基于优化样本分布抽样集成学习的半监督文本分类方法研究*

徐禹洪[1,2]，黄沛杰[1]

（1. 华南农业大学数学与信息学院，广东 广州 510642；

2. 华南理工大学计算机科学与工程学院，广东 广州 510006）

摘要：针对现有文本分类方法在即时性文本信息上的挑战，考虑到即时性文本信息具有已标注数据规模小的特点，为了提高半监督学习的分类性能，本文提出一种基于优化样本分布抽样集成学习的半监督文本分类方法。首先，通过运用一种新的样本抽样的优化策略，获取多个新的子分类器训练集，以增加训练集之间的多样性和减少噪声的扩散范围，从而提高分类器的总体泛化能力；然后，采用基于置信度相乘的投票机制对预测结果进行集成，对未标注数据进行标注；最后，选取出适量的数据来更新训练模型。实验结果表明，该方法在长文本和短文本上都取得了优于研究进展方法的分类性能。

关键词：文本分类；半监督学习；集成学习；样本抽样策略

中图分类号：TP391　　**文献标识码：**A

Semi-supervised Classification Method Based on Ensemble Learning with Optimizing Sample Distribution Sampling

XU Yuhong[1,2], HUANG Peijie[1]

(1. College of Mathematic and Informatics, South China Agricultural University, Guangzhou 510642, China;

2. School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China)

Abstract: This paper address the challenge of existing text classification methods on instant text information. Considering the instant text information has marked characteristics of small scale labeled data, in order to improve the classification performance of semi-supervised learning, this paper proposes a semi-supervised classification method based on ensemble learning with optimizing sample distribution sampling. First, through a new optimal sampling strategy, several new sub classifier training sets are obtained, in order to increase the diversity between training set and reduce the diffusion range of noise, so as to improve the generalization ability. Then, the voting mechanism based on confidence multiplication is used to integrate the prediction results, and the unlabeled data are labeled. Finally, the appropriate amount of data is selected to update the training model. The experimental results show that our approach has better classification performance in long text and short text.

Key words: text classification; semi-supervised learning; ensemble learning; sampling strategy

1 引言

随着现代信息技术的不断发展，信息传播的途径变得方便和快捷，使更多的人们通过互联网进行信息的接收和传播[1]，产生了越来越多类似新闻标题、微博评论和产品评价等即时性文本信息。而这些即时性文本信息，实际上很难在短时间内得到大量的已标注数据，因此对类似新闻的即时性文本信息的自动分类就成为了当前的研究热点[2]。同时，相对于长文本来说，含有文本特征更少的短文本将带来更大的挑战[3]。

众所周知，监督学习的文本分类算法依赖于已标注数据作为模型的训练集，其分类性能也会随着训练集规模增大而逐步提升，然而上述的即时性文本分类任务面临的重大问题是极

*收稿日期：　　定稿日期：

基金项目：国家自然科学基金(71472068)；广东省大学生创新训练计划项目(201510564281)

少的已标注数据和大量的未标注数据[4]，因此监督学习无法被应用到此类小规模已标注数据的分类任务上。显然，如果只利用少量的已标注数据来对分类模型进行训练，一般很难得到具有较强泛化能力的分类模型，对海量的未标注数据进行人工标注必定是一项巨大的额外开销。如果弃用大量“低廉的”未标注数据，又极大地浪费了这些数据的存在价值。

由于上述原因，近年来，半监督学习成为了研究热点。半监督学习通过对未标注数据进行预测和赋予标注，从而增加了已标注数据的规模，使分类器的分类性能得到提升，即在少量已标注数据的基础上，利用大量且易获取的未标注数据来提高分类模型的泛化能力[5]。然而，运用半监督学习方法也产生了另一个重要的问题，如何去确保通过少量已标注数据训练的分类模型预测出来的标注的正确性。Co-training是目前分类任务中最常见的半监督学习方法，该方法要求满足两个条件：（1）分类器之间存在一定的独立性；（2）每个分类器都是一个强分类器。由于文本没有天然存在的多个独立视图，经典的做法是划分特征子空间方法，但是通过划分特征子空间后，部分分类器因为不包含足够的有用特征信息而退化成弱分类器，最终导致总体分类性能的下降。为了提高分类器的性能，使分类模型预测出来的标注更为正确，集成学习成为了一种有效的途径[6]。对若干个子分类器进行集成，不仅提升了分类器性能，也使分类器更具普适性，其中，子分类器之间的多样性对于集成学习的成效是至关重要的[7]。

