面向半监督情感分类的特征选择方法研究[1]

王志昊 ， 王中卿 ， 李寿山 ， 李培峰 ， 施寒潇
（苏州大学 计算机科学与技术学院 ， 江苏 苏州 215006）
（浙江工商大学 计算机与信息工程学院 ， 浙江 杭州 310018）

摘要：特征选择旨在降低高维度特征空间，进而简化问题和优化学习方法。已有的研究显示特征提取方法能够有效降低监督学习的情感分类中的特征维度空间。同以往研究不一样的是，本文首次探讨半监督情感分类中的特征提取方法，提出一种基于二部图的特征选择方法。该方法首先借助二部图模型来表述文档与单词间的关系。然后，结合小规模标注样本的标签信息和二部图模型，利用标签传播（LP）算法计算每个特征的情感概率。最后，按照特征的情感概率进行排序进而实现特征选择。多个领域的实验结果表明，在半监督情感分类任务中，基于二部图的特征选择方法明显优于随机特征选择，在保证分类效果不下降（甚至提高）的前提下有效降低了特征空间维度。

关键词：情感分类；半监督学习；二部图；标签传播；特征选择
中图分类号：TP391　　**文献标识码：**A

Feature Selection Method for Semi-Supervised Sentiment Classification

WANG Zhihao，WANG Zhongqing，LI Shoushan，LI Peifeng，SHI Hanxiao
(School of Computer Sciences and Technology，Soochow University，Suzhou，Jiangsu 215006，China)
(School of Computer Science and Information Engineering，Zhejiang Gongshang University，Hangzhou， Zhejiang 310018，China)

Abstract: Feature selection aims to reduce the high-dimensional feature space so as to simplify the problem and improve the learning method. Existing studies have shown that feature selection is effective in reducing feature space in sentiment classification. In this paper, we focus on feature selection method. Different from all previous studies, we attempt to conduct the research on feature selection on semi-supervised sentiment classification. We propose a novel feature selection method based on bipartite graph which focuses on semi-supervised sentiment classification. First, we formulate the relations between documents and words with the help of bipartite graph model. Then, with a small amount of labeled data and the bipartite graph, a label propagation algorithm is applied to calculate the feature probabilities belonging to sentimental categories. Third, the features are then selected according the sentimental probabilities. The experimental results across multiple domains demonstrate that our feature selection method achieves much better performances than random feature selection method. Our approach is capable of significantly reducing the dimension of the feature vector without any loss in the classification performance.

Key words: Sentiment classification; Semi-supervised learning; Label Propagation; Bipartite graph; Feature selection

---

[1] **基金项目：**国家自然科学基金资助项目（61070123, 61003155）；中科院自动化所模式识别国家重点实验室开放课题项目资助；教育部人文社会科学研究青年基金（12YJC630170）；浙江省自然科学基金资助项目(LY13F020007，Z1110551) 。

作者简介：王志昊（1988—）男，硕士研究生，主要研究方向为自然语言处理，情感分析与意见挖掘。

1 引言

随着微博、社交网络、电子商务等互联网应用的迅猛发展，人们习惯于在网络中表达观点或抒发情感。与此同时，网络中的用户行为数据成倍增长，这些数据包含着大量情感信息。传统的基于主题的文本分类系统无法满足对这些主观文本分析的需求，情感分类在这种背景下受到越来越多人的重视[1]。情感分类任务是指对文本自身情感倾向性进行分类，例如，判断某一评论是“赞扬” 或“批评”[2-3]。近年来，情感分类在自然语言处理研究领域已经成为一个热点研究问题[1]。

情感分类的主流研究方法是机器学习方法，大致可以分为无监督学习、监督学习和半监督学习三种。目前大多数研究都基于监督学习并且已经取得了非常好的效果[4]，但是由于监督学习依赖于大量人工标注的训练样本，使得监督学习的分类系统具有很高的标注代价。相对而言，无监督学习方法不需要人工标注训练样本，是最小化标注量的一种解决方案，但由于其分类效果不佳，通常难以达到实际要求[5-6]。半监督学习是采取综合利用少量已标注样本和大量的未标注样本来提高学习性能的情感分类方法[7-9]，它兼顾了人工标注成本和分类效果，被视为一种折中方案。本文主要围绕半监督情感分类方法进行展开。

