基于 IDSSL 的文本情感分析研究

王 刚[1,2]，李宁宁[1]，杨善林[1,2]
（1. 合肥工业大学 管理学院，安徽 合肥 230009； 2. 过程优化与智能决策教育部重点实验室，安徽 合肥 230009）

摘要：随着社交媒体的不断普及，网络上出现了大量用户创造的文本信息。这类文本所包含的用户的观点、意见和态度等情感信息，对于互联网用户有着重要的作用，已受到越来越多的重视，并已提出大量有监督的文本情感分析方法来利用这些数据。但文本情感分析中存在大量无标记样本，如何利用大量无标记样本和少量有标记样本进行学习的问题，已成为了文本情感分析领域亟待解决的问题之一。为此，本文提出一种改进的半监督文本情感分析方法 IDSSL（Improved Disagreement-based Semi-Supervised Learning）。该方法以基于分歧的半监督方法为框架，首先利用 Random Subspace 的方式构建多个初始分类器，然后以“多数帮助少数”的方式利用无标记样本训练分类器。最后，在情感分析经典数据集上进行了实验，结果证明了本文提出的方法的有效性，而且取得了比其它半监督学习方法都好的实验结果。

关键字：文本情感分析；半监督学习；多分类器；Random Subspace
中图分类号：TP391 **文献标识码：**A **文章编号：**1004-6062(2018)03-0126-008
DOI: 10.13587/j.cnki.jieem.2018.03.015

0 引言

近年来，随着互联网的快速发展，互联网用户大规模增加。根据最新的《中国互联网发展状况统计报告》显示，截至 2014 年 6 月底，我国网民规模达 6.32 亿，互联网普及率为 46.9%[1]。互联网的广泛普及带动了博客、论坛和社交网络等社交媒体的飞速发展，同时产生了大量源于用户创造的主观性文本[2,3]。这类文本所包含的用户的观点、意见和态度等情感信息，对于互联网用户有着重要的作用[2-4]。例如，消费者在互联网上购买某项产品或服务的时候，一般会参考之前购买者的评论信息，来辅助自己的购买决策行为[3,4]。这些主观性文本的数量急速增加，仅靠人工进行分析需要消耗大量的人力和时间。因此，如何利用信息技术来有效地收集、存储和分析这些主观性的文本所表达的情感信息，已成为当前迫切需要解决的问题，而文本情感分析技术正是解决这一问题的有效工具[2-6] 。

文本情感分析是当前机器学习和数据挖掘领域的研究热点之一，它主要是指利用自然语言处理和文本挖掘技术，分析和抽取主观性文本中所表达的情感信息，识别出其情感倾向[3,4]。目前，对文本进行情感分析主要有两类方法：基于情感知识的方法和基于机器学习的方法[3-6]。 基于情感知识的方法主要依靠一些已有的情感词典和语言知识，比如 SentiWordNet、General Inquire、POS Tragger 等，来对文本的情感倾向进行分类。这类方法主要以自然语言处理为基础，但由于目前自然语言处理领域还存在一些关键技术需要突破，并且基于情感知识的文本情感分类需要事先构建情感知识库，大大限制了基于情感知识方法的进一步发展[3,6]。因此，越来越多的研究者开始关注基于机器学习的方法。基于机器学习的方法主要依靠机器学习中的分类方法来对文本中的情感进行分析，主要包括两个主要步骤：第一步，通过特征构建技术提取主观性文本的情感信息；第二步，使用分类技术对这些文本中所包含的情感信息进行挖掘[5,6]。首先，对于文本情感分析的特征构建，目前使用最多的是在基于词袋（Bag-of-Words）的框架下进行的，在词袋框架下文本被看作是无序词汇的集合。主要使用 N-gram 作为词语特征，其中 Pang 首次将机器学习方法用于篇章级的文本情感分析，并发现使用 Uni-gram 特征得到了最好的分类结果[5]。其次，在文本情感分析中常用的分类技术主要有 NB (Naive Bayes)、ME (Maximum Entropy)和 SVM (Support Vector Machine)等[4-7]。虽然基于机器学习的方法在文本情感分析中获得广泛的应用，但若要得到具有较强泛化能力的分类器，则需要大量的有标记样本，并且在实践中获取有标记的样本需要花费大量的时间和精力。然而，获取无标记样本却十分容易[8,9]。因此，如何利用少量有标记样本和大量无标记样本进行学习已成为了文本情感分析领域亟待解决的问题[8,9]。

当前，利用无标记样本的学习方法主要包括主动学习（Active Learning）、直推学习（Transductive Learning）和半监督学习（Semi-Supervised Learning）[10-13]。其中，主动学习的过程需要“神谕（Oracle）”的参与，直推学习又是建立在“封闭世界（Closed-world）”上的方法，而半监督学习基于“开放世界(Open-world)”的假设，利用大量的无标记样本辅助学习，以提高学习器的泛化能力，因此本研究中主要关注半监督学习的方法。对于文本情感分析中的半监督学习问题，已有一些学者对其进行了初步研究[14-17]。最初，半监督学习被用于基于情感知识的方法，对文本情感分类中的情感词扩

收稿日期：2015-01-05 **修回日期：**2015-10-09
基金项目：国家自然科学基金资助项目（71471054、91646111）；安徽省自然科学基金资助项目（1608085MG150）
作者简介：王刚（1980—），男，江苏连云港人；博士，副研究员；研究方向：商务智能和数据挖掘。