本文提出了一种基于优化样本分布抽样集成学习的半监督分类方法（Semi-supervised classification method based on ensemble learning with optimizing sample distribution sampling，以下简称SSEOS）。为了增加子分类器之间的多样性和提高子分类器的分类性能，SSEOS采用了优化样本分布的抽样策略。一方面，尝试从样本分布的角度上增加样本之间的多样性，对训练集按照不同分布的抽样，不仅增加了子训练集之间的多样性，还降低了噪声数据的扩散范围；另一方面，对于每一个通过不同样本分布抽样而来的子训练集，采用基于Bagging的优化分布重抽样方法构建多个子分类器训练集。其中，由同一个样本分布的多个子分类器训练集学习而来的一系列子分类器，通过运用基于置信度相乘的集成方法组合为代表该样本分布的分类器，然后多个代表不同样本分布的分类器再通过运用基于置信度相乘的集成方法组合为最终分类器。通过运用半监督学习方法，每次迭代都对所有未标注数据进行预测，同时标注适量置信度高的未标注数据并将其添加到训练集里，从而更新训练模型。实验结果表明，不管对于长文本还是短文本，SSEOS不仅优于经典的监督学习分类方案，而且比半监督学习分类方案的效果要好。

本文后续部分安排如下：第 2 节详细介绍文本分类的相关工作；第 3 节提出基于优化样本分布抽样集成学习的半监督分类方法；第 4 节给出实验结果及分析；最后，第 5 节给出总结，并对下一步工作进行展望。

2 相关工作

文本分类在组织文本中起着重要的作用，是自然语言处理领域（NLP）中的一个研究热点问题，被有效地应用在万维网、网络新闻和电子邮件上。许多机器学习算法已应用到这个问题上，并已在许多文本分类领域上取得成功效果[8-9]。

一般情况而言，文本分类任务对于不同的文本分析粒度可以归纳成 3 类：篇章级、句子级和词语级[10]。而在本文的文本分类中，主要是针对篇章级的长文本分类和句子级的短文本分类。

基于无监督的分类方法，即不依赖已标注数据来生成分类模型。常用的方法是借助领域词典来进行无监督分类，而领域词典的构建需要花费巨大的开销，且领域词典的适用领域范围小。Turney 采用基于关键词的方法，使用词与词之间的点互信息和词的语义倾向对评论语料进行无监督学习的分类[11]。朱嫣岚等人利用语义相似度对词语的褒贬类别进行分类，

实验结果表明该方法在常用词中的性能较好[12]。刘鸿宇等人通过分析出不同类型情感句中的结构，制定出各类别所对应的情感倾向判断规则[13]。陈涛等人通过句法和语义特征得到句子模式，把句子模式分成 3 个一级类别和 105 个二级类别，然后根据不同句子模式进行相应的处理操作[14]。虽然，无监督学习方法具有简单和方便的优点，且不需要已标注的数据，但缺点是分类效果不佳，难以很好地适应所有领域的文本分类。

与无监督文本分类方法相比，监督学习文本分类方法能够得到更优的分类性能，但该方法非常依赖训练集的规模，这就会导致花费巨大的额外开销来建立足够规模的训练集。监督学习方法在文本分类中取得了很好的效果[15]。最早把监督学习运用在情感分类任务中，文本结构以短语的形式表达出来，运用不同分类方法并取得了很好的效果[16]。在监督学习中，从文本特征选择方面展开思考，详细比较和分析了各类统计选择法所产生的效果和差异原因[17]。有学者尝试利用联合最优方式来约束错误的传播，进而采用条件随机场方式来提高分类的准确率[18]。有学者对文本处理的各个阶段的多种做法进行效果比较，分析表明如果训练数据规模合适，运用二元短语结构、信息增益和 SVM 的搭配可以得到相对满意的效果[19]。

近年来，半监督的分类方法逐渐受到越来越多专家学者的重视。在半监督学习研究的早期，Blum 和 Mitchell 提出了协同训练方法[20]，并证明了如果两个视图是充分冗余且相互独立的话，通过协同训练方法可以利用未标注数据使弱分类器的性能得到提升。然而在实际任务中，理论上所要求的视图之间的条件独立性，通常并不能得到保证。其中，Wan 提出把外语与中文当成相互独立的视图，根据相互独立的视图运用基于半监督的分类方法[21]。而 Li 等人提出以个人视图和非个人视图将用户评价分成两个不同的视图的协同训练分类方法[22]，在 Book 和 Kitchen 领域表现出很好的结果。代大明等人尝试将情感词和情绪词作为不同的视图来进行协同训练的情感分类也获得了一定的成效[23]。经过进一步的研究，Li 等人构建 Textual 和 Social 两个独立视图进行协同训练[24]，实验也表明，基于这两个视图的半监督学习方法十分有效，其中提出的算法的 F-score 比 baseline（Textual）的 F-score 高出 13%。苏艳等人尝试提出了一种动态随机特征子空间方案[25]，通过一系列实验发现该方案取得了比一般静态随机特征子空间方案更好的效果。Wang 和 Zhou 证明协同训练方法的条件定理[26]，并提出协同训练需要的并不是真正的多个视图，而是分类器之间需要存在一定的分歧性。有学者专门介绍基于分歧的半监督方法，表明了子分类器之间的多样性对分类效果起着关键作用[4]。然而在实际任务中，一般不会有多个天然相互独立的视图，采用随机特征子空间方法会导致特征子空间中包含的信息量不足，使分类器退化成弱分类器。