情感分类任务同其他文本分类一样，面临着高维度特征空间的问题，同时半监督情感分类任务中的训练分类模型的过程也因高维度而变得漫长。特征选择是解决高维问题的一种有效手段，可以在不降低分类效果的前提下达到降维的目的。特征选择方法在文本分类研究中占有非常重要的地位。同时，已有的研究显示特征选择方法能够有效的降低监督情感分类中的特征维度空间[10]。然而针对半监督情感分类，特征选择方法的研究还未涉及。在半监督情感分类中，由于标注样本规模太小，特征在类别中的分布并不能可靠获得。因此，传统的基于监督情感分类的特征提取方法没法直接应用。如何在半监督情感分类中进行特征选择是一个新的具有挑战的问题。

本文首次探讨半监督情感分类中的特征提取方法，提出一种基于二部图的特征选择方法。该方法首先借助二部图模型来表述文档与单词间的关系。然后，结合小规模标注样本的标签信息和二部图，利用标签传播（LP）算法计算特征的正负类的情感概率。最后，按照特征的情感概率进行排序进而实现特征选择。基于选择出的特征，利用标签传播算法进行半监督情感分类。该方法在英文和中文领域都有明显的降维效果，在部分领域的评论语料中，其分类效果超过了使用全特征的半监督学习方法，最高有 4 个点的提升。此外，我们还针对随机特征选择的结果进行了比较研究，实验结果显示我们的方法优势明显。

本文结构安排如下：第二节介绍了半监督情感分类及特征选择方法的相关工作；第三节提出基于二部图的半监督特征选择方法；第四节给出实验结果及分析；第五节给出相关结论。

2 相关工作

2．1 半监督情感分类

近几年来，基于半监督学习的情感分类渐渐受到广大研究者们的重视。文献[8]将两种不同语言（英语和汉语）作为两个不同的视图，采用协同训练方法进行半监督情感分类；Li 等则是把评价语句分为个人视图（Personal View）和非个人视图（Impersonal View）并同样采用协同训练方法进行半监督情感分类[9]。Dasgupta 和 Ng 将谱聚类、主动学习、直推学习和集成学习引入到半监督学习中[7]，但仍未获得较高的分类准确率（在初始标注样本为 100 时，Book 和 DVD 领域的准确率只有 60%）。苏等对协同训练方法进行改进，提出了基于动态随机特征子空间的协同训练算法，并实验验证了当特征子空间数目为 4 左右的时候，该半监督分类方法能够取得最佳性能[11]。Li 等则基于限制性非负矩阵分解（Constrained Non-negative Tri-factorization）的方法实现了这种方式的半监督学习情感分类任务[12]。此外，高等提出了一种基于一致性标签的集成方法，该方法对两种主流的半监督情感分类方法：基于随机特征子空间的协同训练方法和标签传播方法进行了融合，从而有效降低对未标注样本

的误标注率，获得比任一种半监督学习方法更好的分类效果[13]。

2．2 情感分类中的特征选择

情感分类任务作为一种特定的文本分类任务，同其他文本分类一样面临着高维度特征空间的问题。特征选择用于降低高维度特征空间，让文本分类变得更快速，分类更精确[14]。相关研究表明，将特征选择方法 CHI 应用于大规模在线产品评论，可以在不损失性能的前提下减少特征向量维度[15]。Ng 等将 WLLR 方法用于电影评论的情感分类，取得很好的分类效果[16]。此外，Li 等将 DF、MI、IG 等特征选择方法用于主题文本分类和情感分类问题中，有效降低了维度[10]。然而，上述特征选择方法都是基于监督学习的，依赖于人工标注的结果。在半监督情感分类中，如何利用少量标注样本寻找出大量未标注样本中的有效特征是值得研究的问题。据我们所知，在半监督情感分类问题上还没有关于特征选择方法的研究。

