{i=1..n}$进行聚类,得到 k 个类 $C=\{\boldsymbol{c}_1,\boldsymbol{c}_2,\cdots,\boldsymbol{c}_k\}$;

⑤ 返回 C;

结束

本文的基于特征聚类的半监督学习方式的情感分类方法的核心算法如算法 2 所示. 算法 2 涉及到的 spectral 思想和步骤已经在图 1 中详细说明. 在算法 2 中,共现矩阵 $\boldsymbol{S}$ 中的元素 $\boldsymbol{S}_{ij}$ 表示情感特征 i 和情感特征 j 共现的次数. 如果 $\boldsymbol{S}_{ij}=0$,那么表示情感特征 i 和情感特征 j 无共现. 对角矩阵 $\boldsymbol{D}$ 是由共现矩阵 $\boldsymbol{S}$ 获得. 每个训练实例(经过预处理的消费者评论)x_t 都对情感分类结果 y_t,$y_t\in\{-1,+1\}$,其中 -1 表示消极情感类别,$+1$ 表示积极情感类别. 未标签情感类别的实例在数据集合 U 中.

算法 2. 基于情感特征聚类的情感分类算法.

输入:训练数据集合 $\mathrm{T}=\{(\boldsymbol{x}_t,y_t)_{t=1}^{m}\}$、未标签数据集合 $U=\{\boldsymbol{u}_1,\boldsymbol{u}_2,\cdots,\boldsymbol{u}_y\}$;

输出:情感分类器 f.

开始

① 根据情感特征抽取规则从 T 和 U 抽取情感特征;由训练数据集合 T 构建训练情感特征矩阵 $\boldsymbol{M}$,$\boldsymbol{M}\in i^{m\times n}$,$m$ 是训练数据集的实例个数,n 表示情感特征的个数,在未标签数据集合上构建矩阵 $\boldsymbol{N}$,$\boldsymbol{N}\in i^{z\times n}$,$z$ 是未标签数据实例的个数;

② 根据情感特的共现频率构建共现矩阵 $\boldsymbol{S}$,$\boldsymbol{S}\in i^{n\times n}$;/* 如果 2 个情感特征共同出现在同样的评论中,且其上下文都有相同的名词单元出现,称它们是共现的 */

③ 获得对角矩阵 $\boldsymbol{D}$,$\boldsymbol{D}_i=\sum_{1\leqslant j\leqslant n}\boldsymbol{S}_{ij}$

④ 获得 Laplacian 矩阵 $\boldsymbol{L}$,$\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{S}$;

⑤ 计算矩阵 $\boldsymbol{L}$ 的前 k 个最大特征值的对应的特征向量 $\boldsymbol{u}_1,\boldsymbol{u}_2,\cdots,\boldsymbol{u}_k$,$\boldsymbol{u}_i$ 是列向量,这些特征向量形成矩阵 $\boldsymbol{V}$,$\boldsymbol{V}\in i^{n\times n}$;

⑥ 计算矩阵 $\boldsymbol{B}=\boldsymbol{M}\times\boldsymbol{V}$,$\boldsymbol{B}\in i^{m\times k}$,如果 $\boldsymbol{B}$ 的元素大于 0,那么把它设置为 1;

⑦ 计算矩阵 $\boldsymbol{E}=\boldsymbol{N}\times\boldsymbol{V}$,$\boldsymbol{E}\in i^{z\times k}$,如果 $\boldsymbol{E}$ 的元素大于 0,那么把它设置为 1;

⑧ 根据训练数据集合,使用 SVM 分类器训练获得情感分类器 f_1;

⑨ 使用 SVM 在经过扩展特征后的数据集$\{([\boldsymbol{x}_t,\boldsymbol{B}_t],y_t)_{t=1}^{m}\}$上训练获得情感分类器 f_2;

⑩ 使用分类器 f_1 分类未标注情感类别的数据集合 U,得到分类结果 $\boldsymbol{x}$,$\boldsymbol{x}$ 是向量;

