多分类器组合的情感分析方法研究

王超名 陈旭东

（北京交通大学 软件学院 软件工程）

摘要：近年来，随着 O2O 的迅速发展，外卖领域产生了大量的文本评论。在大数据时代，这些包含用户主观情感倾向的文本蕴含着大量有用的信息，人工分析文本的情感倾向效率极低，所以，利用计算机对文本进行情感分析具有巨大的应用价值。本文针对外卖领域的文本评论，对情感极性分析的常用方法进行了系统研究。1）、本文首先分析了基于情感词典的情感极性分析算法。在人工分析了近三百条评论的基础上，构建了一个外卖领域的情感词典。同时，本文提出一个新词典——短语词典，可以自动识别更多句式和短语。2）、接着，本文分析了基于 k-近邻、朴素贝叶斯、最大熵模型、支持向量机的情感极性分析算法，发现基于最大熵模型和支持向量机的算法准确率最高。3）、最后，本文在观察各个分类器分类结果的基础上，发现单个分类器都有各自的缺点，于是本文尝试将多个分类器组合起来以期克服其缺点。本文探究了多分类器组合的两种方式和组合后的两种决策策略，在标准康奈尔影评数据集上，最好的组合策略能将准确率提升 6.5%。另外，在本文构建的外卖数据集上，该方法同样有效。

关键词：多分类器融合；情感分析；K-近邻；贝叶斯；最大熵；支持向量机；情感词典

Sentiment Analysis Research on Multiple Classifiers Combination

Chaoming Wang, Xudong Chen

1 引言

近年来，随着 O2O 的迅速发展，外卖领域产生了大量的文本评论。在大数据时代，这些包含用户主观情感倾向的文本蕴含着大量有用的信息，然而人工分析文本的情感倾向效率极低，所以，利用计算机对文本进行情感分析具有巨大的应用价值。

情感分析在粒度上可以有三种划分：词语粒度、句子粒度和篇章粒度。[1]而每种粒度下，又有其相应的几个子任务。[2]

1）、篇章粒度。篇章粒度上的情感分析主要考虑整个篇章的情感极性。Lee et al. 将传统机器学习方法应用于电影评论的情感分析，最后得到从篇章中提取出 unigrams 后 SVM 的分类效果最好。[3]这三种机器学习方法都是有监督的方法。而 Peter D. Turney 提出了一种无监督的方法进行篇章情感极性的自动分类。[4]篇章粒度有几个子任务，都有相应的研究尝试解决。比如，主客观分类[5]、观点有效性检测[6]、垃圾观点侦测[7]。

2）、句子粒度。句子粒度的情感分析首先是判断句子是否主观或客观，然后将主观语句进行极性分类。Weibe et al. 提出主观性分析在句子中的有效性及其在自然语言处理的应用。[8]熊德兰等人提出了基于 HowNet 的语义距离和语法距离相结合的句子褒贬倾向性计算方法利用夹角余弦法对语义倾向进行了改进。[9]句子粒度的子任务有：观点检索[10]、观点问答[11]、观点摘要[12]。

3）、词语粒度。词语粒度的情感分析主要是提取出评价对象，评价观点，实体，属性，

特征及其情感倾向。[13][14][15]

本文的工作是尝试对外卖领域的文本，进行句子粒度的情感分析，归纳总结基于词典的方法和基于机器学习的有监督的情感极性分析方法。最后提出多种分类器并联+串联的混合融合策略，使用置信平均决策或投票决策，来决定句子的情感极性。结果发现，准确率有极大提升。

本文内容组织如下：第二部分是情感分析的预处理技术，第三部分是基于情感词典的情感极性分析方法，第四部分是基于机器学习（k-近邻算法、朴素贝叶斯、最大熵模型、支持向量机）的情感分析方法，第五部分是本文提出的多分类器融合的情感分析方法。

2 预处理

2.1 构建爬虫

传统爬虫向服务器发送请求后，服务器返回相应的静态网页信息。而如今大多数网站，很多信息都利用 js 进行动态加载，传统爬虫根本无法获取到。所以，本文利用 Python 上 selenium 和 BeautifulSoup，构建了一个终极爬虫，能自动打开浏览器，模拟实际的网页点击事件和浏览事件，与 js 网页动态交互，能成功解析 js 动态网页，爬取外卖网站上的文本评论。

另外，在爬取的文本评论内，有大量的垃圾评论、重复评论、短小的无效评论。本文设计了专门的算法，来保证文本评论语料的有效性、可靠性。首先判断评论是否太短，若小于一定阈值（如 4 个字），则去除；判断是否是无效的如“aaaaaa”之类的英文单词评论；判断是否是“23333333”之类的数字评论；判断是否是句子内部的重复单词评论，如“啊啊啊啊啊”、“好吃好吃好吃好吃”；判断是否是重复句子，有些顾客直接复制粘贴前一位评论者的评论进行发表，所以需要去掉重复句子。其次，根据用户打分，5 分为正向评论，3 分以下为负向评论，最后，我们得到了一个高效可靠的评论语料库。

2.2 特征选择

文本处理中，特征选择是一个非常重要的步骤，其目的在于从原始特征信息中，挑选出最具有代表性的、分类性能优异的特征进行分类，故合适的特征选择方法将很大程度上决定最终分类效果的好坏。通常情况下，文本分类领域都是选取词语作为特征，通过计算词语与类别之间的关系，度量各个词语对类别的贡献度大小，从而归属于某个类别。由于在文本数据处理中，往往词语的数量非常多，若把所有的词语都选为特征项，则特征空间的维度过大，不仅会增加运算量，还会影响分类的精度。因此须对文本内容作降维处理，合适的特征选择方法就尤为重要。情感分类中的特征选择，一方面需要去除与情感无关、类别关联度较小的特征，排除不必要干扰。另一方面，特征选择方法要能获取与情感分类有关联的特征信息，才能提高情感倾向性判别的准确性。[16]因此，必须针对外卖评论选择合适的特征抽取方法，才能提高情感识别的分类效果。本文的特征选择采用X^2统计方法：

$$X^2(t,C)=\frac{N(AD-BC)^2}{(A+C)(B+D)(A+B)(C+D)}$$

X^2统计方法度量特征词语 t 和句子类别 c 之间的相关程度，并假设 t 和 c 之间符合具有一阶自由度的X^2分布。如果 A 表示包含词条 t 且属于类别 c 的句子频数，B 为包含 t 但是不