3 基于二部图的半监督特征选择方法

3．1 总体框架

为了本文的表述清楚，我们首先给出基于二部图的半监督特征选择方法的总体框架图，如图 3-1 显示。该方法以二部图模型为基础，首先构建情感文本的二部图表示，建立文档到特征的正负类转移概率矩阵。通过多次迭代并对转移概率差值进行排序，选择出那些区分度最高的特征。这些被选择出的特征构成了半监督学习和情感分类过程中的特征空间。此外，在半监督学习中我们选用标签传播算法（LP）作为学习策略，情感分类方面则选用贝叶斯分类器。下文将详细介绍我们提出的基于二部图的半监督特征选择方法的具体实施步骤。

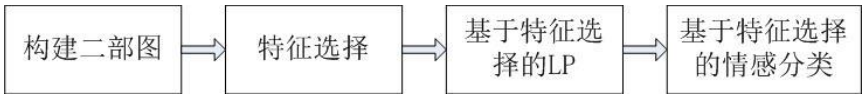

图 3-1：总体框架图

3．2 基于二部图的情感文本表示

情感分类中，文档通常用词袋（Bag-of-words）模型化并用向量形式描述，其缺点是文档与单词间的关联是不清晰的。本文采用的二部图是图论中的一种特殊模型，其顶点集v 可分割为两个互不相交的子集，并且图中每条边依附的两个顶点都分属于这两个互不相交的子集。图 3-2 显示了文档-单词的二部图表示，其中文档用$d_1, d_2 ....., d_n$ 表示，文档中包含的单词用$w_1, w_2 ....., w_n$ 表示。文档-单词的二部图仅存在文档到词及词到文档的连接关系，一篇文档可以包含多个单词，一个单词会在多个文档中出现。显然，通过构建这种文档-单词的二部图可以很清晰地表述文档与单词间的关系。