展。如 Turney 和 Littman 利用一些少量的褒义和贬义种子词，分别通过计算面向语义的点互信息和面向语义 LSA 的分数与种子词的相关度，接着预测新的词的情感倾向，然后重复这个过程直到没有新的词可以被标记[14]。这种利用少量标记数据的扩展情感词的方法，本质上是一个自学习（Self-training）的过程。随着研究的开展，研究者也开始探索将半监督学习用于基于机器学习的文本情感分析中。如 Jin 成功将协同训练算法应用于句子级别的文本情感分类，通过选择两个分类器标记句子样本一致性去鉴别相机评论的情感倾向[15]。Wan 针对跨语言的文本情感分类问题，将汉语和英语看作两个不同的视图，然后使用协同训练进行分类[16]。Li 等为了满足协同训练需要两个视图的条件，将文本中的句子分为个人视图（Personal View）和非个人视图（Impersonal View），随后又根据协同训练中的差异性定理，提出了一种利用动态随机子空间方式产生两个视图的协同训练算法[17]。目前，应用于文本情感分析中的半监督学习方法主要还是基于分歧的方法（Disagreement-based Method），并且相对于半监督学习中的其它方法，如生成模型法（Generative Method）、基于图的方法（Graph-based Methods）和 S3VMs（Semi-Supervised Support Vector Machines）等[7-10]，基于分歧的方法较少受到模型假设、损失函数非凸性和数据规模问题的影响，并已被成功应用于很多领域，因此本文也主要在基于分歧的方法的框架下对文本情感分析进行研究[13,18]。

基于分歧的半监督学习方法主要主要通过使用两个或两个以上的分类器来对未标记数据进行利用，最初的基于分歧的方法是 Blum 和 Mitchell 在 1998 年提出的 Co-training 算法，又被称为标准协同训练算法[18,19]。Co-training 算法要求训练样本具有两个条件独立性的视图，但在实际应用中，即使样本具有两个视图也难以满足条件独立性这一条件。故一些学者针对 Co-training 算法这一问题提出了许多改进方法[20-22]，在众多改进方法中，Zhou 等提出的多分类器的方法是一个重要的研究方向[21,22]。多分类器的方法首先构造三个或三个以上的基分类器，然后在学习过程中以“多数帮助少数”的方式为少数分类器提供训练样本。多分类器的方法始于 Zhou 等在 2005 年提出的 Tri-training 算法[21]。Tri-training 算法使用重抽样技术（Bootstrapping）[23,24]训练得到三个分类器[11,21]，然后利用无标记样本进行训练。随后，Zhou 等又将其扩展到更多的分类器的情况，提出了以 Random Tree 作为固定基础分类器的 Co-forest 算法[22]。目前，基于分歧的半监督方法中的多分类器方法主要通过 Bootstrapping 的方式产生，最近 Wang 等通过理论研究发现多个分类器之间的差异性对于基于多分类器的半监督学习方法至关重要[25]。另一方面，对于文本情感分析问题来讲，该问题中存在大量的高维、冗余数据，通过 Random Subspace[26] 的方式相比较于 Bootstrapping 的方式能够得到差异性更大的初始分类器[6]。因此，本文在 Tri-training 和 Co-forest 算法的基础上，提出一种改进的基于分歧的半监督方法（Improved Disagreement-based Semi-Supervised Learning，IDSSL），用于解决带有大量无标记样本的文本情感分析问题。该方法首先在有标记样本中随机生成多个子空间，并在每个子空间上训练得到一个分类器；其次，利用训练好的分类器对未标记样本进行标记，根据“多数帮助少数”的原则，利用多数分类器对置信度（Confidence）较高样本进行标记，并把标记后的样本加入少数分类器的有标记训练集中，以便少数分类器利用这些新标记的样本对分类器进行更新；该过程不断迭代进行，直到达到停止条件；最后，将这些分类器以主投票的方式组合成一个分类器。本研究结合文本情感分析问题的自身特点，以及基于分歧的半监督学习方法的研究前沿，提出了一个新的文本情感分析方法 IDSSL，一方面对于利用少量有标记样本和大量无标记样本进行文本情感分析研究具有重要的理论价值，另一方面也拓展了基于分歧的半监督学习方法的应用领域，具有重要的应用意义。

1 基于IDSSL的文本情感分析模型

本研究主要以基于分歧的半监督学习方法为基础，提出一种基于IDSSL的文本情感分析模型，用于解决文本情感分析研究中如何利用大量无标记样本提高分类器的性能的问题。接下来，我们首先对基于分歧的半监督学习方法的文本情感分析问题进行建模，然后，以此为基础对基于分歧的半监督学习方法进行理论分析，最后，提出基于IDSSL的文本情感分析方法。

1.1 基于分歧的半监督学习方法的文本情感分析建模

文本情感分析通过挖掘和分析网上主观性文本中的立场、观点、情绪等情感信息，对于文本中的情感倾向做出类别判断[3,4]。对于文本情感分析问题，基于机器学习的方法已经取得较好的结果，但若要得到具有强泛化能力的分类器，则通常需要大量的有标记样本。然而在实际文本情感分析应用中，对样本进行标记需要耗费大量的人力物力。与此同时，大量的无标记样本很容易获得。为此，如何利用大量的无标记样本来改善学习性能已成为了当前文本情感分析研究中最受关注的问题之一[8,9]。本研究主要利用半监督学习的方法解决上述问题。在众多半监督学习方法中，基于分歧的方法是一种有着成熟的理论基础和大量实验验证支持的方法，因此本文利用基于分歧的方法来解决文本情感分析中半监督学习问题。

对于文本情感分析中的基于分歧的半监督学习问题，其形式化定义如下：首先，对网上的主观性文本经过分词、去停用词、词干提取等操作后得到文本情感语料集：$D=\{x_1,x_2\ldots,x_{l+u}\}\subset X$；接着，对文本情感语料集 D 中的部分样本进行标记，得到有标记样本集：$L=\{(x_1^{'},y_1^{'}),(x_2^{'},y_2^{'}),\ldots,(x_l^{'},y_l^{'})\}\subset X\times Y$，语料集 D 中剩余的样本作为无标记样本集：$U=\{x_1^{'},x_2^{''},\ldots,x_u^{''}\}\subset X$，这里的 x_i、$x_i^{'}$和$x_i^{''}\in X$，均为表示为 d 维向量的分类特征，y_i和$y_i^{'}\in Y$ 是样本 x_i 的标记，l、u 分别是 L、U 中的样本的数量，且 $u\gg l$；基分类器 $f(x)$，类似于集成学习中的基分类器，即基于分歧的算法中的每一个分类器，可以是同种分类器，也可以是异种分类器。最后，利用有标记样本集 L 和无标记样本集 U 进行学习，期望得到一个可以准确地对样本 x 预测标记 y 的文本情感分类器 $F(x)$。这里的 $F(x)$ 是基分类器 $f(x)$ 的组合分类器，组合方式很多，有乘积、主投票等[18-22]。

