微博观点识别及其情感分析

张红喜 吴明晖 金苍宏

浙江大学城市学院，浙江 310015；

通信作者，E-mail:hx.zhang@live.cn

摘要 针对微博的情感分析的观点句识别和观点句情感倾向的分类任务，依照决策树模型，在词、句、篇章的层面提取分类特征，提出了一种基于神经网络算法的分类策略。由《自然语言处理与中文计算会议》所提供的已标注样例用逐步减小 η 值的训练方法得到分类器参数，然后在分类器中增加松弛因子以期达到更好的分类效果。实验结果表明，算法具有较高的准确率。

关键字 情感分析；决策树；神经网络；松弛因子

Recognition And Sentiment Analysis O f Perspective Sentences From Microblog

Abstract In order to fulfill the task of recognition and sentiment analysis of perspective sentences from microblog, we propose an algorithm which based on Neural Network Algorithm, and use information come from word, sentence and chapter. Classifier's parameters are trained by use date provided by nlpcc2012, in the training process, we decrease the value of η , Simultaneously, we add a Slack factor to get a better result. The following experiment demonstrates that, the algorithm perform excellent in accuracy rate.

Key words Sentiment Analysis; decision tree; neural network; Slack factor

由于许多文本情感分析存在的许多问题和其广阔的应用情景，引起学术界和商业界的关注。[1] 在微博评论中，观点据表达了作者对于某对象的情感，是情感分析的一个重要对象。作者的情感状态可以是赞赏、批评；支持、反对；高兴、沮丧等。对于观点句的情感状态，大体上可以将它们分为正面态度和负面态度两大类。情感分析的任务就是区分文本中正面和负面的语言描述。按照这样的思路，分析相关微博的评论分析，从中发现具有评论观点的句子，将无观点句与有观点句进行分类。在此基础上，再对观点句解析，辨析其属于正面观点还是属于负面观点。用决策树模型来描述算法过程：微博评论首先由观点分类器进行分类，得到观点句和非观点句；然后观点句交由情感分类器，判断其情感的正负性。分类器的设计包括 svm，n-gram，遗传算法等分类模型，比如类似微博的 Twitter，Pak[4] 和 Paroubek[4] 在其情感分析中使用了 n-gram 算法。

1 分类模型

算法采用决策树的分类策略，先用观点句式别分类器从微博文本中区分观点句与非观点句，然后利用情感正负分类器给观点句进行正负面判断。建立文本的合适的向量模型，在次基础上，可建立文本的相应度量。[3] 两个分类器的设计思路都使用了神经网络的方法进行参数的设置。分类器也是建立在文本向量化的基础上。算法模型如下所示：

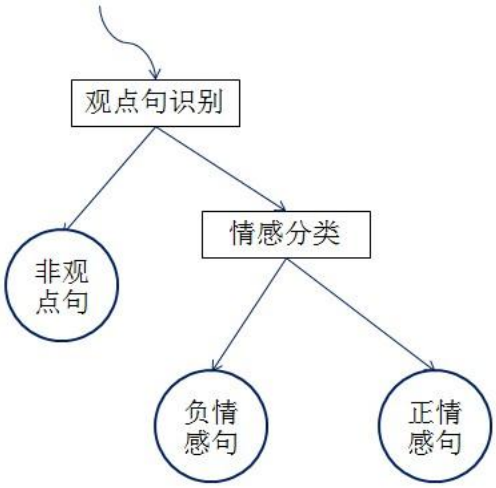

图 1 算法模型

1.1 神经网络感知器

1.1.1 训练法则

设输入向量 $X=(x_0,x_1,x_2,...,x_n)^T$，则 n 个输入分量在几何上构成一个 n 维空间。由方程 $\sum_{i=0}^{n} w_i x_i=0$ 确定一个 n 维空间上的超平面。此平面可以将输入样本分为两类。一个最简单的单计算节点感知器具有分类功能。其分类原理是将分类知识存储于感知器的权向量（包含了阈值）中，由权向量确定的分类判决界面将输入模式分为两类。