一般而言，通过集成学习，分类器性能往往比只使用一个子分类器要好[27]。目前，比较著名的集成学习方法包括了 Bagging 算法[28]，Random Subspace 算法[29]，Boosting 算法[30]和 Stacking 算法[31]。Hansen 提出了当子分类器具有较好的分类性能且有一定相对独立时，集成出来的分类器性能将会有所提高[32]。有学者认为集成学习可以在一定程度上改善分类器的性能，它将会变成一个极具研究价值的分支[15]。Zhou 等人提出使用“选择性集成”方法，并且证明出组合部分子分类器会比组合所有子分类器要好，即采用“选择性集成”方法就可以取得较好的表现[33]。在后续的研究中，Zhou 和 Li 提出了利用子分类器结果和训练集之间的互信息去评价子分类器的性能，以及利用子分类器之间的互信息评价子分类器之间的独立性[34]。Yu 等人提出了一种多目标优化的半监督集成分类的方法，基于随机子空间下利用特征采样计算所有子空间的最优分类器，有效地提高了分类器的性能[35]。李寿山和黄居仁将四种不同的分类方法应用在中文情感分类任务上[6]，并且发现对于所有的领域，运用 Stacking 方法都可以得到优于基分类器的效果。通过进一步的研究，Li 等人将两个半监督学习算法通过 Stacking 算法进行集成[5]，同时将已标注的数据分成训练集和验证集，实验表明该做法能大大提高了分类器的整体性能，其中准确率平均比 baseline 高出 4.95%。有学者运用基于一致性的集成学习方法，对未标注数据进行预测，同时规定只有预测标注一致时才能

被加入训练集，实验表明这个方法提高了对未标注数据标注的正确率，比任一种半监督方法都要表现得好[36]。但是对于噪声数据，同样存在被所有子分类器同时学习的风险。

本文提出的方法尝试从改变训练样本分布的角度上，增加子分类器训练集之间的多样性，同时降低噪声数据的扩散范围。在集成学习的基础上，运用半监督学习方法将未标注数据利用起来，使分类性能得到进一步提高。

3 基于优化样本分布抽样集成学习的半监督分类方法

3.1 基于集成学习的监督学习与半监督学习分类方案面临的问题

虽然基于集成学习的监督学习与半监督学习分类方案在一定程度上能够优于仅使用基于监督学习与半监督学习的分类方案。但这些基于集成学习的监督学习与半监督学习分类方案仍然存在很多不足之处，主要体现在以下几个方面：

（1）监督学习的分类性能很大程度上依赖已标注数据的数目，而在实际的文本中，已标注数据的数量是很少的。对于在已标注数据很少的条件下学习而来的分类器，往往泛化能力不好，从而导致分类性能不能满足要求。

（2）虽然可以运用半监督方法，但因为已标注数据规模很少，起始分类器一般不能具备较好的泛化能力，导致对未标注数据预测错误。相比于原始的已标注数据，这些从未标注数据中经分类器预测后学习而来的样本数据的正确性不能有很好的保障。如果原始的已标注数据因未被抽样到而导致分类器未能学习到该部分样本数据，这就会对分类器的分类性能产生很大的影响。

（3）在实际任务中，一般不会有多个天然相互独立的视图，如果采用随机特征子空间方法把总空间划分为多个独立视图，就会导致特征子空间中包含的信息量不足，使分类器退化成弱分类器，最终导致分类器分类性能的降低。在半监督学习中，弱分类器由于预测错误产生的噪声会在后续的迭代过程中被不断放大，最终导致分类器的总体分类性能的不断降低。

（4）如果采用重抽样采样的方式，从已标注数据集中获得多个子训练数据集，在一定程度上是可以增加子训练数据集之间的多样性，每个子分类器所包含的文本特征信息也不会缺失。但是一般的重抽样采样方法不能很好保证各子训练数据集之间存在较大的多样性。另一方面，因为这种方式是在所有的已标注数据中进行重抽样采样，有可能导致噪声数据被所有子分类器同时学习，从而造成所有子分类器在这个局部阶段上同时降低了分类性能。对于在半监督学习中，分类器在这个局部阶段时预测错误的噪声会被加入到已标注数据中，同样，这些噪声会在后续的迭代过程中被不断地放大。

本文将提出一种基于优化样本分布抽样集成学习的半监督分类方法（SSEOS），尝试去解决上述几个存在的问题。

3.2 基于优化样本分布的抽样方法

Bagging 算法是由 Breiman 提出的一种集成学习方法[28]，在个体样本生成阶段，运用随机重复抽样方式，获取若干个子训练数据集，这种方法有利于提升总体泛化能力。但是这种重抽样采样方法不能很好地保证子训练数据集之间存在较大的多样性。为了增加子训练数据集之间的多样性，在 SSEOS 中，提出基于优化样本分布的抽样方法，从已标注数据中按照不同的样本分布对其进行采样，获得多个不同样本分布的子训练集。为了避免原始的已标注数据因未被抽样到而导致分类器未能学习到该部分标注信息明确的训练数据，对于子分类器的训练数据集，还提出了基于 Bagging 的优化分布重抽样方法。