1.2 基于分歧的半监督学习方法的理论分析

基于分歧的方法是目前很流行的一种半监督学习方法。

为了使得基于分歧的方法能够在文本情感分析领域中有效地运行，即可以有效地利用无标记样本提升分类器的泛化能力，本文提出一种新的文本情感分析方法IDSSL。IDSSL算法的核心思想主要体现在两方面，一方面由于初始的基于分歧的方法，即Co-Training方法的假设太强，故IDSSL采用多分类器的方法，通过构造多个不同的分类器，以“多数帮助少数”的方式利用无标记样本进行学习。另一方面，进一步考虑到文本情感分析问题是一个高维的分类问题，存在着成千上万的分类特征，Tri-training和Co-forest产生多个分类器的Bootstrapping方法存在着不足，故IDSSL中采用Random Subspace的方式来构造基础分类器。下面从理论上对以上两个方面进行分析。

1.2.1 基于分歧的半监督学习方法中的多分类器方法分析

多分类器的方法是Zhou等对于初始的基于分歧的方法，即Co-training算法的改进方法。多分类器的方法，首先在有标记样本上生成多个不同的分类器，然后利用多数分类器的信息，来排除少数分类器的不确定性，从而改善各个分类器性能.多分类器的方法相比较于Co-training算法，泛化能力更强，而且不需要两个条件独立的视图，适用于范围更广、效率更高。下面我们对于多分类器的方法的相关定义和公式进行介绍。

定义 1 损失函数[13]：对于样本 x，真实标记为 y，分类器 f 对于样本 x 的预测为 $f(x)$，则损失函数定义为：

$$c(x,y,f(x))\in[0,\infty) \tag{1}$$

对于采用多分类器的基于分歧的方法，我们假设在 L 上产生了 k 个分类器，这 k 个分类器的半监督学习模型的学习目标是使下面的函数最小化。Ω 是正则化项，用于惩罚分类器 f 的复杂参数以避免模型过拟合，$\lambda>0$ 为正则化参数，用于调节损失函数和正则化项的比例。

$$\left(f_1^*,\ldots,f_k^*\right)=argmin_{f_1,\ldots,f_k}\sum_{j=1}^{k}\left(\sum_{i=1}^{l}c\left(x_i,y_i,f_j(x_i)\right)+\lambda_1\Omega_{SL}\left(f_j\right)\right)+\lambda_2\sum_{u,j=1}^{k}\sum_{i=1}^{u}c\left(x_i^{'},f_u(x_i),f_j(x_i)\right) \tag{2}$$

一般来讲，上述公式是有可行解的，即多分类器的方法可以有效地利用无标记样本提高分类器的泛化能力。半监督学习领域的专家 Zhu 在文献[13]中的评论证明了这一点。但是，Zhu 没有给出多分类器的方法奏效的具体的关键因素，在后续的研究中，Wang 等通过差异性定理说明了多分类器的方法有效的原因。

1.2.2 基于分歧的半监督学习方法中分类器之间的差异性分析

自从 Co-training 算法被提出以来，学者们一直试图从理论上说明基于分歧的方法可行的原因。Wang 等通过差异性定理证明了分类器之间具有较大差异是基于分歧的方法能够奏效的关键因素[25]。下面我们对差异性理论的相关定义和定理进行简单介绍。

定义 2 差异性[25]：假设有两个分类器 f 和 g，则它们之间的差异性 $d(f,g)$ 定义如下：

$$d(f,g)=P_{x\in D}\left(f(x)\neq g(x)\right) \tag{3}$$

当知道一个分类器的错误率后，两个分类器之间的差异性可以帮助估计出另一个分类器的错误率上界。假设分类器 $f_j^i\in H_j(j=1,2)$ 表示第 j 个分类器在第i轮训练得到的分类器，v 表示每次训练添加的无标记样本数量，$H_j(j=1,2)$ 表示假设空间(Hypothesis Class)。令 a_i 和 b_i 表示分类器 f_1^i 和 f_2^i 泛化错误率，利用初始的有标记样本 L 学习得到的 f_1^i 和 f_2^i 错误率满足 $a_0<\frac{1}{2}$，$b_0<\frac{1}{2}$，并且初始有标记样本数目l满足 Angluin 和 Laird 提出的噪声模型，即 $l\geq max(\frac{2}{a_o^2}ln\frac{2|H_1|}{\delta},\frac{2}{b_o^2}ln\frac{2|H_2|}{\delta})$，令 A_i 表示分类器 f_1^i 的错误率上界，令 B_i 表示分类器 f_2^i 的错误率上界，$A_0=a_0,B_0=b_0$。研究基于分歧的方法是否能够利用未标记数据提高泛化性能，需要分析分类器 f_1^i 和 f_2^i 的错误率的上界 A_i 和 B_i。下面给出定理来说明满足何种条件时基于分歧的方法可以利用无标记样本提高泛化性能。

定理[25] 设 δ 表示学习过程中的置信度参数，那么基于分歧的方法中的分类器 f_1^i 和 f_2^i $(i\geq1)$ 错误率的不等式成立：

$$P\left(d\left(f_1^i,f^*\right)\leq A_i\right)\geq1-\delta?A_i=a_0-\frac{v\cdot\Theta_i}{2l}and\ i\leq\frac{v\cdot\Theta_i^2+4l\cdot a_0\cdot\Theta_i}{4l\cdot a_0^2} \tag{4}$$

$$P\left(d\left(f_2^i,f^*\right)\leq B_i\right)\geq1-\delta?B_i=b_0-\frac{v\cdot\Delta_i}{2l}and\ i\leq\frac{v\cdot\Delta_i^2+4l\cdot b_0\cdot\Delta_i}{4l\cdot b_0^2} \tag{5}$$