对于处理层中任一节点，其净输入为所得的文档向量，输出 o_j 为节点净输入与阈值之差的函数，离散型单计算层感知器的转移函数一般采用符号函数。

$$o = sgn(\sum_{i=0}^{n} w_i x_i) \quad (1)$$

其中权向量 $W=(w_0,w_1,w_2,...,w_n)^T$ 是决定神经网络感知器输出分类的直接影响因素。在具有实例样本的情况下，我们可以用感知器法则来训练权向量的合适取值。其中，我们假

设训练空间用 D 来表示。

$$w_i = w_i + \Delta w_i \tag{2}$$

$$\Delta w_i = \sum_{x \in \mathrm{D}} \eta (t - o) x_i \tag{3}$$

t 是当前训练样例的目标输出，o 是感知器的输出，η 是一个正常数称为学习速率。这种权值的学习方法叫梯度下降法则。

1.1.2 训练原理

尽管很多方法用于定义误差，一个常用的训练误差的度量指标为

$$E[\vec{w}] = (1/2)\sum_{x \in D} (\mathrm{t}_x - \mathrm{o}_x)^2 \tag{4}$$

由几何知，这是一个多维空间的抛物曲面，具有单一全局最小值，如图所示，

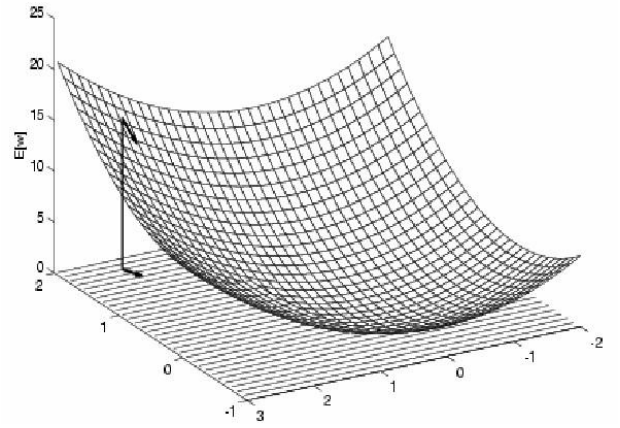

图 2 误差抛物曲面

当然，具体的抛物面形状依赖于具体的训练样例集合。

梯度下降法则的目的是将误差沿误差曲面最陡的的方向下降，可以通过计算 E 相对 $\vec{\mathrm{W}}$ 每个分量的导数来得到这个方向。这个方向导数叫做 E 对与 $\vec{\mathrm{W}}$ 的梯度。

$$\nabla E(\vec{w}) = \left(\frac{\partial E}{\partial w_0}, \frac{\partial E}{\partial w_1}, \ldots, \frac{\partial E}{\delta w_{\mathrm{n}}}\right) \tag{5}$$

既然梯度去顶了 E 最陡峭的方向，那么梯度下降的训练法则是：

$$\vec{w} = \vec{w} + \Delta \vec{w} \tag{6}$$

$$\Delta \vec{w} = -\eta \nabla E(\vec{w}) \tag{7}$$

其中：

$$\Delta \vec{w}_i = -\eta \frac{\partial E}{\partial w_i} \quad (10)$$

$$\frac{\delta E}{\delta w_i} = \frac{\partial}{\partial w_i}(1/2)\sum_{x\in D}(\mathrm{t}_x - \mathrm{o}_x)^2 \quad (11)$$

$$= \sum_{x\in D}(\mathrm{t}_x - \mathrm{o}_x)\frac{\partial}{\partial w_i}(t_x - \vec{w}.\vec{x})$$

$$= \sum_{x\in D}(\mathrm{t}_x - \mathrm{o}_x)(-x_i)$$

将公式（11）带入公式（7），然后将公式（7）带入（6）便得到了梯度下降权值更新法则，

$$\vec{w} = \vec{w} + \sum_{x\in D}(\mathrm{t}_x - \mathrm{o}_x)(-\vec{x}) \quad (12)$$