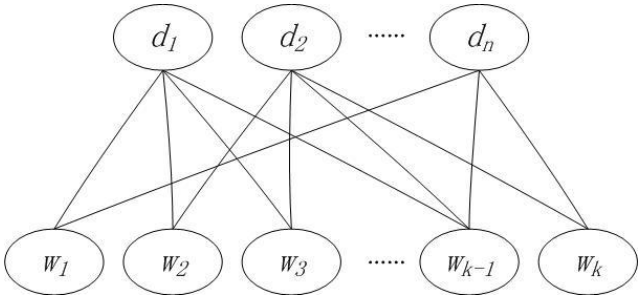

图 3-2：文档与单词的二部图模型

3．3 基于二部图和标签传播的特征选择

上文提到，在半监督情感分类的问题上还没有关于特征选择方法的研究，目前大多数特征选择方法都是围绕监督学习展开的。相关研究表明，特征选择可以在不降低分类效果的前

提下达到降维目的，这也是我们提出这种方法的动机所在。

对于每个文档和文档中所含单词，或文档所包含的特征之间具有如下转移概率：如果文档 d_i 包含 m 个特征且特征 t_k 的权重为 w_{ik}，则文档 d_i 到特征 t_k 的转移概率 $p(d_i \to t_k) = \frac{w_{ik}}{\sum_{k=1}^{m} w_{ik}}$；同理，若特征 t_k 在 n 个文档中出现且文档 d_j 的权重为 w_{kj}，那么该特征到文档 d_j 的转移概率为 $p(t_k \to d_j) = \frac{w_{kj}}{\sum_{j=1}^{n} w_{kj}}$。显然，所有文档到特征 t_k 的转移概率之和 $p(t_k) = \sum_{i=1}^{l+u} \frac{w_{ik}}{\sum_{k=1}^{m} w_{ik}}$，所有 v 个特征到某个文档 d_i 的转移概率之和 $p(d_i) = \sum_{k=1}^{v} \frac{w_{kj}}{\sum_{j=1}^{n} w_{kj}}$。由于半监督情感分类中具有少量已标注样本，包括正类和负类样本，我们由上述转移概率公式求得文档到特征 t_k 的正类转移概率之和 $p_{pos}(t_k) = \sum_{i=1}^{l+u} \left(p_{pos}(d_i) \times \frac{w_{ik}}{\sum_{k=1}^{m} w_{ik}} \right)$，负类转移概率之和 $p_{neg}(t_k) = \sum_{i=1}^{l+u} \left(p_{neg}(d_i) \times \frac{w_{ik}}{\sum_{k=1}^{m} w_{ik}} \right)$，$p_{pos}(t_k)$ 和 $p_{neg}(t_k)$ 的初始值 $P_T^{\,0}$ 都为 0。上述公式中，$p_{pos}(d_i)$ 表示特征到文档 d_i 的正类转移概率之和，$p_{pos}(d_i) = \sum_{k=1}^{v} \left(p_{pos}(t_k) \times \frac{w_{kj}}{\sum_{j=1}^{n} w_{kj}} \right)$，$p_{neg}(d_i)$ 则表示特征到文档 d_i 的负类转移概率之和，$p_{neg}(d_i) = \sum_{k=1}^{v} \left(p_{neg}(t_k) \times \frac{w_{kj}}{\sum_{j=1}^{n} w_{kj}} \right)$。对于标注样本，正类和负类转移概率初始值 $P_L^{\,0}$ 为固定常数（本文中取 $P_D^{\,0} = 1$），对于没有情感倾向的未标注样本，正类和负类转移概率的初始值 $P_U^{\,0}$ 设为 0，在标签传播的过程中，样本到特征以及特征到样本的正负类转移概率被不断更新。本文所提出的特征选择方法认为，某一特征的正类和负类转移概率之和的差值越大，该特征所包含的情感区分度越高，差值的具体计算公式为：

$$\Delta_{opinion}(t_k) = \left| \sum_{i=1}^{l+u} \left(p_{pos}(d_i) \times \frac{w_{ik}}{\sum_{k=1}^{m} w_{ik}} \right) - \sum_{i=1}^{l+u} \left(p_{neg}(d_i) \times \frac{w_{ik}}{\sum_{k=1}^{m} w_{ik}} \right) \right| 。$$

我们选择那些区分度最高的特征作为半监督学习和情感分类过程中使用的特征，特征的极性由转移概率之和高的一方决定。图 3-3 给出了基于二部图的半监督特征选择方法流程。

3．4 基于标签传播的半监督情感分类方法

半监督学习方面，本文使用的是标签传播算法。标签传播算法是 Zhu 等人于 2002 年提出的，它是一种基于图的半监督学习方法，其基本思路是用已标记节点的标签信息去预测未标记节点的标签信息[18]。同样的，这里我们也采用文档-词的二部图来表述文档与单词的关系。文档 d_i 到文档 d_j 的转移概率是由文档 d_i 通过该文档里面的所有词到达文档 d_j 的概率之

和，即 $p(d_i \to d_j) = \sum_{k=1}^{m} \frac{w_{ik}}{\sum_{k=1}^{m} w_{ik}} \cdot \frac{w_{kj}}{\sum_{j=1}^{n} w_{kj}}$。得到文档间的转移概率之后，通过建立标注矩阵和文档-特征的概率转移矩阵计算出未标注样本的标签[18]。此外，本文中的标签传播算法建立在上一步挑选出来的特征，而并不是所有的特征。值得一提的是，实验过程中我们发现，使用了特征选择后的标签传播算法执行效率更高，特别是当特征数目很少时，这种优势非常明显。

基于二部图和标签传播的半监督特征选择算法流程：

输入：

已标初始注样本集合 L，未标注样本集合 U：

所有样本的特征集合 T；

输出：

选择出的特征集合 T^S；

程序：

(1) 初始化：

P：$(l+u+v)\times 2$ 的标注矩阵，同时矩阵前 $l+u$ 行标识文档属于正负类别的概率，后 v 行标识特征属于正负类别的概率

P_L：P 的前 l 行对应的 l 个标注样本

P_U：P 的 $l+1$ 行至 $l+u$ 行对应 u 个未标注样本

P_T：P 的后 v 行对应的 v 个特征

$\overline{T}$：$(l+u+v)\times(l+u+v)$ 的概率转移矩阵，$\overline{t_{ij}},\ i \in (0, l+u)\ and\ j \in (l+u, l+u+n)$ 表示从特征 i 到文档 j 的转移概率，$\overline{t_{ij}}(l+u < j \le l+u+n, 0 < i \le l+u)$ 表示从文档 i 到特征 j 的转移概率，文档到文档的转移概率和单词到单词的转移概率记为 0