⑪ 使用分类器 f_2 分类经过扩展特征后的数据集$\{([u_t,\boldsymbol{E}_t],y_t)_{t=1}^{z}\}$,得到分类结果 $\boldsymbol{y}$,$\boldsymbol{y}$ 是向量;

⑫ FOR $i=1$ TO $|\boldsymbol{x}|$ DO

⑬ IF $\boldsymbol{x}_i==\boldsymbol{y}_i$ THEN

⑭ $T=T\cup\{\boldsymbol{u}_i\}$;

⑮ ENDIF;

⑯ ENDFOR;

⑰ 使用 SVM 在 T 上训练新的情感分类器器 f;

⑱ 返回 f;

结束

2 实验与评价

2.1 数据准备

实验数据来自 Amazon① 和 Tripadvisor②. 其中,手机和笔记本电脑的评论来自 Amazon,关于酒店的评论来自 Tripadvisor. 这些消费者评论经过爬虫下载后由程序自动抽取出来,然后使用开源软件 OpenNLP③ 对这些评论进行句子切分和生成 POS 标签. 关于实验数据的统计如表 1 所示. 每种数据都包括 2 000 个评论,其中每种类型的评论数据集合包括 1 000 个正面评论和 1 000 个反面评论. 表 1 也给出了从每种评论数据集合中获得的句子的个数.

Table 1 Dataset

表 1 数据集

Data	Positive Review #	Negative Review #	Sentence #	Source
phone	1 000	1 000	28 811	Amazon
laptop	1 000	1 000	14 814	Amazon
hotel	1 000	1 000	18 694	Tripadvisor

2.2 对比的方法

2.2.1 基于 self-learning SVM 的方法

最早使用机器学习方法 SVM 进行情感分类的研究工作是 Pang 等人的研究[1]. 本文把评论分解成 uni-gram 来充当 SVM 的训练特征;SVM 的特征值就是这个 uni-gram 在评论中出现的频率. self-learning SVM 是一种自举的学习方式,其基本思想就是先用少量标注了情感类别的消费者评论作为学习训练数据集,用这些训练集学习获得的分类器来分类未标签的消费者评论的情感属性. 最后,从分类器的分类结果中选择最有可能是正确的评论,放入到学习训练集中. 用获得的新训练数据再次训练情感分类器. 重复以上步骤,直到没有未分类标签可以加入到训练数据集为止. 有很多关于其他的 SVM 的主动学习方法[12]. 但本文中的基于 self-learning SVM 的方法其实是比较简单的一种主动学习的方法,因为这里仅仅考虑未标注情感类别的数据集中的实例距离 SVM 的分类超平面的距离. 距离越大,越被认为可能分类是正确的. 基于 self-learning SVM 方法的情感分类算法如算法 3 所示,用到的 SVM 工具是 TinySVM④.

算法 3. 基于 self-learning SVM 的情感分类.

输入:训练集 $L=\{\boldsymbol{l}_1,\boldsymbol{l}_2,\cdots,\boldsymbol{l}_x\}$,$L$ 包括积极情感评论以及消极情感评论;未标注情感类别的数据集 $U=\{\boldsymbol{u}_1,\boldsymbol{u}_2,\cdots,\boldsymbol{u}_y\}$;

输出:情感分类器 C;

开始

① d 是选择最可能分类结果的个数;

② 使用 SVM 在 L 上得到情感分类器 C;

③ WHILE{有未标签的数据记录}

④ 使用情感分类器 C 分类未标注情感类别的数据集合 U 中的记录:得到积极集合 P 和消极集合 N;

⑤ 如果 $|P|\geqslant d$,那么从 P 中选择 d 个最可能分类正确的记录放入到 L 中,$L=L\cup P_d$;否则把 P 中的全部记录放入到 L 中,$L=L\cup P$;

⑥ 如果 $|N|\geqslant d$,那么从 N 中选择 d 个最可能分类正确的记录放入到 L 中,$L=L\cup N_d$;否则把 N 中的全部记录放入到 L 中,$L=L\cup N$;