其中 $\Theta_i=\sum_{k=0}^{i-1}(d(f_1^i,f_2^k-b_k))$，$\Delta_i=\sum_{k=0}^{i-1}(d(f_1^k,f_2^i-a_k))$，$f^*$ 为错误率为 0 的最优分类器。

上面的定理表明不管是利用多视图的标准协同训练，还是利用不同学习算法的单视图的方法，例如Statistic Co-leaning，或是采用多分类器的方法，如Tri-training和Co-forest等，都是为了使初始的分类器具有较大的差异性，而分类器之间的差异性是如何产生的并不是很重要，只要基于分歧的方法中初始的分类器具有较大的差异性，就能够利用无标记样本辅助有标记样本提高泛化性能，而且分类器之间差异性越大，基于分歧的方法的提升分类器性能的效果越好。因此，对于采用多分类器的基于分歧的方法，需要构建多个具有较大的差异性的初始分类器。

对于多分类器方法如何构建多个差异性较大的分类器，集成学习中的一些方法提供了一些思路[6,23,24,26]。对于集成学习，分类器间的差异性同样很重要，差异性越大，集成的效果越好。故而，Zhou等提出了利用Bagging中的Bootstrapping方法构建三个差异性分类器的Tri-training算法，随后又将其扩展到了多个分类器，提出了Co-forest算法。最新的研究表明，对于文本情感分类问题而言，Random Subspace方法相比于Bootstrapping方法更合适[6]。主要原因在于文本情感分析问题的一个重要特征：和文本分类问题一样，文本情感分析问题的分类特征往往成千上万，并且存在一定的噪声。相

对于Bootstrapping的方法，Random Subspace能通过将分类特征划分为不同子集的方式使得分类器之间的差异性更大，同时能够减轻高纬度和噪声问题。故而IDSSL采用Random Subspace方法在文本情感分析数据中构建多个差异性分类器。

1.3 基于IDSSL的文本情感分析方法

IDSSL 算法是本文对于基于分歧的半监督学习方法的一种改进方法，主要针对文本情感分析领域中的半监督学习问题。IDSSL 算法中的主要核心思想是采用多分类器的方法和 Random Subspace 方法。 IDSSL 算法主要分为三个步骤：首先，用有标记样本集L构建初始分类器；然后，利用多数分类器对无标记样本集 U 的预测，来扩大少数分类器的有标记样本集，重新训练分类器；最后输出最终分类器 $F(x)$ 。

1.3.1 构建初始分类器

IDSSL 算法构建初始分类器的步骤如下：对于 d 维样本 $x \in L$ ，即 $x=\{e_1,e_2,\ldots,e_d\}$ ，随机子空间为r维，满足 $r<d$ 。首先在原始 d 维特征空间中随机选择 r 个特征构建 r 维随机子空间样本 $x_{sub}=\{e_1^s,e_2^s,\ldots,e_r^s\}$ 。这样就可以构建由 r 维样本 x_{sub} 组成的有标记样本的随机子空间集合 L_{sub} 。重复此过程 K 次，得到样本特征空间的 K 个子空间， $L_{sub_k}(1\le k\le K)$ 。然后，在每个子空间集合 L_{sub} 上使用分类技术训练得到一个分类器 f_i ，这样构建 K 个不同的分类器 $f_i(1\le k\le K)$ 。

1.3.2 扩大有标记样本集

对于IDSSL方法利用无标记样本的训练过程，本文的方法借鉴了Tri-training和Co-forest方法，以“多数帮助少数”的方式为少数学习器产生伪标记样本。为了提高多数分类器为少数分类器提供的伪标记样本的准确率，本研究利用 $K-1$ 个分类器为一个分类器提供伪标记样本。具体来讲，在每次迭代中，为了选择合适的无标记样本来扩大少数分类器的训练集，用 $K-1$ 个分类器对未标记样本进行预测，将 $K-1$ 的分类器的预测结果集成起来，并选择集成结果中置信度最高的样本加入到剩余的一个分类器的训练集中。为了进一步减少添加样本的错误率，本文采用多分类器的方法中特有的置信度 φ 控制添加的样本。具体来讲，对于对于 $K-1$ 个分类器，如果对于一个无标记样本所做标记的分类器比例超过预先设置的置信度 φ ，就将这个样本连同所做的标记一起加入到剩余的一个分类器的训练集中。

定义 3 置信度 ：假设有 N 个分类器 f ，对于某个样本 x ，对于二分类问题，若有 m 个分类器预测为正例，有 $(N-m)$ 个分类器预测为反例，置信度 φ 为：

$$\varphi=\frac{max(m,N-m)}{N} \tag{6}$$

1.3.3 集成分类器

训练完成后，得到 K 个分类器，分别对文情感分析的测试样本进行分类，最后以主投票的方式将 K 个分类器进行集成。

$$F(x)=arg\,max_{y\in Y}\sum_{i=1}^{K}1(y=f_i(x)) \tag{7}$$

综上所述，IDSSL算法如图1所示。

Input: Labeled example set　L %标记样本
Unlabeled example set　U %无标记样本
Base learning algorithm　f %基分类器
Maximum number of learning rounds　T %最大迭代次数
Number of selected feature rate sub_{rate} %子空间维度比率
The add number of each class　p and n %各类别添加数量
Number of subspace K　%子空间个数