如图 1 所示，为本文提出的基于优化样本分布的抽样方法。其中，基于优化样本分布的抽样方法的思路为：首先，将所有原始的已标注数据添加到每一个子训练集；接着，将经未标注数据学习而来的数据按不同分布进行抽样。即将经未标注数据学习而来的数据划分成 T

个不同分布的数据集（为了增加训练样本之间的多样性，T 的值通常设置成比较小），然后每次将 T-1 个不同分布的数据集添加到对应的子训练集，形成新的子训练集。其中，有阴影的数据代表被选中作为新的子训练集的组成数据；下一步，对新的子训练集进行基于 Bagging 的优化分布重抽样方法抽样，得到最终的子分类器训练集。与一般重抽样不同，采用基于 Bagging 的优化分布重抽样方法对标注信息明确的原始的已标注数据和从未标注数据中学习而来的数据之间按比例进行重抽样采样。即如果其中一个部分含有 m 个训练数据，则从该部分的数据中重抽样 m 次，最终得到一个大小为 m 的训练数据集。最后，把这两部分按比例重抽样的数据合并起来，作为新的子分类器训练集。

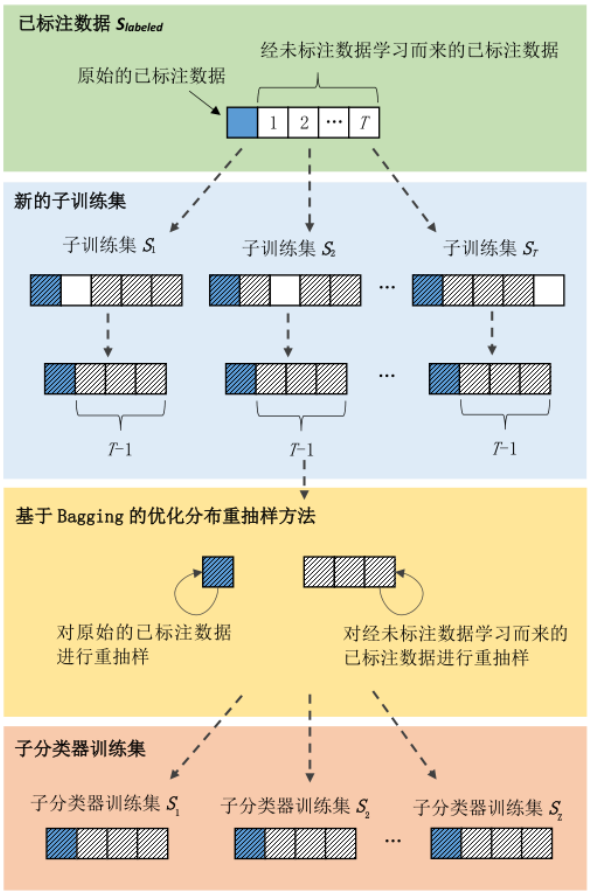

图 1　基于优化样本分布的抽样方法（子训练集的数目为 T）

因此，基于优化样本分布的抽样方法，不仅可以增加子训练集之间的多样性，还可以控制已标注数据中噪声的扩散范围，从而避免噪声数据被所有子分类器同时学习。另一方面，每一个子训练集中都包含所有原始的已标注数据，确保了子分类器都能学习到这些标注信息明确的数据。对于每一个子训练集的数据，采用基于 Bagging 的优化分布重抽样方法采样得到一组子分类器训练集，不仅可以确保每一个子分类器都能学习到足够的原始的已标注数据，还可以提高子分类器训练集之间的多样性，从而提升分类器的泛化能力。

3.3 基于优化样本分布抽样集成学习的半监督分类方法

在 SSEOS 中，首先运用 jieba 分词器[1]对中文文本进行分词，文本特征选取基于词的 Uni-gram 特征。对于分词完毕后的文本内容，再进行去停用词操作。

由于在实际的文本中，已标注的数据量比较小，如果根据已标注数据使用词频、文档

[1]https://github.com/fxsjy/jieba

频率、CHI、MI 等方法进行特征选择的话，部分未能出现在已标注数据的特征就会受到严重的影响。但是如果采用全文本特征的话，文本特征空间将会变得很巨大，造成较大的训练开销和存储开销，且全文本特征空间会夹杂着许多噪声词，影响分类器的性能。

SSEOS 对文本特征选择，仅仅采用了基于 DF 方法，选择在所有样本数据中 DF 大于 5 且小于 1000。这样做不仅为了避免因为已标注数据的数量过少而导致无意间损失了部分重要的文本特征信息，还可以降低特征空间大小，减少训练和存储带来的开销。其中，本文设计的 SSEOS 的总体结构框架如图 2 所示。