⑦ 使用新的训练集 L 训练得到情感分类器 C;

⑧ ENDWHILE;

⑨ 返回 C;

结束

2.2.2 基于 co-training SVM 的方法

co-training[13] 是可用来实现半监督文本分类的算法. co-training 算法是典型的自举算法. co-training 算法与主动学习算法类似:首先标注少量训练数据集,然后用标注的训练数据集以及未标注的数据集进行学习训练学习器. 目前已经存在一些半监督方式的使用基于 co-training 方法进行情感分类算法[14-15],这些研究说明 co-training 框架方法在情感分类问题中是有效的. 为了证明本文方法的有效性,本文实现了基于 co-training[13] 的算法,如算法 4 所示.

本文实现的 co-training 方法需要 3 个分类器. 首先把文本进行预处理,对文本进行分词,也就是获得所有的 uni-grams. 去掉其中的长度小于 3 的 uni-gram 和停止词,剩下的 uni-gram 在每个文本中出现的频率作为训练的特征. 那么所有的分词单元得到

① http://www.amazon.com ② http://www.tripadvisor.com

③ http://opennlp.apache.org/ ④ http://chasen.org/~taku/software/TinySVM/

的分类器为 f_1.为了获得第 2 个分类器 f_2,获得消费者评论的 POS 标签,把这些标签为"JJ","JJR","JJS","VB","VBN","VBG","VBZ",以及"VBP"的词看成是情感词.如果在其左边窗口范围[-3,0]之内的否定指示词,那么情感词与否定指示词共同构成一个否定的情感特征.如果在其左边窗口内未出现否定词,那么这个情感词就单独构成一个情感特征.这些否定词包括:"not","no","donot","don't","didn't","didnt","wasn't","wasnt","isn't","isnt""weren't","werent","doesn't""doesnt","hardly","never","neither"和"nor"等.如果情感单元在所在的评论里面出现,那么特征值就是 1,否则特征值是 0.为了获得第 3 个分类器 f_3,首先根据第 2 个分类器 f_2 获得情感特征的方法获得所有的情感特征,然后由式(6)从初始训练集中(只从初始训练数据集中计算)求出这些情感特征的平均评分:

$$f_{i_r}=\frac{\sum\limits_{1\leqslant i\leqslant|R_i|}r_i}{|R_i|},\tag{6}$$

这里 f_{i_r} 是特征 f_i 的平均评分,R_i 表示包含特征 f_i 的评论集合,r_i 表示包含特征 f_i 的评论的分数.这些评论的平均分数作为情感特征的值来训练获得分类器 f_3.

算法 4. 基于 co-training SVM 的情感分类算法.

输入:训练集 $L=\{\boldsymbol{l}_1,\boldsymbol{l}_2,\cdots,\boldsymbol{l}_x\}$,$L$ 包括积极情感评论以及消极情感评论;未标注情感属性的数据集 $U=\{\boldsymbol{u}_1,\boldsymbol{u}_2,\cdots,\boldsymbol{u}_y\}$;

输出:情感分类器 C;

开始

① WHILE{有未标签的数据记录}

② 使用标注了情感类别的训练集 L 训练得到分类器 f_1;

③ 根据情感特征得到分类器 f_2;

④ 根据由式(6)得到的初始训练集情感特征评分得到分类器 f_3;

⑤ 使用分类器 f_1 分类未标注情感类别的数据集 U,获得情感类别为积极的集合 P_{f_1} 以及分类情感属性为消极的集合 N_{f_1}(取最有可能的前 50 个分类结果);

⑥ 使用分类器 f_2 分类未标注情感类别的数据集 U,获得情感类别为积极的集合 P_{f_2} 以及分类情感属性为消极的集合 N_{f_2}(取最有可能的前 50 个分类结果);

⑦ 使用分类器 f_3 分类未标注情感类别的数据集 U,获得情感类别为积极的集合 P_{f_3} 以及分类情感类别为消极的集合 N_{f_3}(取最有可能的前 50 个分类结果);