梯度下降法则的优点是，在训练样本线性可分得情况下，算法可以保证收敛到满足条件的权值；而在不可分的情况下，将收敛到目标概念的最佳近似。

2 算法详情

2.1 特征选择

词语级特征

词语级的特征选取须根据各分类器所分类对象的特点来设置。分析观点句的结构，发现其一般具有被评价对象[2]（指代）、判断词、评价词，其中，评价词是具有情感倾向的词，代表了评论者的对被评论对象的情感取向，这类词语对于分析作者观点具有非常关键的作用。例如，“他是国家的栋梁之材。”句中“栋梁之材”表达了作者对车的褒义评价。而作为观点句，各要素表现在：“这”作为车的指示代词，“是”为判断词，“栋梁之才”为评价词。也有诸如被评价对象+评价词，或者直接评价词等句式。情感词不仅可以是形容词，动词、名词、副词都可以成为情感词。在情感词部分中，副词往往充当了非常重要的角色，如“他真的是太幼稚了。”，“幼稚”表达了负面的情感，而前面“太”作为副词，给感情色彩取到了极大地加强的作用。

句子级特征

词语的情感倾向判断有其局限性，由于语言表述丰富多彩，表达形式多种多样，相同的

引语所表现的褒贬倾向在不同的上下文环境下会产生变化。另外，随着时代的变化，词语含义的变化、新词的产生都将使得只针对词语的倾向判断工作的难度加大。但有一些句子特征是可以收集成为特征考虑的。如感叹句和反问句常常带有强烈的情感和观点，连词语句如“连。。。都。。。”对作者的观点取到加强的作用。对于评价词和评价词加强副词的顺序也可以在句子层面加以考虑。

篇章级特征

篇章级情感倾向判断一般应用于主题单一、倾向一致且非常明确的文章中。该任务类似于文本分类问题，对于所有文章，分为褒、贬两大类，通过各种分类算法将文本归入其中某个类别。我们可以利用词语、句子级的倾向判断结果，通过统计的方法进行篇章级倾向分析。对于主题明显的文章，比如针对某一事件的微博评论，可以加入前文中，评论的感情倾向，并虽着情感分析的进行让其更新，让全文拥有一个基本的感情基调。

2.2 参数估计

我们设定，已标注的待训的样本集合为T，待分类样本集合为D。并设定分类器中观点句满足：

$$\sum_{i=0}^{n} w_i x_i > 0$$

非观点句：

$$\sum_{i=0}^{n} w_i x_i < 0$$

计算各特征的权值向量，可以利用实验所提供的已标注数据用感知器法则来进行训练。设定初始值 $W=(0,0,...,0)^T$

对没句文本提取其对应的文档向量 $X=(x_1,x_2,...,x_n)^T$

通过公式

$$w_i = w_i + \Delta w_i \tag{13}$$

$$\Delta w_i = \sum_{x \in D} \eta (t-o) x_i \tag{14}$$

进行训练。η的取值将直接影响训练的速度和训练的收敛效果，如果太大，权向量梯度下降搜索过程就会越过最优解的危险，取值太小则会导致收敛速度过慢的后果，为了综合两者的优点，本算法采取的取值方式是在训练的过程中，逐步减小η的值。