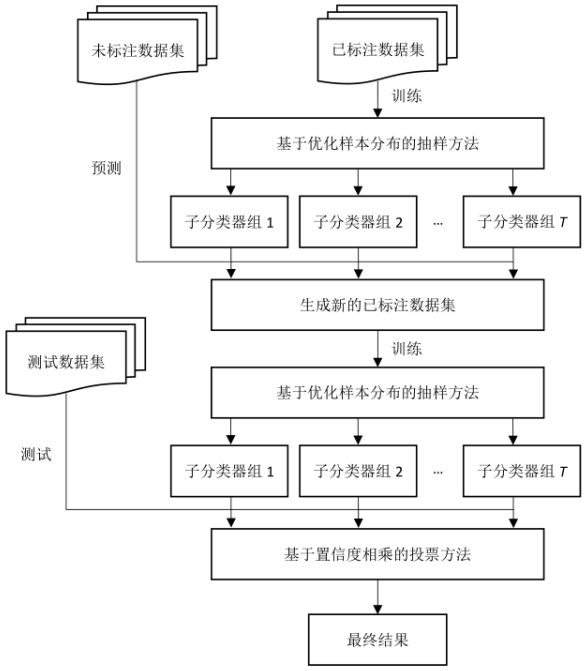

图 2　SSEOS 的总体结构框架图

在图 3 中，用伪代码详细描述基于优化样本分布的抽样方法。其中子分类器组的数目 T 和子分类器组含有子分类器的数目 K 都是预先设定好的。

在基于置信度相乘的投票方法中，数据$\boldsymbol{x_i}$被预测为第 z 个类别的置信度$\boldsymbol{C_z}$可以表示为：

$$\boldsymbol{C_z}(\boldsymbol{x_i}) = \prod_{t=1}^{T*K} \boldsymbol{H_{zt}}(\boldsymbol{x_i}) \quad (1)$$

采用基于置信度相乘的投票方法，最终对数据$\boldsymbol{x_i}$的预测结果如下：

$$\boldsymbol{y_i} = (\arg \max_{z\in\{1,2,\ldots,Z\}} \boldsymbol{C_z}(\boldsymbol{x_i})) \quad (2)$$

其中，$\boldsymbol{x_i}$为样本数据，Z为类别数目，$\boldsymbol{y_i}$是对$\boldsymbol{x_i}$的预测标注结果，$\boldsymbol{H_{zt}}(\boldsymbol{x_i})$表示第 z 个类别的第 t 个子分类器将$\boldsymbol{x_i}$预测为第 z 个类别的置信度。

最后，通过基于置信度相乘的投票机制进行结果集成，每次从各类别中对结果Top-100的未标注数据赋予相应标注，然后加入到已标注数据集中进行下一轮的迭代。

算法： 基于优化样本分布的抽样方法

输入： 原始的已标注数据集 $S_{labeled} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$

经未标注数据学习而来的已标注数据集

$$S_{unlabeled} = \{(x_{n+1}, y_{n+1}), (x_{n+2}, y_{n+2}), \ldots, (x_{n+m}, y_{n+m})\}$$

子分类器组的数目 T

类别数目 Z

一个子分类器组含有的子分类器的数目 K

输出：（Z*T*K）个子分类器训练集

过程：

```
1.  将S_unlabeled划分成 T 个子数据集
2.  for t=1 to T do
3.    将所有原始的已标注数据集S_labeled添加到新的子训练集
      S_train−t中
4.    将S_unlabeled中除了第 t 个以外的 T-1 个子数据集添加到新的
      子训练集S_train−t中
5.  end for
6.  for z=1 to Z do
7.    for t=1 to T do
8.      for k=1 to K do
9.        对第 t 个新的子训练集S_train−t的前 n 个数据，进行基于
          Bagging 的重抽样，然后将抽样的数据添加到子分类器
          训练集S_subtrain−z−t−k中
10.       对第 t 个新的子训练集S_train−t的除前 n 个以外的数据，
          进行基于 Bagging 的重抽样，然后将抽样的数据添加到
          子分类器训练集S_subtrain−z−t−k中
11.     end for
12.   end for
13. end for
```

图 3　基于优化样本分布的抽样方法的伪代码图

4 实验

4.1 实验设置

本文中实验的数据集采用了全网中文新闻数据（SogouCA）的“新闻标题”和“新闻内容”文本数据[2]。选用其中的 6 个主题的“新闻标题”和“新闻内容”分别进行短文本和长文本的分类实验，具体包括：“体育”，“健康”，“教育”，“时尚”，“汽车”和“财经”。其中，在“新闻标题”的短文本实验中，已标注的数据集有 1200 条，未标注的数据集有 1800 条，测试集有 9000 条。在“新闻内容”的长文本实验中，已标注的数据集有 600 条，未标注的数据集有 1800 条，测试集有 9000 条。