a） 设置迭代标记 $t=0$，根据标注样本设定 P_L^0 的值；

b） 初始化 P_U^0 和 P_T^0；

(2) 循环迭代 N 次直到收敛；

a） 文档及特征的正负类标注信息依据公式 $P^{t+1} = \overline{T}P^t$；

b） 还原标注实例的标注信息，即用 P_L^0 替代 P_L^{t+1}；

(3)根据 $\arg\max_{tag}\left(abs\left(p_{neg}(t_k), p_{pos}(t_k)\right)\right)$ 选择区分度最大的特征并得到该特征的正负类标签，添加到 T^S 中；

(4)对 T^S 做平衡处理。

图 3-3：文档与单词的二部图模型

4 实验

4．1 实验设置

实验数据使用了中文和英文两组数据集：其中英文语料采用亚马逊收集的四个不同领域的产品评论，具体为 Book、DVD、Electronics 和 Kitchen，每个领域包含 1100 篇正类和 1100 篇负类评论。实验随机选取正类和负类样本各 100 篇作为初始标注样本，随机选取正类和负类各 800 篇作为未标注样本，剩余的正负各 200 篇作为测试样本。中文语料同样涵盖四个领域的产品评论，分别是化妆品、箱包、电脑和电子产品，每个领域包含 1000 篇正类和 1000 篇负类评论，实验随机选取正类和负类样本各 100 篇作为初始标注样本，随机选取正负各 700 篇作为未标注样本，剩余的正负各 200 篇作为测试样本。实验采用 MALLET 机器学习工具包中的贝叶斯分类器，分类算法的所有参数都设置为默认值。分类选取词的二元特征（Bigram）作为特征。除了与几种常见分类方法作比较外，实验中我们还加入了随机特征选择方法用于对比研究，考虑到该方法的随机性问题，每次实验我们取 5 次实验结果的平均值作为最终结果。

实验中使用准确率（A-ccuracy，Acc.）衡量分类效果，其中，TP 和 TN 代表了被正确分类的正类样本和负类样本，FP 和 FN 代表了被错误分类的正类样本和负类样本。准确率的计算公式如下：

$$Acc. = \frac{TP + TN}{TP + FP + TN + FN}。$$

我们实现以下常见方法的比较研究：

全监督学习：直接使用标注样本及其所有特征训练分类模型（未使用非标注样本）。

使用所有特征的半监督学习：运用标签传播算法后的所有样本直接训练分类器，不使用特征选择方法。

使用特征选择的半监督学习：使用基于二部图的半监督特征选择方法选择特征并构成特征空间，用于标签传播算法和训练分类过程，方法过程参见 3.1。

我们首先实验观察了基于二部图的半监督特征选择方法在英文语料中的表现，之后，在中文领域我们也安排了相应的实验。下面的实验结果中，每张曲线图中横纵坐标表示的含义相同，横坐标为特征数百分比，即选择的特征数目除以总特征数，纵坐标为准确率。

4．2 结果和分析

图 4-1 显示了基于二部图的半监督特征选择方法在英文语料中表现。可以看到，当选择的特征数量大于总特征数的 10%时，我们的方法基本保持稳定的分类效果，很好的达到降维目的。此外，在 kitchen 和 electronics 两个领域，我们的方法表现优异，在保持性能的情况下有 1%-2%的提升。