⑧ $L=L\cup P_{f_1}\cup N_{f_1}\cup P_{f_2}\cup N_{f_2}\cup P_{f_3}\cup N_{f_3}$;

⑨ ENDWHILE;

⑩ 返回 C;

结束

2.3 实验结果

基于 self-learning SVM 的方法使用的特征是去除了停止词的 uni-gram. co-training SVM 方法和本文提出的方法在前面已经给出.在每种实验数据集上的实验数据都被划分为训练数据集、未标签数据集和测试数据集.首先对每种类型数据从数据集中随机抽样 800 个实例作为测试数据集,800 个测试实例中包括 400 个褒义实例和 400 个贬义实例.实验的每种类型产品的训练数据集实例的个数分为 100,200,300 和 400 四种情况,每种划分都是平衡数据集.例如,对于手机(phone)数据集(有 2 000 个实例)而言,如果训练数据集中的实例个数如果是 100,表示有 50 个情感为褒义的实例,还有 50 个情感为贬义的实例;其他 800 个随机采样的平衡实例集作为测试数据集,剩余 1 100 个数据实例属于未标签数据集.

本文实验所用到的 SVM 实现是 TinySVM (http://chasen.org/~taku/software/TinySVM/).本文实验的 3 种方法均使用 SVM 分类未标签数据集中的实例,然后把最有可能被正确分类的实例放到训练集中.选择距离 SVM 超平面最远的实例为最有可能分类正确的实例.评价方法为准确度,即

$$Accuracy=\frac{\text{测试集中分类正确的实例个数}}{\text{测试集的实例个数}}.\tag{7}$$

图 2 是在手机数据集上得到的实验结果.从图 2 可以看出,基于 co-training SVM 的方法结果要比基于 self-learning SVM 的方法性能高,但仍然不如本文提出的方法.

本文提出的方法在训练实例个数为 100,200,300,400 这 4 种情况下得到的准确度明显比其他 2 种方法高.

图 3 是在笔记本电脑(laptop)数据集上得到的实验结果.从图 3 可以看出,基于 co-training SVM 的方法结果和基于 self-learning SVM 的方法得到的准确度较为接近.当训练实例大于 200 时,基于

co-training SVM 方法的分类准确度比基于 self-learning SVM 方法得到的准确度稍微好点,但仍然不如本文提出的方法.很明显,本文提出的方法在笔记本电脑数据集上得到的准确度在所有情况下得到的准确度明显比其他 2 种方法高.

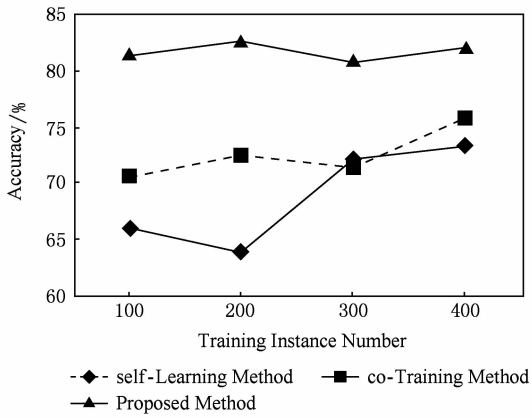

Fig. 2 Experimental results on the phone dataset.

图 2 在 phone 数据集上的分类结果

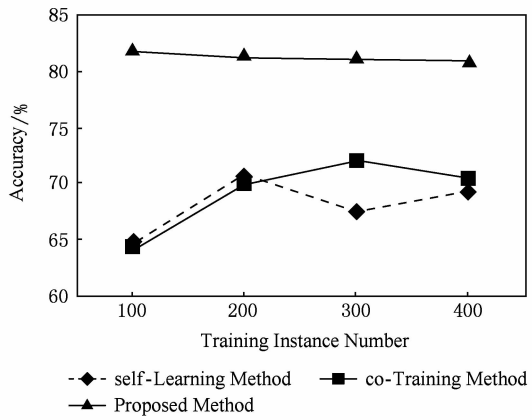

Fig. 3 Experimental results on the laptop dataset.