属于类别 c 的句子频数，c 表示属于类别 c 但是不包含 t 的句子频数，N 表示语料中句子总数，D 表示既不属于 c 也不包含 t 的句子频数。[17]

实际上，如果类别是一个二分类别，则$X^2(t,c1) = X^2(t,c2)$.

total_score	pos_score	neg_score	pos_num	neg_num	word
2616.32556	1308.16278	1308.16278	2762	240	不错
2139.55902	1069.77951	1069.77951	2853	397	很
1905.06747	952.53373	952.53373	2219	241	好
1744.33030	872.16515	872.16515	2450	367	好吃
1624.06672	812.03336	812.03336	1459	69	很好
1276.75156	638.37578	638.37578	1219	75	很快
1237.64500	618.82250	618.82250	1362	129	快
1183.08199	591.54099	591.54099	3234	3808	了
1130.73832	565.36916	565.36916	1186	99	非常

图 1：X^2值前几名的词语。能看出这些词都是一些有效的情感词。“了”这样的词出现在其中，说明可以去除一些停用词，来进一步提高分类精度。

0.09101	0.04551	0.04551	10	6	肥肠
0.08866	0.04433	0.04433	24	15	软
0.08773	0.04386	0.04386	8	6	奶茶
0.08773	0.04386	0.04386	8	6	做生意
0.08773	0.04386	0.04386	8	6	食品
0.08773	0.04386	0.04386	8	6	素
0.08773	0.04386	0.04386	8	6	考虑

图 2：X^2值后几名的词语。能看出这些词的分类作用不是很大。

3 基于词典的情感分析

3.1 情感词典

文本的情感倾向大多通过情感词语来体现，情感词典能否覆盖全面在一定程度上影响着情感分类效果，故情感词典的构建是情感分类研究的基础。文本情感分析研究领域还没有一部完整且通用的情感词典。若构建一个面向外卖领域的情感词典，一方面须对当前的已有相关资源进行总结与整理，另一方面需要构建一个基于外卖的领域情感词典。本文选择中国科学院董振东教授构建的知网（HowNet）为基础情感词典，在此基础上，人工分析了近三百条评论，加上副词词典、连词词典、否定词词典，完成情感词典的构建。另外，本文提出新的短语词典，能自动识别更多句式和短语。

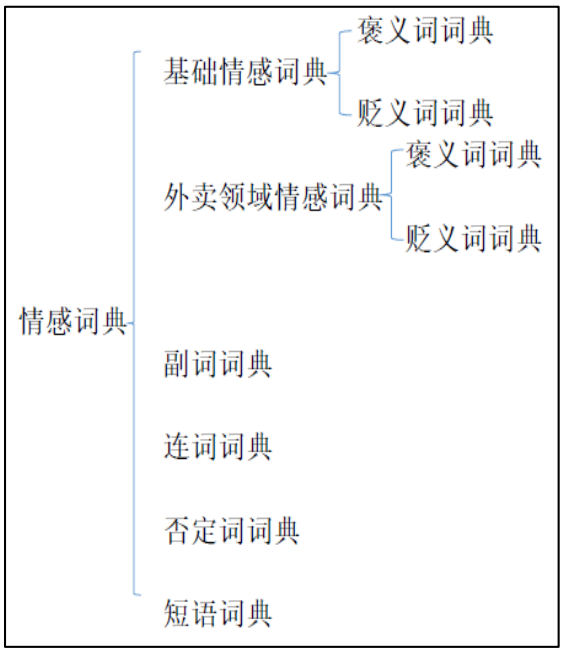


图 6：外卖领域情感词典的构成

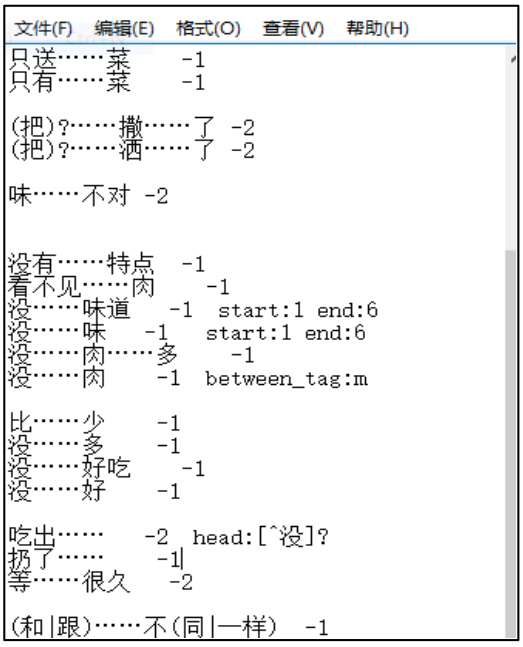


图 7：部分短语词典内容

3.2 情感分析算法

```
1.  Sentiment_analysis_from_corpus_file (path, encoding)
2.      Sentiment_analyse_a_sentence (sentence)
3.          Divide_sentence_into_clauses (the_sentence)
4.          sentiment_analyse_a_clause(the_clause, seg_result)
5.              clause_is_pattern_1(the_clause)
6.              clause_is_pattern_2(the_clause, seg_result)
7.                  emotional_word_analyse(core_word, value, segments, index)
8.              sentiment_analyse_a_word(the_word, seg_result=None, index=-1)
9.                  word_is_conjunction(the_word)
10.                 word_is_punctuation(the_word)
11.                 word_is_positive(the_word, seg_result, index)
12.                     emotional_word_analyse(core_word, value, segments, index)
13.                 word_is_negative(the_word, seg_result, index)
14.                     emotional_word_analyse(core_word, value, segments, index)
15.             clause_is_pattern_3(the_clause)
```