Process

129 EMBED Equation.DSMT4　$For\ i=1,2,\ldots,K$
$L_i=L$
$Seed_i$ =*Random Seed(L, sub_{rate})* %产生随机数
L_{sub_i} = Subspace(L_i, sub_{rate}, $Seed_i$) %构建子空间
$f_i=f(L_{sub_i})$ %训练分类器
End
$For\ t=1,2,\ldots,T$
$For\ i=1,2,\ldots,K$
%将分类器 $f_1, f_2, \ldots f_j,..f_K$ 利用主投票的方式合成一个分类器 $F_i^{'}(i\neq j)$
$F^{'}(x)=arg\,max_{y\in Y}\sum_{j=1}^{K}1(y=f_j(x))(i\neq j)$
%$F_i^{'}$ 选择正例置信度最高的 p 个样本和反例置信度最高的 n 个样本，并且满足 $\varphi(x)>\varphi$ ，则标上
%相应的伪标记，加入临时标记样本集 $L_i^{'}$ 中
$F_i^{'}$ selects p number of the most confident positive and n number of the most confident negative
examples from U to label, and if $\varphi(x)>\varphi$, then add them into $L_i^{'}$ with pseudo label
$L_i=L_i\cup L_i^{'}$　%将临时标记样本集 $L_i^{'}$ 与分类器 f_i 所对应的训练集 L_i 合并
End
$L^{'}=L_1^{'}\cup L_2^{'}\cup\ldots\cup L_K^{'}$
$U=U-L^{'}$　%从无标记样本集中删除已添加的样本
$For\ i=1,2,\ldots,K$　%重新训练分类器
$L_{sub_i}=Subspace(L_i,k,Seed_i)$
$f_i=f(L_{sub_i})$
End
End

Output : $F(x)=arg\,max_{y\in Y}\sum_{i=1}^{K}1(y=f_i(x))$

图 1 IDSSL 算法

2 实验设计

2.1 实验数据集和评价指标

为了验证 IDSSL 方法在文本情感分析领域中应用的有效性，本文选取了四个经典的情感语料集进行实验，该语料集是 Blitzer 教授收集的多领域情感语料集，主要包括 Book，DVD，Electronic 和 Kitchen 四个子集，每个子数据集分别包括 1000 个正面评论和 1000 个负面评论。

实验的评价指标采用目前文本情感分析领域常用的评价指标：平均分类精度（Average Accuracy）指标，定义如下：

$$Average\ Accuracy=\frac{TP+TN}{TP+FP+FN+FN} \tag{8}$$

上式中 TP（True Positive）表示分类模型正确预测的正样本数，TN（True Negative）表示分类模型正确预测的负样本数，FP（False Positive）表示分类模型错误预测的正样本数，FN（False Negative）表示分类模型错误预测的负样本数。

2.2 实验流程

本研究采用的实验环境为：计算机 CPU 3.10 GHz AMD FX(tm)-8120 Eight-Core CPU，内存 8GB，操作系统 Microsoft Windows 7，数据挖掘软件包 WEKA 3.7.0。首先使用 WEKA 自带的 StringToWordVector 函数，剔除停用词，将原始的文本语料转化为 WEKA 所识别的 ARFF 文件格式，最终，Book、DVD、Electronic 和 Kitchen 等四个语料集分别获得 23319、23892、12298 和 10371 个分类特征。

在实验中选用了目前在文本情感分析领域常用的分类器 SVM 作为基础分类器，对 IDSSL 方法在文本情感分析领域中有效性进行验证。本文提出的 IDSSL 方法属于基于分歧的半监督方法，为此，对比实验中我们选取了目前四个最新的基于分歧的半监督学习算法进行比较研究：Self-training、Co-training、Tri-training 和 Co-forest 方法。其中 Self-training、Co-training、Tri-training 和 IDSSL 选取 SVM 作为基础分类器，Co-forest 根据算法要求使用 Random Tree 作为基础分类器。SVM 算法通过 WEKA 下的 SMO 模块来具体实现。由于 SVM 不是概率型分类器，故通过 WEKA 自带的 SetBuildLogistic 函数使得 SVM 对于每个样本都有不同的分类置信度。Tri-training 和 Co-forest 算法来自于 Zhou 和 Li 的源代码[21,22]，Self-training、Co-training 和 IDSSL 算法是在 Eclipse 环境下自行编程实现。由于 Co-training 算法需要两个天然的视图，本实验参照 Nigam 的方法将属性集随机划分成两个大小相近的互斥子集的方式来构建两视图[19]。Self-training、Co-training、Tri-training 和 IDSSL 等算法最大迭代次数设置为 50。IDSSL 的随机子空间数量设置为 10，子空间维数比例根据前期初步实验设置为 0.5，置信度阈值亦根据前期初步实验设置为 0.8。由于添加样本的数量对于 Self-training、Co-training 和 IDSSL 等算法有着重要影响，但是最优的样本添加量无法确定，并且由于实验数据集的正负类别比例为 1:1，故本实验将每个类别的样本添加量分别设置为 1，5，10，15。

为了提高实验结果的可信度和有效性，实验过程使用 5 次 10 倍交叉验证法，即将初始样本集划分为 10 个近似相等的数据集，每个数据集中属于各分类的样本所占的比例与初始样本集中的比例相同，在每次实验中用其中的 1 个数据集作为测试集，用剩下的 9 个数据集组成训练集，其中将训练集按照标记比例选取有标记样本集，剩余的训练集作为无标记样本集，轮转一遍进行 10 次实验。因此，本文的实验结果为 5 次 10 倍交叉验证的平均值。另外，对于训练数据集，我们为了研究不同大小的初始标记样本对于结果的影响，标记比例依次选取 20%，40%，60%，80%。整体实验流程如图 2 所示。