2.3 观点识别强度识别

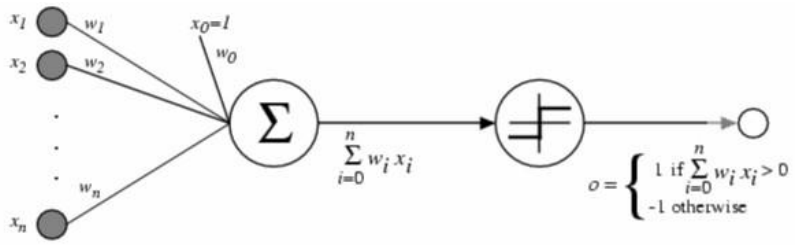

图 3 感知器

如图所示，所述特征向量 $X=(x_0,x_1,x_2,...,x_n)^T$ 所决定的超平面 $\sum_{i=0}^{n} w_i x_i=0$ 中， $w_0 x_0$ 为调整系数，其中 x_0 取值常数 1， w_0 的取值由通过对样本的训练得到。为观点句分类器选定的大都为正面特征， w_0 是小于 0 的。为了达到更好的区分度，加入了松弛因子变量使得超平面函数为：

$$w_0 x_0 *(\log(Ls/Lw)-C)+\sum_{i=1}^{n} w_i x_i=0 \qquad (15)$$

上式 $\log(Ls/Lw)-C$ 为松弛因子， Ls 为当前所分析语句的句子长度， Lw 为当前句中的特征词长度总和， C 为常量因子。通过如此修改算式，可以对含观点词比例较大的语句更高的观点句分来概率。

2.4 情感正负识别

情感倾向强度就是在区分褒贬倾向的基础上，进一步细分，用量化的标准将情感倾向非常明显、比较明显和不太明显的情况表示出来。也就是说，对于某个对象的念度，希望能够通过打分的形式给出其强度信息。对于态度倾向不是简单的判断其褒贬态度，而是给出强度信息，由此作为情感倾向分析结果输出的排序依据。对于文本摘要或过滤系统来说，通过对强度设置阈值，进行更精确更实用的操作。

在情感的正负识别算法中，默认所分析微博评论的感情倾向是中性的，所以没有观点句分类器中的 $v_0 x_0$ 项，情感正负分类器的权向量用 $V=(v_1,v_1,...,v_m)$ 来表示。情感句特征向量 $X=(x_1,x_2,...,x_m)^T$ 直接从观点句的向量值直接中提取的到。计算方式为： $\sum_{i=1}^{m} v_i x_i>0$ 则所分析微博评论为正面情感，否则为负面。

3 结果和结论

《自然语言处理与中文计算会议》是由中国计算机学会（CCF）主办的 CCF 中文信息技术专业委员会年度学术会议。《自然语言处理与中文计算会议》专注于自然语言处理及中文计算领域的学术和应用创新，致力于推动该领域学术界和工业界研究、创新与应用的发展，成为覆盖全国、具有国际影响力的学术与创新交流平台，并与 2012 第一届举办，算法的实验是在其提供的测评数据和标注数据的基础上进行的。

参加《自然语言处理与中文计算会议》微博情感分析评测的结果表明，在观点句的抽取中，准确率达 78%，所参赛 53 队中，排第 8 名的成绩，而召回率有待提升，仅 45.5%。在情感倾向评测的结果中，准确率达 87.9%，在所参赛 53 队中，排名第 2，40%的召回率依然有待提升。

较高的准确率说明了算法的有效性和准确性，对分类的特征选取的比较恰当，算法设计的比较合理。而算法具有较低召回率，是由于特征值的样本容量较小导致的。要改进算法的测试效果，可添加更多的分类特征及收集更多的样本，也可对参数训练的方法加以改进。

参考文献

[1] B. Liu. opinion mining. Proceedings of LREC. Sentiment Analysis and Subjectivity. Handbook of Natural Language Processing, Second Edition, (editors: N. Indurkhya and F. J. Damerau), 2010.

[2] Liu. Sentiment analysis: a multi-faced problem. to appear in IEEE Intelligent Systems,2010;Accessed 01/11/10, 2010.

[3] Andrew L. Maas, Raymond E. Daly, Peter T. Pham. Learning Word Vectors for Sentiment Analysis. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics，2011，pages 142–150.

[4] Alexander Pak, Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. Proceedings of LREC. 2010.