程序的算法分析如下：

1、首先将语句分成几个子句，分句为后续的情感分析做铺垫，若分句错误，后续的情感分析将无从谈起。另外，分句要能识别出各种句式，如“如果……就……”。接下来对每一个分句进行分析，得到每个分句的情感分值，整句的情感分值为每个分句情感分值之和。

2、判断分句是否是句式一，即“如果……就”，“要是……就好了”等假设句式。若是，则不管句式内有多少正向词语，直接判定为负向情感。比如，“要是肥牛再多点就好了”。

3、判断分句是否是句式二，句式二是被写进文件的各种短语。每次读取文件，解析成正则表达式，进行匹配。比如，文件内为:“提高了……质量 2”会被解析成“提高了[\u4e00-\u9fa5]* 质量”，为正向短语，分值为 2；而“希望……提高……质量 -1”会被解析成“希

望[\u4e00-\u9fa5]*提高[\u4e00-\u9fa5]*质量”，为负向短语，分值为-1。另外，还有一些特殊短语的解析，如：“没……味道 -1 start:1 end:6”会解析成“没[\u4e00-\u9fa5]{1， 6}味道”，中间最少 1 个字，最多 6 个字。而“没……肉 -1 between_tag:m”会解析成“没[\u4e00-\u9fa5]*肉”，并且中间必须有一个数词（词性 tag 为 m），匹配“没有几块肉”之类的句子。

4、将分句分词，分词利用 jieba 分词。然后逐个分析分词。首先判断是否是连词，是，则记录；判断是否是标点符号，是，则记录。判断是否是正向情感词，是则深度分析前三个视窗内的词。若有奇数个否定词，则正向情感分值变负；偶数个否定词，情感分值乘 2 加倍。若有副词，判断副词强度，“很”之类的词情感分值加倍，“不太”这类的副词，情感分值减半。

5、然后，乘上之前记录的连词的倍数，标点符号的倍数，综合得到整个分句的情感分值。

6、所有分句情感分值分析完成之后，将所有分句分值统计相加，得到整个句子情感分值。若大于 0，为正向情感，若小于 0，为负向情感，等于 0，为中性。

3.3 实验结果

本文对正负各 1000 条外卖评论进行测试，实验结果如下：

Positive			Negative			total
P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	
79.82	80.7	80.26	80.49	79.6	80.04	**80.15**

可以发现，基于词典的情感分析方法，代码量大，人工构建量大，如果要构建一个高精度的基于情感词典的情感分析方法，人工工作量相当大。所以我们对基于机器学习的情感分析方法展开了研究。

4 基于机器学习的情感分析

4.1 k-近邻算法

k-近邻算法简单直观。给定一个训练数据集，对新的输入实例，在训练数据集中找到与该实例最近邻的 k 个案例，这 k 个案例的多数属于某个类，就把该输入实例分为这个类。[18]

基于 K-近邻算法的情感分析的一般步骤为：

1、 计算已知类别的数据集中的每个句子与当前句子之间的距离；
2、 按照距离递增次序进行排序；
3、 选取与当前句子距离最小的 k 个句子；
4、 确定这 k 个句子所在类别的出现频率；
5、 返回前 k 个句子出现频率最高的类别作为当前点的预测分类。

4.2 朴素贝叶斯

假设句子 X 包含词汇（$a_1, a_2, \cdots\cdots, a_m$）。Y 为情感极性类别的集合（$y_1, y_2$）。则 X 的情

感极性为$argmax\big(P(yi|x)\big) = argmax(P(yi|x1, x2, \dots, xm))$.由贝叶斯公式

$$P(y_i|x) = \frac{P(x|y_i)P(y_i)}{P(x)}$$

$P(x)$是相同的，可直接忽略。所以，只需要找到使得$P(x|y_i)P(y_i)$最大的y_i。

朴素贝叶斯（Naïve Bayes）被称为 naïve 是因为假设各个词之间是相互独立的，于是：

$$P(x|y_i)P(y_i) = P(a_1|y_i)P(a_2|y_i) \dots P(a_m|y_i)P(y_i) = P(y_i)\prod_{j=1}^{m} P(a_j|y_i)$$

Naïve Bayes 的条件独立假设看上去很傻很天真，但结果却很强大。因为有些独立假设在各个分类之间的分布都是均匀的，所以对于似然的相对大小不产生影响；即便不是如此，也有很大的可能性各个独立假设所产生的消极影响或积极影响互相抵消，最终导致结果受到的影响不大。[19]

4.3 最大熵

最大熵原理指出，当我们需要对一个随机事件的概率分布进行预测时，我们的预测应当满足全部已知的条件，而对未知的情况不要做任何主观假设。在这种情况下，概率分布最均匀，预测的风险最小。因为这时概率分布的信息熵最大，所以人们称这种模型叫"最大熵模型"。以下分别介绍了从数学上如何 1）、满足全部已知条件，2）、对未知条件尽可能均匀。