本文的实验将从整体正确率（Accuracy）和每个类别对应的 F-score 评价分类效果，F-score 是准确率（P）与召回率（R）的调和平均数。

对于每一个类别，这里用 TP 代表真正例；这里用 TN 代表真负例；这里用 FP 代表假正例；这里用 FN 代表假负例。

其中，Accuracy 和 F-score 的公式如下：

[2]http://www.sogou.com/labs/resource/ca.php

$$Accuracy = \frac{\sum_{i=1}^{Z} TP_i}{\text{测试集数目}} \quad (3)$$

$$F = \frac{2P \cdot R}{P+R} \quad (4)$$

4.2 实验结果与分析

为了证明所提出算法的有效性，本文实现了基于集成学习的半监督学习方案和常见算法来进行对比研究。实验运用了五折交叉验证方法来选取模型参数，实验对比方法包括了：

（1）**基于 Logistic 的监督学习方案**：不对未标注数据做任何操作，采用基于 Logistic 的监督学习方法训练分类模型。

（2）**基于 Self-training 的半监督学习方案**：结合未标注的数据，采用基于 Self-training 的半监督学习的方法训练分类模型[37]。

（3）**基于 Random Subspace 集成的监督学习方案**：不对未标注数据做任何操作，将文本特征空间进行随机划分[29]，接着运用监督学习方法训练分类模型，最后，运用置信度相乘投票法进行结果的集成。

（4）**基于 Random Subspace 集成的半监督学习方案**：将文本特征空间进行随机划分，接着运用半监督学习方法训练分类模型[21]，最后，运用置信度相乘投票法进行结果的集成。

（5）**基于 Bagging 集成的监督学习方案**：不对未标注数据做任何操作，将已标注数据进行重抽样采样获得子训练集[28]，接着运用监督学习方法训练分类模型，最后，运用置信度相乘投票法进行结果的集成。

（6）**基于 Bagging 集成的半监督学习方案**：将已标注数据进行重抽样采样获得子训练集，接着运用半监督学习方法训练分类模型[38]，最后，运用置信度相乘投票法进行结果的集成。

（7）**SSEOS**：本文所提出的基于优化样本分布抽样集成学习的半监督分类方法，首先，采用所提出的基于优化样本分布的抽样方法获得子训练集；然后，采用基于 Bagging 的优化分布重抽样方法获得子分类器训练集；接着，采用半监督学习的方法训练分类模型；最后，运用置信度相乘投票法进行结果的集成。

在长文本实验中，各分类方案在所有主题上的总体正确率如表 1 所示，在 6 个主题上的实验结果如图 4 所示。

表 1　各分类算法的实验结果对比

算法方案	长文本 Accuracy (%)		短文本 Accuracy (%)	
	Data1（监督）	Data1+Data2（半监督）	Data1（监督）	Data1+Data2（半监督）
Logistic	88.12	-	71.11	-
Self-training	-	89.26	-	72.17
Random Subspace	87.55	88.13	62.28	70.58
Bagging	88.63	89.27	71.46	72.50
SSEOS	**89.00**	**89.50**	**71.97**	**72.97**

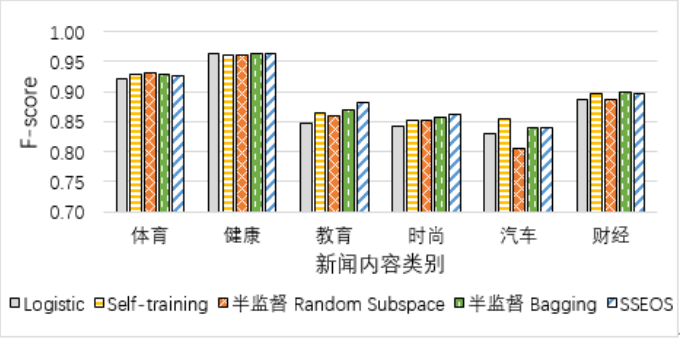

图 4　长文本中各分类算法的各类别的 F-score 对比

其中 Data1 表示 600 条已标注数据集，Data2 表示 1800 条未标注数据集。从实验结果可以看出基于半监督学习方案的结果比基于监督学习方案的结果要好。其原因主要是半监督方案能够把大量的未标注数据利用起来，证明了未标注数据的存在价值。不管是监督学习的方案还是半监督学习的方案，大部分引入集成学习方法之后，分类的整体正确率和各类别 F-score 都有所提升。其中本文所提出的实验方案 SSEOS 在 6 个主题上的整体分类正确率表现最佳，优于所有的对比方法。

另外，如图 4 所示，SSEOS 在 6 个主题的 F-score 上综合表现也最佳，较为显著的是，SSEOS 在“教育”主题，相对于基于 Logistic 的监督学习方案提高了 4.1%，相对于基于 Self-training 的半监督学习方案提高了 1.8%，相对于基于 Random Subspace 的半监督学习方案提高了 2.5%，相对于基于 Bagging 的半监督学习方案提高了 1.2%。