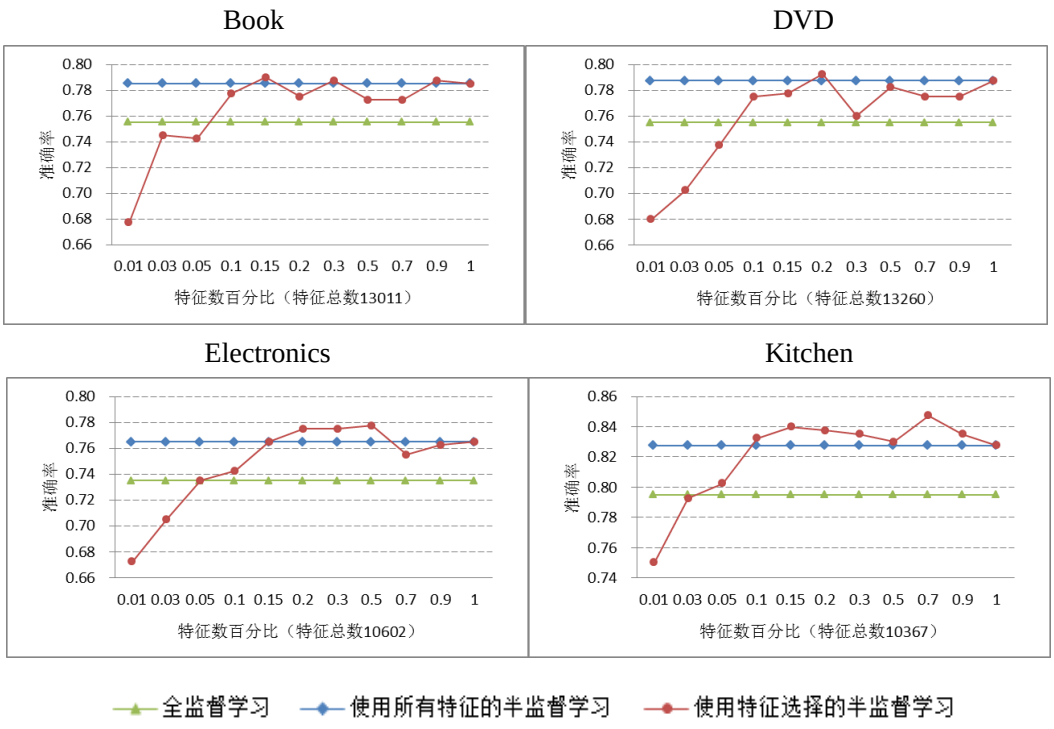

图 4-1：基于二部图的半监督特征选择方法的分类性能比较（英文领域）

图 4-2 显示了我们的方法在中文语料下的分类效果。实验结果表明，基于二部图的半监督特征选择方法在中文语料里也有不俗表现。其中化妆品和箱包两个领域在特征数相对较少时也有很好的分类性能，达到了降维的效果。值得一提的是，在电子产品领域我们的方法表现突出，分类效果基本不低于使用全特征的标签传播算法。在特征数百分比为 0.03，即 136 个特征时达到峰值，分类效果有 4 个点的提高。

图 4-3 和图 4-4 分别显示了选择 200 个和 500 个特征时，随机特征选择和我们提出的方法在中英文领域下的分类性能比较。实验结果清楚的表明：基于二部图的半监督特征选择方法比随机特征选择方法的分类效果好很多，特别是在特征数目很少的时候这种优势越大。除了上图中 200 和 500 这两个固定的特征数目点以外，我们对其他特征数也做了类似的实验。实验结果都显示，我们的方法始终优于随机特征选择方法。