4.3.1 满足所有已知条件

我们假设 p 为最大熵模型的学习后的所有条件概率，$p(y|x)$为 p 中的一个元素。训练集内包含大量的样本，$(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)$。我们定义$\tilde{p}(x, y)$为样本$(x, y)$出现的概率。从训练集中我们可以得到很多有用的特征，我们定义一个特征函数：

$$f(x, y) = \begin{cases} 1 & \text{满足特征条件} \\ 0 & \text{不满足特征条件} \end{cases}$$

特征函数在满足成为一个特征时取值 1，否则取值 0。则特征函数的经验概率为：

$$\tilde{p}(f) \equiv \sum_{x,y} \tilde{p}\,(x, y) f(x, y)$$

在最大熵模型中，特征函数的期望概率为：

$$p(f) \equiv \sum_{x,y} \tilde{p}\,(x) p(y|x) f(x, y)$$

最大熵模型的训练，就是满足这样一个约束条件——模型的期望概率与训练集中得到的经验概率相等：

$$p(f) = \tilde{p}(f)$$

所有这样的约束组成一个集合 C。

4.3.2 对未知条件尽可能均匀

现在，最大熵模型已经满足了所有约束条件了。那对于未知条件，我们要尽可能均匀（uniform）。那怎样才能达到“尽可能均匀”呢？一个衡量均匀分布的数学度量为熵：

$$H(p) \equiv -\sum_{x,y} \tilde{p}\,(x)p(y|x)\log(p(y|x))$$

当熵最大时，则对位置条件就尽可能均匀了。

最大熵模型的训练，就是要找到一个模型$p^* \in C$，并且具有最大熵$H(p)$。即，

$$p^* = \underset{p\in C}{\operatorname{argmax}}\, H(p) = \underset{p\in C}{\operatorname{argmax}}(-\sum_{x,y} \tilde{p}\,(x)p(y|x)\log(p(y|x)))$$

4.4 支持向量机

支持向量机的基本思想是：求解能够正确划分训练数据集并且几何间隔最大的分离超平面。对于线性可分的训练数据集，线性可分离超平面有无穷多个（等价于感知机），但是几何间隔最大的分离超平面却只有一个。间隔最大的直观解释是：对训练集找到几何间隔最大的超平面意味着以充分大的确信度对训练数据集进行分类。[19]以下分别介绍了从数学上如何1）、划分数据集，2）、最大化间隔。

4.4.1 划分数据集

一个线性分类器的学习目标就是要在 n 维的数据空间中找到一个分类超平面，其方程可以表示为：

$$W^T x + b = 0$$

于是，我们得到一个分类函数：

$$f(x) = W^T x + b$$

为了便于计算距离，我们将分类标签设为 1 和-1。

4.4.2 最大化间隔

这个分类超平面要满足这样的条件：找到离分割超平面最近的点，确保它们离分割超平面的距离尽可能远。[20]

令 d 为离分割超平面最近的点 X 到分类超平面的距离：

$$d = \frac{W^T x + b}{\|W\|} = \frac{f(x)}{\|W\|}$$

d 小于 0 表示点在超平面负侧，标签$y = -1$；d 大于 0 表示点在超平面正侧，标签$y = 1$。而我们需要不是 d，而是 d 的绝对值，因此乘上对应的类别 y 得：

$$\tilde{d} = \mathrm{y} * \mathrm{d} = \frac{f(x)}{\|W\|}$$

我们的目标便是：$\max \tilde{d}$。方便推导和优化的目的，我们可以令最近的那些点的$f(x) = 1$。于是上述目标函数$\max \tilde{d}$转化为：

$$\max \frac{1}{\|W\|}, s.t.\, y_i * (W^T + b) \geq 1, i = 1, \dots, n$$

这个问题等价于：

$$\min \frac{1}{2}\|W\|^2, s.t.\, y_i * (W^T + b) \geq 1, i = 1, \dots, n$$

4.5 实验结果

对于外卖领域的文本评论，本文采用正负各 3500 条作为训练数据，正负各 1000 条作为测试数据，特征数为 4000。其中基于支持向量机的情感分析方法利用了 Python 机器学习包 scikit-learn 中的 SVC。最终实验数据统计如下：

Method	Positive			Negative			total
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	
K-NN（k=1）	**99.84**	60.8	75.57	71.82	**99.9**	83.56	80.35
K-NN（k=3）	93.02	58.6	71.9	69.78	95.6	80.68	77.1
Bayes	91.11	69.7	78.98	75.47	93.2	83.4	81.45
MaxEnt	91.91	80.7	85.94	82.80	92.9	87.56	**86.8**
SVM	91.92	**81.9**	**86.62**	**83.68**	92.8	**88.0**	**87.35**

对于标准的康奈尔影评数据集，本文采用正负各前 500 篇作为训练数据，正负各 200 篇作为测试数据，特征数为 4000。其中基于支持向量机的情感分析方法利用了 Python 机器学习包 scikit-learn 中的 SVC。实验数据统计如下：

Method	Positive			Negative			total
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	
k-NN（k=1）	**100**	65.5	79.15	74.35	**100**	85.29	82.75
k-NN（k=3）	77.17	71	73.96	73.15	79	75.96	75
Bayes	99.24	65	78.55	73.98	99.5	84.86	82.25
MaxEnt	90.5	**82**	**86.09**	**83.56**	91.5	87.35	**86.75**
SVM	99.34	75.5	85.80	80.24	99.5	**88.84**	**87.5**