在短文本实验中，各分类方案在所有主题上的总体正确率如表 1 所示，在 6 个主题上的实验结果如图 5 所示。

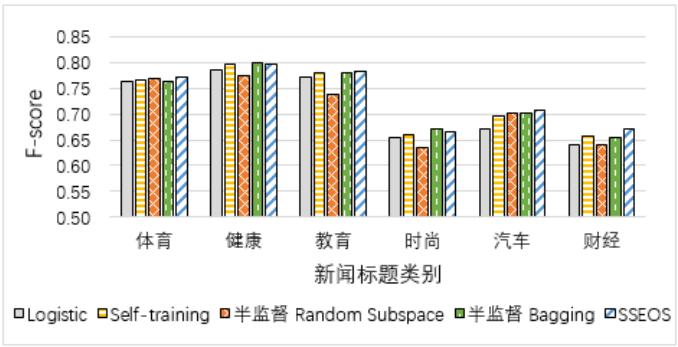

图 5　短文本中各分类算法的各类别的 F-score 对比

短文本实验结果同样表明基于半监督学习方案优于基于监督学习方案，并且不管是监督学习的方案还是半监督学习的方案，大部分引入集成学习方法之后，分类的整体正确率和各类别 F-score 都有所提升。其中本文所提出的实验方案 SSEOS 在 6 个主题上的整体分类正确率表现最佳，同样优于全部对比方法。

另外，SSEOS 在 6 个主题的 F-score 上综合表现也最佳。较为显著的是，SSEOS 在“汽车”主题，相对于基于 Logistic 的监督学习方案提高了 5.3%，相对于基于 Self-training 的半监督学习方案提高了 1.8%，相对于基于 Random Subspace 的半监督学习方案提高了 0.7%，相对于基于 Bagging 的半监督学习方案提高了 1.0%；SSEOS 在“财经”主题，相对于基于 Logistic 的监督学习方案提高了 4.5%，相对于基于 Self-training 的半监督学习方案提高了

1.8%，相对于基于 Random Subspace 的半监督学习方案提高了 4.5%，相对于基于 Bagging 的半监督学习方案提高了 2.5%。

不管是在短文本实验中还是在长文本实验中，基于 Random Subspace 的方案总体表现不佳，特别是在短文本的实验中，其中原因是：在短文本实验中，数据集是具有高度概括性的新闻标题，由于新闻标题的字数长度基本被控制在 20 字以内，本身能表现出来的文本特征很稀疏，采用 Random Subspace 的集成学习方法，使子分类器所能掌握的文本特征变得更少，导致分类性能下降，另一方面，实验是在少量的已标注训练数据的基础上进行的，所以该方案表现不好。但随着未标注数据的加入，基于 Random Subspace 的集成学习半监督方法在分类性能上还是能有所提高的。

5 总结与展望

本文研究基于集成学习的监督与半监督文本分类问题，提出一种基于优化样本分布抽样集成学习的半监督分类方法（SSEOS）。该方法可以得到具有多样性的子分类器训练集，同时可以控制已标注数据中噪声的扩散范围。实验结果表明，我们的方法能够进一步提高半监督文本分类的分类准确率，在短文本和长文本的分类上都具有比较好的分类性能。

本文实验中，所使用的特征是词的一元特征（Uni-gram），在下一步的工作中，我们将尝试把词的二元特征（Bi-gram）应用到文本分类中，进一步提高分类性能。另外，在集成学习方面，文本提出的优化方案 SSEOS 只在样本数据抽样的策略方面进行了优化。在下一步工作中，我们将引入语义信息，更深层次地挖掘文本中存在的语义关系，避免获取不到文本中有用信息的问题。探索多分类器的集成问题，尝试让子分类器自动适应并设定置信度权重，进一步对集成学习方式进行研究。

参考文献

[1] 薛春香, 张玉芳. 面向新闻领域的中文文本分类研究综述[J]. 图书情报工作, 2013, 57(14):134-139.

[2] Krishnalal G, Rengarajan S B, Srinivasagan K G. A New Text Mining Approach Based on HMM-SVM for Web News Classification[J]. International Journal of Computer Applications, 2010, 1(19):98-104.

[3] Sriram B, Fuhry D, Demir E, et al.. Short text classification in twitter to improve information filtering[C]// Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2010), 2010:841-842.

[4] 周志华. 基于分歧的半监督学习[J]. 自动化学报, 2013, 39(11):1871-1878.

[5] Li S, Huang L, Wang J, et al.. Semi-stacking for Semi-supervised sentiment classification[C]// Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL- IJCNLP 2015), 2015:27-31.

[6] 李寿山, 黄居仁. 基于Stacking组合分类方法的中文情感分类研究[J]. 中文信息学报, 2010, 24(5):56-62.

[7] 方丁, 王刚. 基于集成学习理论的文本情感分类[J]. 计算机系统应用, 2012, 21(7):177-181.

[8] Sebastiani F. Machine learning in automated text categorization[J]. ACM Computing Surveys, 2002, 34(1):1-47.

[9] Pang B, Lee L. A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts[C]// Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics (ACL 2004), 2004:271.