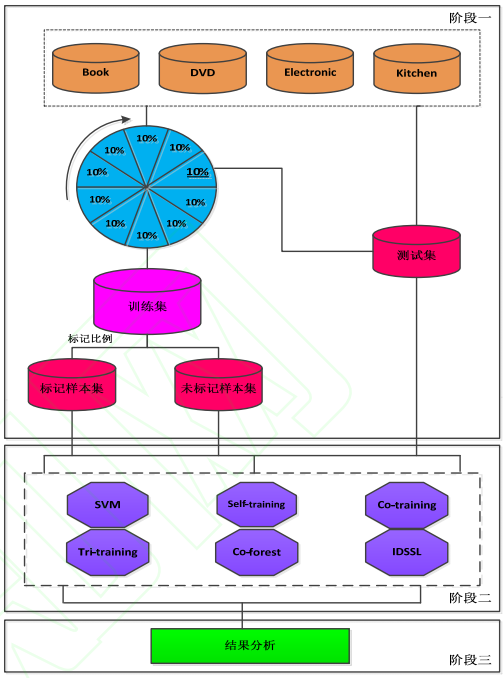

图 2 实验流程

3 实验结果分析与讨论

3.1 实验结果

根据以上实验设计，最终实验结果如表 1 所示。其中 Book、DVD、Electronic 和 Kitchen 分别表示实验用的四个数据集，SVM、Self-training、Co-training、Tri-training、Co-forest 和 IDSSL 分别表示 SVM、Self-training、Co-training、Tri-training、Co-forest 和本文提出的方法。20%、40%、60%和 80%分别表示标记比例。

表 1 实验结果

数据集 \ 方法		SVM(%)	Self-training(%)	Co-training(%)	Tri-training(%)	Co-forest (%)	IDSSL (%)
Book	20%	69.67	71.57	71.68	70.19	52.24	72.58
	40%	72.65	74.82	75.42	73.71	54.68	75.60
	60%	74.33	76.37	76.47	75.77	56.31	77.40
	80%	75.68	76.61	76.74	76.33	57.92	78.11
DVD	20%	69.79	71.98	72.74	71.06	53.15	73.50
	40%	74.27	76.51	76.47	74.90	57.87	77.18
	60%	75.82	77.71	77.53	76.22	59.06	78.90
	80%	77.90	78.83	78.19	78.24	60.54	79.85
Electronic	20%	73.80	75.71	75.60	74.21	62.39	76.49
	40%	77.24	78.78	78.11	77.63	65.19	79.36
	60%	78.40	79.85	79.66	78.68	67.02	80.60
	80%	79.22	80.42	80.61	79.65	68.41	81.12
Kitchen	20%	75.65	78.37	77.02	76.29	64.72	78.59
	40%	79.31	80.98	80.56	79.80	67.66	81.72
	60%	80.15	81.48	81.25	81.36	69.17	82.45
	80%	81.46	82.93	83.27	82.49	70.98	83.55

根据表 1 的结果来看，首先，相对于 SVM，以及采用 SVM 作为基分类器的 Self-training、Co-training、Tri-training 和 IDSSL 方法，由于 Co-forest 采用 Random Tree 作为基础分类器，取得较差的实验结果，这与前人的研究结果也相一致[2,4-6]，也进一步证明了决策树方法在文本情感分析领域应用效果较差。并且由于 Co-forest 的实验结果明显劣于其它几种方法，在后续的分析和讨论中我们就不再重点讨论。其次，四种半监督学习方法 Self-training、Co-training、Tri-training 和 IDSSL 在分类精度上较基础分类器 SVM 都有了显著提高，例如对于 Book 数据集，当标记比例为 20%时，Self-training 的分类精度为 71.57%，Co-training 的分类精度为 71.68%，Tri-training 的分类精度为 70.19%，IDSSL 方法的分类精度为 72.58 % ，均高于基础分类器 SVM 的分类精度 69.67 %，这说明了半监督学习方法在文本情感分析中应用的有效性。并且 IDSSL 方法在四个数据集中分别在不同的标记比例下，均取得了最好的结果，例如对于 DVD 数据集，当标记比例分别为 20%、40%、60%和 80%时，IDSSL 方法均取得了最高的分类精度 73.50 % ，77.18 % ，78.90 %和 79.85 %，这也说明了本研究中提出的 IDSSL 方法的有效性。

3.2 分析与讨论

3.2.1 不同半监督学习方法的对比分析

进一步为了分析这四种半监督学习方法在文本情感分析中应用的不同效果，我们分别采用式（9）计算不同半监督学习方法相对于基础分类器 SVM 的分类精度提高的百分比，得到图 3。

$$Improvement=\frac{Average\ Accuracy_{\text{半监督学习}}-Average\ Accuracy_{\text{基础分类器}}}{Average\ Accuracy_{\text{基础分类器}}} \tag{9}$$

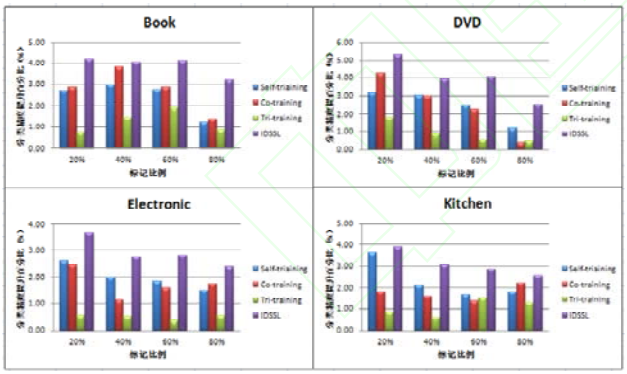

图 3 不同半监督学习方法的分类精度提高结果分析

如图 3 中所示，可以看出：首先，在四种半监督学习中，IDSSL 方法较 Self-training、Co-training 和 Tri-training 方法在文本情感分析中应用更为有效，其主要原因在于采用了多分类器和 Random Subspace 方法。多分类器的方法使得 IDSSL 更加具有泛化能力，同时由于文本情感分析本质上也是属于文本分类的范畴，文本分类中存在大量冗余分类特征，Random Subspace 可以得到差异性比较好的多个分类器，因而使得 IDSSL 能很好地利用无标记样本进行学习。其次，Tri-training 方法相对 Self-training 和 Co-training 方法效果不佳。主要原因是在高维且样本数量较少的情况下，使得 Bootstrapping 所产生的初始分类器差异性不大， 无法发挥多分类器的方法的优势，因而 Tri-training 方法的效果不佳。最后，Co-training 方法在大多数情况下，分类精度要差于 Self-training 方法，此结果与已有的文献结果是一致的[19,21]。这主要是因为 Co-training 方法采用了随机划分属性的方式产生的视图不满足条件独立性假设，因而无法充分发挥多视图学习的优势。