5 多分类器融合的情感分析

5.1 实验观察

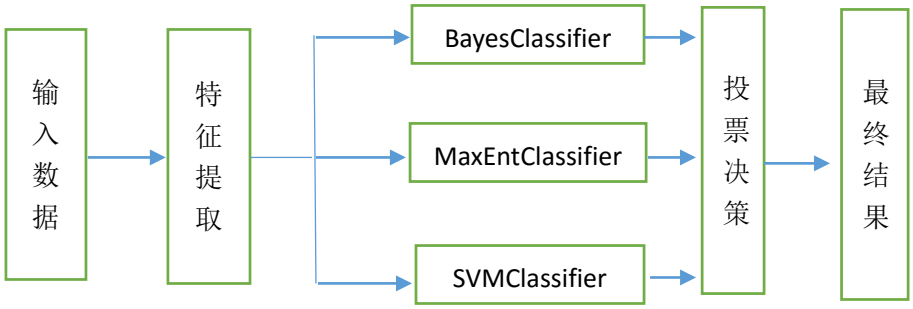
从宏观整体上，可以看出，最大熵、SVM 的准确率最高。而在微观细节上，为了更直观、详细地观察到每一个分类器对每一个测试数据的分类结果，本文做了如下统计工作：

	数据2	数据11	数据202	数据488	数据506	数据510	数据1019	数据1022	数据1029	数据1099	数据1113	数据1203
Origin	1	1	1	1	1	1	0	0	0	0	0	0
K-NN	1	0	1	0	1	0	0	0	0	0	0	0
Bayes	0	0	1	1	0	1	0	0	1	1	0	0
MaxEnt	1	1	1	0	0	1	1	1	0	1	1	0
SVM	1	1	0	1	0	1	1	0	0	1	0	**1**

本文选取了部分外卖数据的分类结果用于展示。“Origin”一行表示测试数据原始标签，“k-NN”一行表示基于 K-近邻算法的情感分析结果，“Bayes”、“MaxEnt”、“SVM”三行如是。可以发现，对于第 2、1029 条数据，Bayes 分类错误，但另外三个分类器分类正确；对于第 202、1230 条数据，SVM 分类错误，另外三个分类器分类正确；对于第 1022 条数据，最大熵分类错误，但是其他三个分类器又分类正确。很自然地，本文想，是否可以将这几种不同的分类器并联在一起，进行投票决策？哪个分类票数多，则结果就属于那个分类。于是，本文提出多分类器并联组合的投票决策策略。

5.2 并联组合：投票决策

但投票决策可能会出现一些问题，如第 11、488 条数据等，正负分类结果各有两个分类器，投票将得不出结果。因此，应该选用奇数个分类器用于投票决策。最大熵和 SVM 性能最好，入围理所当然。在 Bayes 和 K-NN 中，K-NN 内存占用大，计算复杂度高，空间复杂度也高，所以，本文淘汰 k-NN，选用 Bayes+最大熵+SVM 进行并联组合。其并联组合流程为：



若 C_i 为分类为 i 的分类器个数，则多分类器并联组合的投票决策策略下的情感分类为：

$$i = argmax(\mathrm{Ci}).$$

对于外卖领域的文本评论，本文采用正负各 3500 条作为训练数据，正负各 1000 条作为测试数据，特征数为 4000。最后实验数据如下：

Method	Positive			Negative			total
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	
Fusion1	**92.54**	80.6	86.16	82.82	**93.5**	87.83	87.05
Bayes	91.11	69.7	78.98	75.47	93.2	83.4	81.45
MaxEnt	91.91	80.7	85.94	82.80	92.9	87.56	86.8
SVM	91.92	**81.9**	**86.62**	**83.68**	92.8	**88.0**	**87.35**

对于标准的康奈尔影评数据集，本文采用正负各前 500 篇作为训练数据，正负各 200 篇作为测试数据，特征数为 4000。最后实验结果如下：

Method	Positive			Negative			total
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	
Fusion1	**99.39**	81	**89.26**	**83.97**	**99.5**	**91.08**	**90.25**
Bayes	99.24	65	78.55	73.98	**99.5**	84.86	82.25
MaxEnt	90.5	**82**	86.09	83.56	91.5	87.35	86.75
SVM	99.34	75.5	85.80	80.24	**99.5**	88.84	87.5

多分类器并联组合的投票决策在外卖数据集上无效，而在康奈尔影评数据集上性能却提升显著。要寻找原因，只能重新回到细节微观上看每个分类器对每个数据的分类效果及投票

决策效果。以下是外卖数据集上摘取的部分分类结果：

	数据 2	数据 9	数据 32	数据 39	数据 53	数据 417	数据 583	数据 598
Origin	1	1	1	1	1	1	1	1
Bayes	0	0	1	0	0	0	1	1
MaxEnt	1	0	1	1	0	1	0	1
SVM	1	0	0	0	1	0	0	0
Fusion1	**1**	**0**	**1**	**0**	**0**	**0**	**0**	**1**

可以看出，对于第 2、32、598 条数据而言，“3 局 2 胜”的投票决策的结果是正确的，而对于第 39、53、417、583 条数据，“3 局 2 胜”的投票决策的结果却是错误的。于是，本文做了一些反思。

首先，现实生活中的确有很多投票决策的实例，但是，投票决策是建立在投的每一票都是等价、公平的基础之上的。若每个分类器对于正负分类的投票质量都是一致，则少数服从多数理所当然，但是实际情况是，每个分类器判断一个句子为“正”或“负”时，准确率是不同的，甚至是有显著差异。如，由上述外卖语料的实验结果可得，SVM 投“正”票时，其准确率为 91.92%，而投“负”票时，准确率却只有 83.68%。于是，本文尝试提出多分类器并联组合的置信平均策略。