[10] 赵妍妍, 秦兵, 刘挺. 文本情感分析[J]. 软件学报, 2010, 21(8):1834-1848.

[11] Turney P D. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews[C]// Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002), 2002:417-424.

[12] 朱嫣岚, 闵锦, 周雅倩,等. 基于HowNet的词汇语义倾向计算[J]. 中文信息学报, 2006, 20(1):14-20.
[13] 刘鸿宇, 赵妍妍, 秦兵,等. 评价对象抽取及其倾向性分析[J]. 中文信息学报, 2010, 24(1):84-89.
[14] 陈涛, 徐睿峰, 吴明芬,等. 一种基于情感句模的文本情感分类方法[J]. 中文信息学报, 2013, 27(5):67-75.
[15] Dietterich T G. Machine-Learning Research: Four Current Directions[J]. AI Magazine, 1997, 18(4):97-136.
[16] Pang B, Lee L, Vaithyanathan S. Thumbs up?: sentiment classification using machine learning techniques[C]. In Proceedings of Emnlp, 2002:79-86.
[17] 代六玲, 黄河燕, 陈肇雄. 中文文本分类中特征抽取方法的比较研究[J]. 中文信息学报, 2004, 18(1):27-33.
[18] 王根, 赵军. 基于多重冗余标记CRFs的句子情感分析研究[J]. 中文信息学报, 2007, 21(5):51-55.
[19] 唐慧丰, 谭松波, 程学旗,等. 基于监督学习的中文情感分类技术比较研究[J]. 中文信息学报, 2007, 21(6):88-94.
[20] Blum A, Mitchell T. Combining labeled and unlabeled data with co-training[C]// Proceedings of the 7th Conference on Computational Learning Theory (COLT 2000), 2000:92-100.
[21] Wan X. Co-Training for Cross-Lingual Sentiment Classification[C]// Proceedings of the 47rd Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing (ACL- IJCNLP 2009), 2009:235-243.
[22] Li S, Huang C R, Zhou G, et al.. Employing personal/impersonal views in supervised and semi-supervised sentiment classification[C]// Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010), 2010:414-423.
[23] 代大明, 李寿山, 李培峰,等. 基于情绪词与情感词协作学习的情感分类方法研究[J]. 计算机科学, 2012, 39(12):245-248.
[24] Li S, Dai B, Gong Z X, et al.. Semi-supervised Gender Classification with Joint Textual and Social Modeling[C]// Proceedings of the 26th International Conference on Computational Linguistics (COLING 2016), 2016:2092-2100.
[25] 苏艳, 居胜峰, 王中卿,等. 基于随机特征子空间的半监督情感分类方法研究[J]. 中文信息学报, 2012, 26(4):85-91.
[26] Wang W, Zhou Z H. A New Analysis of Co-Training[C].// Proceedings of the 27th International Conference on Machine Learning (ICML 2010), 2010:1135-1142.
[27] 周志华. 机器学习[M]. 清华大学出版社, 2016.
[28] Breiman L. Bagging predictors[J]. Machine Learning, 1996, 24(2):123-140.
[29] Ho T K. The random subspace method for constructing decision forests[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20(8):832-844.
[30] Schapire R E. The strength of weak learnability[J]. Machine Learning, 1990, 5(2):197-227.
[31] Wolpert D H. Stacked generalization[J]. Neural Networks, 1992, 5(2):241-259.
[32] Hansen L K. Neural Network Ensemble[J]. IEEE Trans Pattern Analysis and Machine Intelligence, 1990, 12(10):993-1001.
[33] Zhou Z H, Wu J, Tang W. Ensembling neural networks: many could be better than all[M]. Elsevier Science Publishers Ltd., 2002, 137(1-2):239-263.
[34] Zhou Z H, Li N. Multi-information Ensemble Diversity[C]// Proceedings of the 9th International Conference on Multiple Classifier Systems (ICMCS 2010), 2010:134-144.
[35] Yu Z, Li L, Liu J, et al.. Hybrid Adaptive Classifier Ensemble[J]. IEEE Transactions on Cybernetics (TCYB), 2015, 45(2):177-190.
[36] 高伟, 王中卿, 李寿山. 基于集成学习的半监督情感分类方法研究[J]. 中文信息学报, 2013,

27(3):120-127.

[37] Reichart R, Rappoport A. Self-Training for Enhancement and Domain Adaptation of Statistical Parsers Trained on Small Datasets[C]// Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL 2007), 2007:617-623.

[38] Li Y, Su L, Chen J, et al.. Semi-supervised Question Classification Based on Ensemble Learning[C]// Proceedings of the 6th International Conference on Swarm Intelligence (ICSI 2015), 2015:341-348.

作者简介：

徐禹洪（1994—），硕士研究生，主要研究领域为人工智能、自然语言处理。

Email: alvinhong@stu.scau.edu.cn

黄沛杰（1980—），通讯作者，博士，副教授，主要研究领域为人工智能、自然语言处理、口语对话系统。

Email: pjhuang@scau.edu.cn