3.2.2 样本标记比例对半监督学习方法的影响分析

半监督学习方法的一个重要参数就是样本标记比例，接下来我们就样本标记比例对半监督学习方法的影响做一分析。我们对样本标记比例分别取 20%、40%、60%和 80%，得到图 4 所示的结果。

从图 4 中可以看出：一方面，随着初始标记样本的比例增加，四种半监督学习方法的分类精度也随之提升。另一方面，虽然半监督学习方法能够利用无标记样本提高分类器的精度，但是标记样本的数量对于分类器也还是很重要的。当样本标记比例达到 60%时，最差的分类器取得的分类精度也比样本标记比例 20%时最好的分类器取得的分类精度要好，这充分说明了有标记样本的重要性，在实践中虽然可以通过半监督学习方法来利用无标记样本，但能够获得足够多的准确有标记样本对于机器学习问题至关重要。

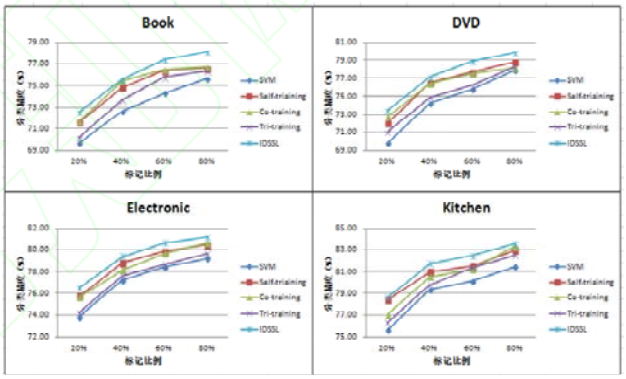

图 4 不同标记比例的结果分析

3.2.3 样本添加量对于 IDSSL 方法的影响分析

IDSSL 方法的一个重要参数就是样本添加量，其取值对 IDSSL 的分类精度有着重要影响。下面我们就对该参数进行分析，根据前面的实验设计，IDSSL 方法中样本添加量分别取 2，10，20，30，得到图 5 所示的结果。

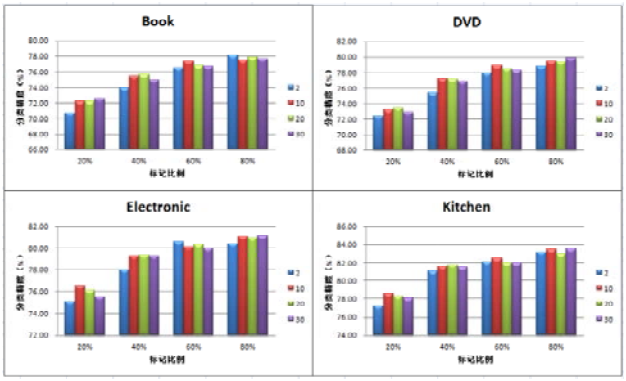

图 5 不同样本添加量的结果分析

从图 5 中可以看出，除了 Book 数据集在 80%的标记样本比例和 Electronic 数据集在 60%的标记样本比例的情况下，

增加样本添加量可以进一步提高 IDSSL 的分类精度。这主要是因为在训练过程中，随着样本添加量增加，使得最终添加到标记样本集中的样本数量在增加，从而可以使得分类器的差异性变大，根据我们在第二节中的理论分析可知，差异性更大的多分类器可以使得 IDSSL 获得更高的分类精度。

4 结束语

随着 Web2.0 的兴起与普及，越来越多的用户乐于在互联网上分享自己的观点或体验，这类评论信息迅速膨胀，仅靠人工的方法难以应对网上海量信息的收集和处理，因此迫切需要计算机帮助用户快速整理和分析这些相关评价信息，文本情感分类技术在这样的背景下应运而生。

对于基于机器学习的文本情感分类问题，如何利用少量的有标记样本和大量无标记样本进行学习是一个亟待解决问题，为此本文提出一个新的半监督文本情感分类方法 IDSSL，对于文本情感分析研究具有重要的理论价值和实践意义。在进一步的研究中，一方面，需要在更广泛的数据集上验证本研究的结论，另一方面，也需要对 IDSSL 做更深入理论分析。

参 考 文 献

[1] 中国互联网信息中心(CNNIC).第 34 次中国互联网统计报告[EB/OL]. http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/hlwtjbg/201407-/P020140721507223212132.pdf, 2014.

[2] 张紫琼, 叶强, 李一军. 互联网商品评论情感分析研究综述 [J]. 管理科学学报, 2010, 13(6): 84-96.

[3] 王刚, 王珏, 杨善林. 电子商务中基于非均衡数据分类和词性分析的意见挖掘研究 [J]. 情报学报, 2014, 33(3): 313-325.

[4] Pang B, Lee L. Opinion mining and sentiment analysis [J]. Foundations and trends in information retrieval, 2008, 2(1-2): 1-135.

[5] Pang B, Lee L, Vaithyanathan S. Thumbs up?: sentiment classification using machine learning techniques [A]. Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10[C]. Association for Computational Linguistics, 2002: 79-86.

[6] Wang G, Sun J, Ma J, *et al*. Sentiment classification: the contribution of ensemble learning [J]. Decision Support Systems, 2014, 57(1): 77-93.

[7] 罗彬, 邵培基和罗尽尧等. 基于多分类器动态集成的电信客户流失预测 [J]. 系统工程学报, 2010, 10(5): 703-711.