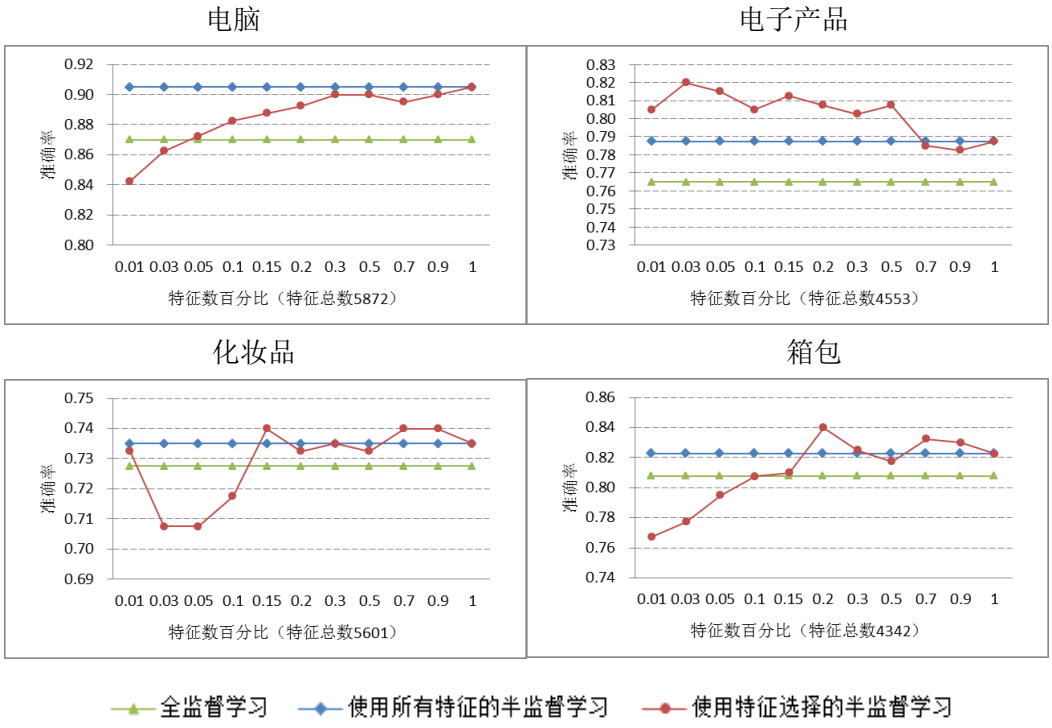

图 4-2：基于二部图的半监督特征选择方法的分类性能比较（中文领域）

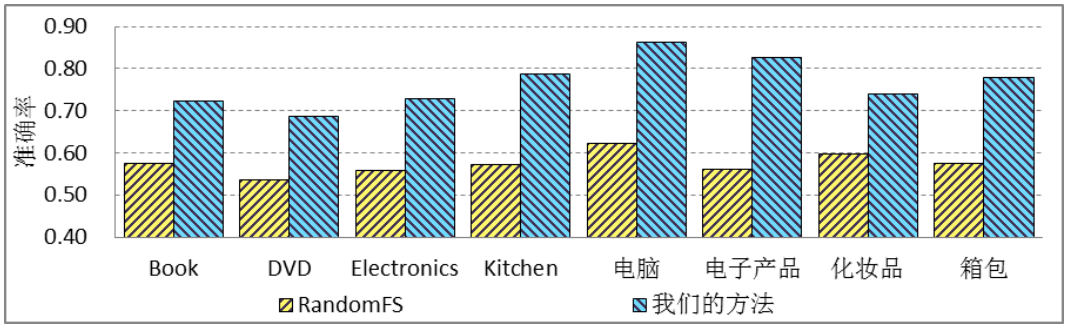

图 4-3：200 个特征下与随机特征选择方法的分类性能比较

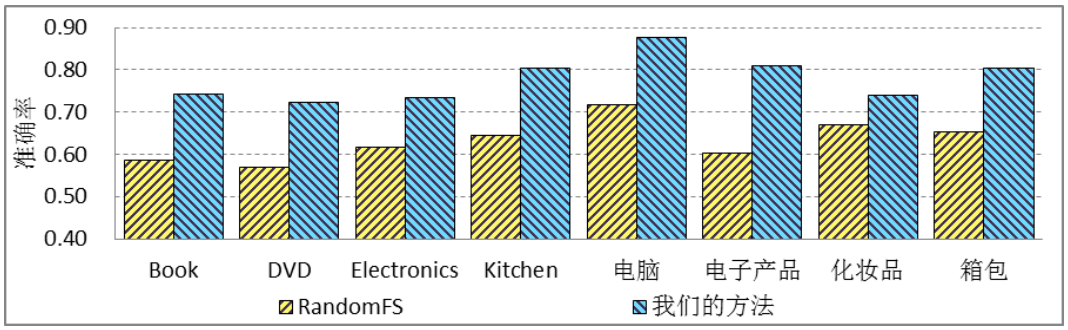

图 4-4：500 个特征下与随机特征选择方法的分类性能比较

4 结语

本文在半监督情感分类中，提出一种基于二部图和标签传播的特征选择方法。该方法首先借助二部图模型来表述文档与单词间的关系。然后，结合小规模标注样本和标签传播算法进行特征提取。实验结果表明，在多个领域的半监督情感分类任务中，基于二部图和标签传播的特征选择方法明显优于随机特征选择。在保证分类效果不下降（甚至提高）的前提下有效降低了特征空间维度。