图 3 在 laptop 数据集上的分类结果

图 4 是在酒店(hotel)数据集上得到的实验结果.从图 4 可以看出,本文提出的基于情感特征聚类方法的准确度在所有情况下得到的准确度比基于 self-learning SVM 方法和基于 co-training SVM 方法在所有实验情况下都要高.当训练数据数量在 300 时,基于 self-learning SVM 方法和基于 co-training SVM 方法得到的准确度较为接近.然而总体而言,基于 co-training SVM 方法要比基于 self-learning SVM 方法要好.

综上所述,本文提出的半监督方式情感分类方法在同样数据集上分类性能要比基于 self-learning SVM 的情感分类方法和基于 co-training SVM 的情感分类方法要好.

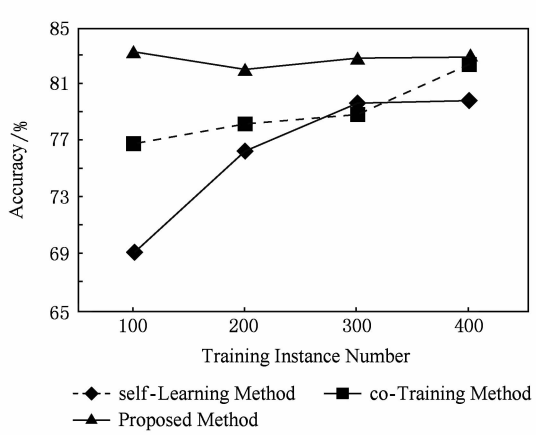

Fig. 4 Experimental results on the hotel dataset.

图 4 在 hotel 数据集上的分类结果

3 相关工作比较和讨论

在文献[14]中,产品评论的内容被分为 2 个视图:一个视图是个人视图(personal view),这种视图是表达了发言者对某个对象的情感;而另外一个视图是非个人视图(impersonal view),表达了对评论对象的评价.利用这 2 个视图,结合 co-training 的算法[13],获得最终的情感分类器.然而,文献[14]的方法在进行情感分类之前必须先构建这 2 个视图,而构建这 2 个视图前必须确定符合条件的文本或句子属于哪个视图.Wan[15]也使用 co-training 的方法实现跨语言情感分类问题,可见基于 co-training 方法在最新的情感分类研究中起着重要的作用.Dasgupta 和 Ng[16]使用频谱技术来挖掘情感明确的评论,然后结合主动学习(active learning)、转换学习(transductive learning)和组装学习(ensemble learning)的方法来分类新的评论的情感倾向.文献[17]使用基于二部图的联合文档和单词情感分析方法最终完成半监督方式的情感分类,但其性能上并不比当时其他半监督方式学习方式更好.Zhou 等人[6]使用深度主动网络进行半监督方式情感分类,也取得了不错的效果,但算法复杂度相对较高.综合考虑,本文的方法比起其他方法而言较为容易实现而且算法复杂度也能接受.

4 结 论

本文提出了一种基于情感特征聚类的半监督情感分类方法.这种方法只需要使用少量训练数据和一些未标签情感类别的数据.首先从标签情感类别

的训练数据集合和未标签情感类别的数据实例中提取情感特征,然后使用 Spectral 算法把情感特征映射成扩展特征.使用普通的文本特征得到一个分类器,普通的文本特征和经过扩展的情感特征合起来训练后得到另一个分类器.用这 2 个分类器未标签情感类别的数据集合进行分类,把分类结果一致的数据实例放入到训练集合中.用新的数据训练得到最终的情感分类器.实验结果表明,在同样数据集上本文方法的情感分类性能优于基于 self-learning SVM 的情感分类方法及基于 co-training SVM 的情感分类方法.