[8] Zheng X, Zhu S, Lin Z. Capturing the essence of word-of-mouth for social commerce: Assessing the quality of online e-commerce reviews by a semi-supervised approach [J]. Decision Support Systems, 2013, 56: 211-222.

[9] Yu N. Exploring Co-training strategies for opinion detection [J]. Journal of the Association for Information Science and Technology, 2014, 65(10): 2098-2110.

[10] 周志华, 王珏. 半监督学习中的协同训练风范 [J]. 机器学习及其应用, 北京: 清华大学出版社, 2007: 259-275.

[11] Zhou ZH, Li M. Semi-supervised learning by disagreement [J]. Knowledge and Information Systems, 2010, 24(3): 415-439.

[12] Chapelle O, Schölkopf B, Zien A. Semi-Supervised Learning[M]. Cambridge, MA, USA: MIT Press, 2006.

[13] Zhu X. Semi-supervised learning literature survey [R]. Technical Report 1530, University of Wisconsin-Madison: 2006.

[14] Turney, PD, Littman ML. Measuring praise and criticism: Inference of semantic orientation from association [J]. ACM Transactions on Information Systems (TOIS), 2003, 21(4): 315-346.

[15] Jin W, Ho HH, Srihari RK. OpinionMiner: a novel machine learning system for web opinion mining and extraction [A]. Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining [C], 2009: 1195-1204.

[16] Wan X. Co-training for cross-lingual sentiment classification [A]. Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1 [C], 2009: 235-243.

[17] Li S, Huang CR, Zhou G, *et al*. Employing personal/impersonal views in supervised and semi-supervised sentiment classification [A]. Proceedings of the 48th annual meeting of the association for computational linguistics [C], 2010: 414-423.

[18] Blum A, Mitchell T. Combining labeled and unlabeled data with co-training [A]. Proceedings of the 11th annual conference on Computational learning theory[C]. ACM, 1998: 92-100.

[19] Nigam K, Ghani R. Analyzing the effectiveness and applicability of co-training[A]. Proceedings of the 9th international conference on Information and knowledge management[C]. ACM, 2000: 86-93.

[20] Goldman S, Zhou Y. Enhancing supervised learning with unlabeled data[A]. Proceedings of the 17th International Conference on Machine Learning,[C] San Francisco, CA, 2000: 327-334.

[21] Zhou ZH, Li M. Tri-training: Exploiting unlabeled data using three classifiers [J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(11): 1529-1541.

[22] Li M, Zhou ZH. Improve computer-aided diagnosis with machine learning techniques using undiagnosed samples [J]. IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans, 2007, 37(6): 1088-1098.

[23] 李毓, 张春霞. 基于 out-of-bag 样本的随机森林算法的超参数估计 [J]. 系统工程学报, 2011, 26(4): 566-572.

[24] Breiman L. Bagging predictors [J]. Machine Learning, 1996, 24(2): 123-140.

[25] Wang W, Zhou ZH. Analyzing co-training style algorithms [A]. Proceedings of the 18th European Conference on Machine Learning[C]. Berlin, Heidelberg: Springer-Verlag, 2007: 454-465.

[26] Ho TK. The random subspace method for constructing decision forests [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20(8): 832-844.

Study of text sentiment analysis based on IDSSL

WANG Gang[1, 2], LI Ning-ning[1], YANG Shan-lin[1, 2]

(1. School of Management, Hefei University of Technology, Hefei 230009, China;

2. The Ministry of Education Key Laboratory of Process Optimization and Intelligent Decision, Hefei 230009, China)

Abstract: With the growing popularity of social media, a large number of user generated content is posted on the Internet. These kinds of texts contain user's points of view, opinions and attitudes, which play an important role for Internet users. Researchers pay increased attention to user-generated content. Subsequently, a lot of supervised text sentiment analysis methods have been proposed to make use of this kind of data. However, there are a lot of unlabeled data in the sentiment analysis. How to use a large number of unlabeled data and a small amount of labeled data has become one of the urgent research problems in the area of sentiment analysis. Therefore, this paper proposed an Improved Disagreement-based Semi-Supervised Learning (IDSSL) method for text sentiment analysis, which is based on the framework of disagreement-based semi-supervised learning.

Firstly, a model for sentiment analysis based on the disagreement-based semi-supervised learning was constructed. First of all, the disagreement-based semi-supervised learning was theoretically analyzed. The analysis found that the multiple-classifiers method is better than original disagreement-based semi-supervised learning method. On the other hand, diversity is the key value of the multiple-classifier disagreement-based semi-supervised learning method. Moreover, Random Subspace method can lead to diversity of the classifiers in the area of sentiment analysis. Therefore, we constructed a sentiment analysis model by combining multiple classifiers method produced with Random Subspace method, namely IDSSL method. IDSSL method consists of three steps: (1) multiple initial classifiers are built based on the Random Subspace method; (2) classifiers are trained by the rule of “majority help minority” to utilize the unlabeled instances; and (3) the base classifier was integrated in majority vote.

Secondly, experiments were carried out using the classic datasets of sentiment analysis. The established standard measure in sentiment analysis was adopted to evaluate the performance of the proposed method. IDSSL method is compared with several disagreement-based semi-supervised learning method, including Self-training method, Co-training method, Tri-training method and Co-forest method. Self-training, Co-training, Tri-training, and IDSSL used SVM as base learner. To minimize the influence of variability in the training set, the 10-fold cross validation was performed five times on the sentiment analysis datasets.

Finally, experimental results proved the effectiveness of our proposed method. Moreover, our proposed method obtained better results than the other semi-supervised learning methods, including Self-training method, Co-training method, Tri-training method, and Co-forest method. In addition, we also discuss different semi-supervised learning methods’ results, the influence of the label rate on semi-supervised learning methods, and the influence of the add-number on the IDSSL method.

Key words: Text sentiment analysis; Semi-supervised learning; Multi classifiers; Random subspace

中文编辑：杜 健；**英文编辑：**Charlie C. Chen