5.3 并联组合：置信平均

本文尝试用引入置信度 C 的概念。置信度 C_{i-j} 表示分类器 j 对于输入数据分类为 i 的置信度。对于外卖语料而言，三种分类器中，最大熵判断一条语句为正向时，其正确的可能性为 91.91%，所以其正向置信度 C 为 0.9191；判断一条语句为负向时，其正确为可能性为 82.80%，所以负向置信度 C 为 0.828。SVM 判断一条语句为正向时，其正确的可能性为 91.92%，所以其正向置信度 C 为 0.9154；判断为负向时，其正确的可能性为 83.68%，所以其负向置信度 C 为 0.8368。贝叶斯分类器也如是。另外，令 N_i 表示分类为 i 的分类器数目。其中 i 取 0 和 1。则，多分类器并联融合下的情感分类为：

$$i=\begin{cases} argmax\left(\dfrac{\sum_j C_{i-j}}{N_i}\right) & \text{若 } N_i \neq 0 \\ 0 & \text{若 } N_1 = 0 \\ 1 & \text{若 } N_0 = 0 \end{cases}$$

对于外卖领域的文本评论的实验数据如下：

Method	Positive			Negative			total
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	
Fusion2	89.03	**87.7**	**88.36**	**87.88**	89.2	**88.54**	**88.45**
Fusion1	**92.54**	80.6	86.16	82.82	**93.5**	87.83	87.05
Bayes	91.11	69.7	78.98	75.47	93.2	83.4	81.45
MaxEnt	91.91	80.7	85.94	82.80	92.9	87.56	86.8
SVM	91.92	81.9	86.62	83.68	92.8	88.0	**87.35**

对于标准的康奈尔影评数据集的实验结果如下：

Method	Positive			Negative			total
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	
Fusion2	91.22	**93.5**	**92.35**	**93.33**	91	**92.15**	**92.25**

Fusion1	**99.39**	81	89.26	83.97	**99.5**	91.08	90.25

实验结果表明，此方法有效。但为了进一步弄清其工作原理，我们需要再次在微观细节上观察每个分类器对每个数据的分类效果及投票决策效果。以下是外卖数据集上摘取的部分分类结果：

	数据2	数据32	数据39	数据53	数据417	数据583	数据1029	数据1037	数据1054	数据1456
Origin	1	1	1	1	1	1	0	0	0	0
Bayes	0	1	0	0	0	1	1	1	1	1
MaxEnt	1	1	1	0	1	0	0	1	0	0
SVM	1	0	0	1	0	0	0	0	0	0
Fusion1	**1**	**1**	**0**	**0**	**0**	**0**	**0**	**1**	**0**	**0**
Fusion2	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**

可以看出，并联组合投票决策下判断正确的第 2、32 条数据，在并联组合置信平均策略下也判断正确；而对于并联组合投票决策下判断错误的第 39、53、417、583 条数据，即便 3 票内有 2 票投票错误，并联组合置信平均策略也可以判断出其正确类别。但是，对于数据 1029、1037、1054、1456，原本 “3 局 2 胜”投票策略分类正确，然而在并联组合置信平均策略下，却判断错误。进一步反思，可以发现，其实每个分类器的正向置信度都高于负向置信度，所以，3 个分类器中只要有 1 个分类器投票为“正”，则整个语句分类结果必为正向。也就是说，置信平均会抛弃投票决策的优势，而不是改进投票决策。那么，怎样才能兼得置信平均与投票决策各自的优势呢？于是，本文尝试引入多分类器并联组合下的置信平均+投票策略。

5.4 并联组合：置信平均+投票决策

原先 3 个分类器组合的构造决定了置信平均与投票决策的不兼容。若想将投票决策与置信平均优势互补，唯一的方法就是改变多分类器组合的构造。要改造分类器的组合方式，其中一个方法就是增加一个分类器。本文增加的分类器为 k-NN。增加的分类器若也是并联组合，则多分类器并联组合下的置信平均+投票策略为：

$$i=\begin{cases} argmax\left(\dfrac{\sum_j C_{i-j}}{N_i}\right) & \text{若 } N_i=2 \\ 0 & \text{若 } N_0=3 \text{ 或 } N_0=4 \\ 1 & \text{若 } N_1=3 \text{ 或 } N_1=4 \end{cases}$$

对于外卖领域的文本评论的实验数据如下：

Method	Positive			Negative			total
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	
Fusion3	**92.77**	82.1	87.11	83.95	93.6	88.5	87.85
Fusion2	89.03	**87.7**	**88.36**	**87.88**	89.2	**88.54**	**88.45**
Fusion1	92.54	80.6	86.16	82.82	**93.5**	87.83	87.05

对于标准的康奈尔影评数据集的实验结果如下：

Method	Positive			Negative			total
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	
Fusion3	**99.42**	85.5	91.94	87.28	**99.5**	**92.99**	**92.5**
Fusion2	91.22	**93.5**	**92.35**	**93.33**	91	92.15	92.25

Fusion1	99.39	81	89.26	83.97	**99.5**	91.08	90.25

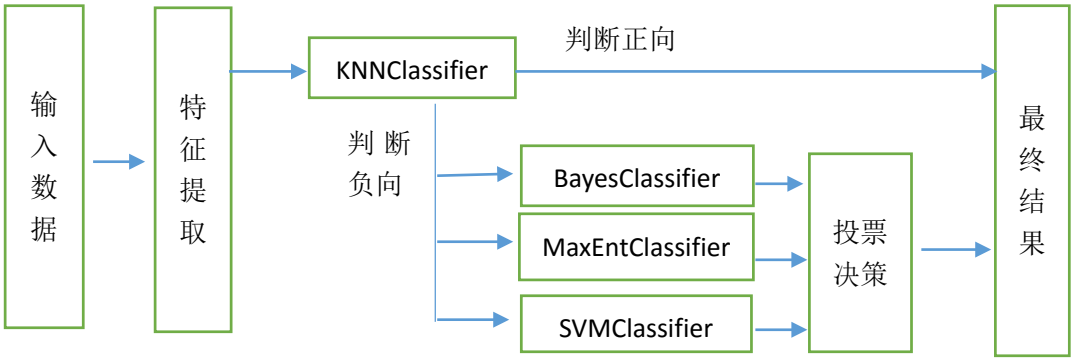
实验结果表明，此方法在不同数据集上并不总是有效。进一步反思原因，可以发现，若将新增分类器 k-NN 与原有分类器并联组合，其实并不是在保证原有分类精度之上做进一步的改善。相反，新增分类器并联组合后反而会增加组合分类器的不确定性。那么，是否可以在保证原有组合分类器精度之上，再添加一个新的分类器，将在置信平均策略下无法正确分类的数据利用投票决策正确分类呢？于是，本文尝试引入多分类器串并联组合。