参 考 文 献

[1] Pang B, Lee L, Vaithyanathan S. Thumbs up? sentiment classification using machine learning techniques [C] //Proc of the 40th Annual Meeting on Association for Computational Linguistics (ACL'02). Stroudsburg, USA: Association for Computational Linguistics, 2002: 79-86

[2] Turney P. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews [C] //Proc of the 40th Annual Meeting on Association for Computational Linguistics (ACL'02). Stroudsburg, USA: Association for Computational Linguistics, 2002: 417-424

[3] Pang B, Lee L. Opinion mining and sentiment analysis [J]. Foundations and Trends in Information Retrieval, 2008, 2(1/2): 1-135

[4] Subasic P, Huettner A. Affect analysis of text using fuzzy semantic typing [J]. IEEE Trans on Fuzzy Systems, 2001, 9(4): 417-424

[5] Rakesh A, Rajagopalan S, Srikant R, et al. Mining newsgroups using networks arising from social behavior [C] //Proc of the 12th Int Conf on World Wide Web (WWW'03). New York: ACM, 2003: 529-535

[6] Zhou S, Chen Q, Wang X. Active deep networks for semi-supervised sentiment classification [C] //Proc of the 23rd Int Conf on Computational Linguistics: Posters (COLING'10). Stroudsburg, USA: Association for Computational Linguistics, 2010: 1515-1523

[7] Xia R, Zong C, Li S. Ensemble of feature sets and classification algorithms for sentiment classification [J]. Information Sciences, 2011, 181(6): 1138-1152

[8] Li S, Hao J. Spectral Clustering-Based Semi-supervised Sentiment Classification [G] //LNCS 7713: Proc of the 8th Advanced Data Mining and Applications. Berlin: Springer, 2012: 271-283

[9] Mohar B. The Laplacian spectrum of graphs [J]. Graph Theory Combinatorics, and Applications, 1991, 2: 871-898

[10] Mohar B, Juvan M. Graph Symmetry: Algebraic Methods and Applications [M]. Berlin: Springer, 1997: 227-275

[11] Ng A, Jordan M, Weiss Y. Advances in Neural Information Processing Systems 14 [M]. Cambridge, USA: MIT Press, 2001: 849-856

[12] Tong S, Koller D. Support vector machine active learning with applications to text classification [J]. Journal of Machine Learning Research, 2002, 2: 45-66

[13] Blum A, Mitchell T. Combining labeled and unlabeled data with co-training [C] //Proc of the 11st Annual Conf on Computational Learning Theory. New York: ACM, 1998: 92-100

[14] Li S, Huang C, Zhou G, et al. Employing personal/impersonal views in supervised and semi-supervised sentiment classification [C] //Proc of the 48th Annual Meeting on Association for Computational Linguistics (ACL'10). Stroudsburg, USA: Association for Computational Linguistics, 2010: 414-423

[15] Wan X. Co-training for cross-lingual sentiment classification [C] //Proc of Joint Conf of the 47th Annual Meeting of the ACL and the 4th Int Joint Conf on Natural Language Processing of the AFNLP. Stroudsburg, USA: Association for Computational Linguistics, 2009: 235-243

[16] Dasgupta S, Ng V. Mine the easy, classify the hard: a semi-supervised approach to automatic sentiment classification [C] //Proc of the Joint Conf of the 47th Annual Meeting of the ACL and the 4th Int Joint Conf on Natural Language Processing of the AFNLP. Stroudsburg, USA: Association for Computational Linguistics, 2009: 701-709

[17] Sindhwani V, Melville P. Document-Word Co-regularization for Semi-supervised Sentiment Analysis [C] //Proc of the 8th IEEE Int Conf on Data Mining (ICDM'08). Piscataway, NJ: IEEE, 2008: 1025-1030

Li Suke, born in 1977. Received his PhD degree in computer science from Peking University, China in 2012. Currently assistant professor in Peking University, China. His research interests include financial data mining, Web mining and retrieval, opinion mining, social networks, and information security.

Jiang Yanbing, born in 1975. PhD of computer science. Associate professor of Peking University. His research interests include software development methodology, object-oriented software development technology, model-driven software development technology, "Cloud + End" mobile Internet software development technology and software reconstruction technology(jyb@ss.pku.edu.cn).