面向半监督分类的特征选择的研究才刚刚起步，存在很多问题需要我们进一步探讨。例如，从上面的实验结果可以看到，相比使用全部特征的半监督学习方法，我们的方法在大多数领域的分类效果没有很大的性能提升。下一步工作中，我们将尝试加入分类器融合策略，用以稳定和提高最终的分类性能。此外，我们会继续关注半监督领域的高维度特征空间问题，寻找更适合半监督情感分类任务的特征选择方法。

参考文献

[1] Pang Bo, Lee L., Vaithyanathan S. Thumbs up? Sentiment classification using machine learning techniques. Proceedings of EMNLP-02, 2002.

[2] Liu Bing, Hu Minqing, Cheng Junsheng. Opinion Observer:Analyzing and Comparing Opinions on the Web. Proceedings of WWW-05, 2005.

[3] Wiebe J., Wilson T., Cardie C. Annotating Expressions of Opinions and Emotions in Language. Language Resources and Evaluation, 2005.

[4] 唐慧丰, 谭松波, 程学旗. 基于监督学习的中文情感分类技术比较研究. 中文信息学报, 2007, 6(2)..

[5] Zagibalov T., Carroll J. Automatic Seed Word Selection for Unsupervised Sentiment Classification of Chinese Test. Proceedings of COLING, 2008.

[6] Yarowsky D. Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. Proceedings of ACL-05, 2005.

[7] Dasgupta S., Ng V. Mine the Easy, Classify the Hard: A Semi-Supervised Approach to Automatic Sentiment Classification. Proceeding of ACL-IJCNLP-09, 2009.

[8] Wan Xiaojun. Co-Training for Cross-Lingual Sentiment Classification. Proceedings of ACL-IJCNLP-09, 2009.

[9] Li Shoushan, Huang Chu-Ren, Zhou Guodong, et al. 2010. Employing Personal/Impersonal Views in Supervised and Semi-supervised Sentiment Classification. Proceedings of ACL-10, 2010.

[10] Li Shoushan, Xia Rui, Zong Chengqing, et al. A Framework of Feature Selection Methods for Text Categorization. Proceedings of IJCNLP-09, 2009.

[11] 苏艳, 王中卿, 居胜峰, 李寿山, 周国栋. 基于随机特征子空间的半监督情感分类方法研究. 中文信息学报, 2012, 26(4):85-92.

[12] Li Tao, Zhang Yi, Sindhwani V. A Non-negative Matrix Tri-factorization Approach to Sentiment Classification with Lexical Prior Knowledge. Proceeding of ACL-IJCNLP-09, 2009.

[13] 高伟, 王中卿, 李寿山. 基于随机特征子空间的半监督情感分类方法研究. Proceedings of YCCL-2012, 2012

[14] Yang Yiming, Pedersen J. A comparative study on feature selection in text categorization. Proceedings of ICML-97, 1997.

[15] Cui Hang, Mittal V., Datar M. Comparative Experiments on Sentiment Classification for Online Product Reviews. Proceedings of AAAI-06, 2006.

[16] Ng V., Dasgupta S., Niaz Arifin S..Examining the role of linguistic knowledge sources in the automatic identification and classification of reviews. Proceedings of the COLING/ACL Main Conference Poster

Sessions, 2006.

[17] 宗成庆. 统计自然语言处理, 清华大学出版社, 2008.5

[18] Zhu Xiaojin, Ghahramani Z. Learning from Labeled and Unlabeled Data with Label Propagation. CMU CALD Technical Report, 2002.