5.5 串并联组合：投票决策

由 5.3 节的分析可得，置信平均策略无法正确分类的数据大部分都是负向评论，因为每个分类器的正向置信度都高于负向置信度，一个正向评论只要 3 个分类器中有 1 个分类器投票为“正”，则整个语句分类就可投票正确，分类为正向。则，我们的关注点落在负向分类。若我们能得到一个高负向准确率的分类器，则可以先用此分类器进行投票，若此分类器投票为负向，则直接判断该语句为负向；若此分类器投票为正向，则利用 8.3 节多分类器并联组合下的置信平均策略进行分类。也就是，将高负向准确率的分类器与并联组合置信平均分类器进行串联组合，即串并联组合下执行置信平均策略。

然而，本文所有分类器都不具备高负向准确率，却有一个 k-NN 分类器具备接近 100%的高正向准确率。但是一个高正向准确率的分类器与置信平均策略的作用其实是相同的，而不是互补。置信平均策略的本质，是在“2 票负向 1 票正向”的投票情况下，将错误的负向投票决策结果，利用置信度，更正为正向投票结果；然而，若数据原本分类的确是负向，则置信投票决策就错误校正了正确的“3 局 2 胜”投票策略的分类结果。

因此，为了充分利用高正向准确率分类器，本文尝试将正向判断良好工作的高正向准确率的分类器与负向判断能良好工作的并联组合投票决策进行串联，优势互补。于是，此种多分类器混合组合策略流程如下：



对于外卖领域的文本评论，最后实验数据如下：

Method	Positive			Negative			total
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	
Fusion4	92.65	84.5	**88.39**	85.75	93.3	**89.37**	**88.9**
Fusion3	**92.77**	82.1	87.11	83.95	**93.6**	88.5	87.85
Fusion2	89.03	**87.7**	88.36	**87.88**	89.2	88.54	88.45
Fusion1	92.54	80.6	86.16	82.82	93.5	87.83	87.05

对于标准的康奈尔影评数据集，最后实验结果如下：

Method	Positive			Negative			total
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	

Fusion4	**99.44**	88.5	**93.65**	89.64	**99.5**	**94.31**	**94**
Fusion3	99.42	85.5	91.94	87.28	**99.5**	92.99	92.5
Fusion2	91.22	**93.5**	92.35	**93.33**	91	92.15	92.25
Fusion1	99.39	81	89.26	83.97	**99.5**	91.08	90.25

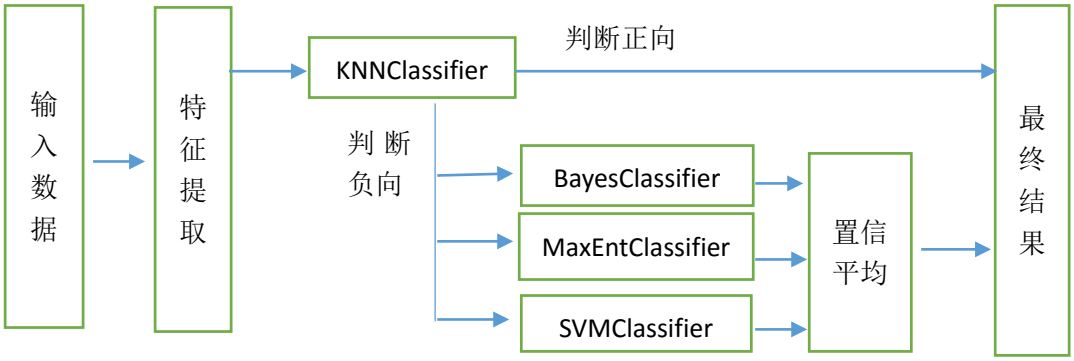
实验结果表明，此方法有效，性能提升显著。但为了弄清此方法是否是在保证并联组合的分类效果的基础上提升了分类精度，我们还需要再次在微观细节上观察每个分类器对每个数据的分类效果及投票决策效果。以下是 5.3 节摘取的同样的外卖数据的分类结果，并添加上 k-NN 的分类结果，汇总如下：

	数据2	数据32	数据39	数据53	数据417	数据583	数据1029	数据1037	数据1054	数据1456
Origin	1	1	1	1	1	1	0	0	0	0
k-NN	1	1	1	0	0	1	0	0	0	0
Bayes	0	1	0	0	0	1	1	1	1	1
MaxEnt	1	1	1	0	1	0	0	1	0	0
SVM	1	0	0	1	0	0	0	0	0	0
Fusion2	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
Fusion4	**1**	**1**	**1**	**0**	**0**	**1**	**0**	**1**	**0**	**0**

可以看出，并联组合置信平均策略下错误分类的第 1029、1054、1456 条数据，在串并联投票决策下被分类正确；在并联组合置信平均策略下正确分类的第 2、32、39 条数据，在串并联投票决策下也能被正确分类。然而，第 53 条数据却被错误分类，这也说明此种组合策略的局限性，丧失了 5.3 节置信平均策略的优势。

5.6 串并联组合：置信平均

k-NN 分类器虽然具有高正向准确率，但是其正向召回率却很低，只有 60.8%。也就是说，虽然 k-NN 判断为正向的结果都是正确的，但是接近 40%本应该判断为正向的语句，却被判断为负向。由 5.5 节的分析可得，多分类器并联组合的置信平均策略也对正向评论具有良好的判断能力。因此，本文尝试将 k-NN 判断为负向的语句利用置信平均策略进行再次判断，以期弥补 k-NN 低正向召回率的缺陷。即串并联的置信平均策略，此种多分类器混合组合策略流程如下：



对于外卖领域的文本评论，最后实验数据如下：

Method	Positive			Negative			total
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	
Fusion5	89.17	**89.7**	**89.43**	**89.64**	86.64	**89.37**	**89.4**

Fusion4	92.65	84.5	88.39	85.75	93.3	**89.37**	88.9
Fusion3	**92.77**	82.1	87.11	83.95	**93.6**	88.5	87.85
Fusion2	89.03	87.7	88.36	87.88	89.2	88.54	88.45
Fusion1	92.54	80.6	86.16	82.82	93.5	87.83	87.05

对于标准的康奈尔影评数据集，最后实验结果如下：

Method	Positive			Negative			total
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	
Fusion5	91.47	**96.5**	**93.92**	**96.30**	91	93.57	93.75
Fusion4	**99.44**	88.5	93.65	89.64	**99.5**	**94.31**	**94**
Fusion3	99.42	85.5	91.94	87.28	**99.5**	92.99	92.5
Fusion2	91.22	93.5	92.35	93.33	91	92.15	92.25
Fusion1	99.39	81	89.26	83.97	**99.5**	91.08	90.25

实验结果表明，此种组合策略也是有效。只是随着实验数据的不同，其性能的提升效果不尽相同。

5.7 结论

本文从外卖数据分类结果的观察出发，得出不同分类器对同一数据有不同的分类效果，单个分类器都有各自的缺点，于是尝试将多个单分类器组合起来以期克服其缺点。最后发现，组合方式的不同、同一组合方式下数据的不同、同一组合方式下决策方式的不同，都对分类效果有影响。最佳的多分类器组合方式应该是将投票决策与置信平均决策两者的优势互补，找到或者构建一个高负向准确率的分类器，与并联组合方式进行串联。在决策时，首先进行负向投票决策，然后再进行正向的置信平均决策。本文最终提出的串并联投票决策和串并联置信平均策略，在康奈尔标准影评数据集上，都有良好表现，准确率比最好的单个分类器能提升 6%以上。

6 参考文献

[1] Pang, Bo, and Lillian Lee. 2008. “Opinion Mining and Sentiment Analysis.” Foundations and Trends® in Information Retrieval 2: 1–135.

[2] K. Vivekanandan, J. Soonu Aravindan. 2014. Aspect-based Opinion Mining: A Survey. International Journal of Computer Applications (0975 – 8887) Volume 106 – No.3, November 2014

[3] Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. 2002. “Thumbs up? Sentiment Classification Using Machine Learning Techniques.” In Conference on Empirical Methods in Natural Language Processing (EMNLP 2002), 79–86.

[4] Turney, Peter D. 2001. "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews.” In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL ‘02 417- 424.

[5] Yu, Lei, Jia Ma, Seiji Tsuchiya, and Fuji Ren. 2008. “Opinion Mining: A Study on Semantic Orientation Analysis for Online Document.” In Proceedings of the World Congress on Intelligent Control and Automation (WCICA), 4548–52.

[6] Liu, Jingjing et al. 2007. “Low-Quality Product Review Detection in Opinion Summarization.” In Computational Linguistics, 334–42.

[7] Lim, Ee-Peng et al. 2010. “Detecting Product Review Spammers Using Rating Behaviors.” Proceedings of the 19th ACM International Conference on Information and Knowledge Management: 939–48.

[8] Wiebe, Janyce et al. 2004. “Learning Subjective Language.” Computational Linguistics 30: 277–308.

[9] 熊德兰，程菊明，田胜利．基于 HowNet 的句子褒贬倾向性研究．计算机工程与应用，2008，44(22)：143—144．

[10] Huang, Shen et al. 2009. “Improving Product Review Search Experiences on General Search Engines.” In Proceedings of the 11th International Conference on Electronic Commerce - ICEC ’09, 107.

[11] Moghaddam, Samaneh, and Martin Ester. 2011. “AQA: Aspect-Based Opinion Question Answering.” In Proceedings - IEEE International Conference on Data Mining, ICDM, 89–96.

[12] Hu, Minqing, and Bing Liu. 2004. “Mining and Summarizing Customer Reviews.” Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining KDD 04 04: 168.

[13] Lu, Yue, ChengXiang Zhai, and Neel Sundaresan. 2009. “Rated Aspect Summarization of Short Comments.” Proceedings of the 18th international conference on World wide web - WWW ’09: 131.

[14] Wong, Tak-Lam, Wai Lam, and Tik-Shun Wong. 2008. “An Unsupervised Framework for Extracting and Normalizing Product Attributes from Multiple Web Sites.” Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval SIGIR 08: 35.

[15] Liu, Bing, and South Morgan Street. 2005. “Opinion Observer : Analyzing and Comparing Opinions on the Web.” In Proceedings of the 14th International Conference on World Wide Web, 342–51.

[16] 硕士论文.基于情感词典的中文微博情感倾向分析研究.陈晓东. 2012 年 1 月 12 日

[17] MADEIRA H．COSTA J，VIEIRA M．The 0LAP and data warehousing approaches for analysis and sharing of results from dependability evaluation experiments[C] International Conference on Dependable Systems and Networks.[S.1.]：IEEE Press，2003：86—91．

[18] 统计学习方法，李航，清华大学出版社,2012.3

[19] Harry Zhang. The Optimality of Naive Bayes. 2004, American Association for Artificial Intelligence

[20] Peter Harrington. Machine Learning In Action.2013. Manning Publications