部分监督的音乐情感分类*

王 静，朱慕华，胡明涵

东北大学 自然语言处理实验室，辽宁 沈阳 110819

E-mail: {wangjing.neu, zhumuhua}@gmail.com; huminghan@ise.neu.edu.cn

摘 要：基于歌词的音乐情感分析可以看作一个分类问题。为了获得较高的性能，情感分类通常需要人工标注一定规模的数据以便训练分类模型。然而人工构建数据需要以大量的时间和精力为代价。本文以二类情感分类为例，研究在没有负类数据的情况（即只有正类数据和无标注数据）下如何进行音乐情感分类。这一问题被称为部分监督的音乐情感分类问题。由于不需要人工标注负类数据，部分监督方法可有效的减少数据标注代价。据我们所知，这一方法在以前的工作中研究较少。本文采用两个部分监督方法（启发式的二元分割方法和基于统计的 Spy-SVM 方法）在公开数据集“5SONGS 歌词标注语料库”上进行了比较实验。实验结果表明两种方法性能上具有可比性，但在正类数据和无标注数据的分布发生较大变化时，Spy-SVM 能够保持性能相对稳定。

关键词：音乐情感分类；部分监督方法

Partially Supervised Song Sentiment Classification

Wang Jing, Zhu Muhua, Hu Minghan

NLP Lab, Northeastern University, Shenyang 110819

Email: {wangjing.neu, zhumuhua}@gmail.com; huminghan@ise.neu.edu.cn

Abstract: Lyric-based sentiment analysis can be formalized as a classification problem. However, it is time-consuming and labor-sensitive to collect enough labeled training to achieve satisfactory classification accuracy. In this paper we take binary sentiment classification as a case study and study the problem in the situation where only positive and unlabeled data are available. Such a problem is named as partially supervised sentiment classification. Due to the fact the negative training data is not required, partially supervised approach to sentiment classification can reduce the efforts put onto data annotation significantly. In this paper, we apply two methods, heuristic bipartite partition method and statistical Spy-SVM method, respectively. Experiments on the public available corpus “5SONGS lyric corpus” show that these two methods achieve comparable results. Experimental results also show that the latter achieves much more stable accuracy with varying distribution over positive and unlabeled data.

Keywords: song sentiment classification; partially supervised method

1 引言

随着计算机网络的不断发展，用户每天在网络上生成大量的数据。出于研究或商业目的，人们对网络数据中所表达的情感倾向产生了越来越浓厚的兴趣。利用计算机自动挖掘数据中的情感倾向的任务定义为观点挖掘（**Opinion Mining**）或者情感分析（**Sentiment Analysis**）。到目前为止，研究人员已经对情感分析任务进行了大量的研究[1]。

音乐是网络上不断增多的数据类型之一。音乐的情感倾向性分析也逐渐受到越来越多的用户及研究人员的关注。本文研究音乐情感自动分类问题，即根据一首歌曲所包含的内容，借助计算机自动地为其赋予情感类别。借助音乐的情感分类，用户可以选择适合自己当前情绪的歌曲，或由系统自动向用户推荐某种情感的歌曲。可见音乐情感分类任务具有实际应用意义。

依据用于分析的特征的来源不同，音乐情感分类研究主要分为下列三个方向：基于音频的情感分类[2]、基于歌词的情感分类以及两者的结合。比较而言，音频较歌词具有更丰富的信息，但

* 本课题研究工作部分得到了国家自然科学基金项目（60873091，61073140）、中央高校基本科研业务费专项资金、高等学校博士学科点专项科研基金（20100042110031）、辽宁省自然科学基金项目（20102063）资助。

需要较高的计算代价，基于歌词的情感分类则具有速度上的优势。某些在线服务中，音乐情感分类系统的运行效率是必须考虑的重要因素之一。本文将重点研究基于歌词的音乐情感分类。

基于歌词的音乐情感分类方法可以分为两类：基于情感词典的方法[3]和基于自动分类模型的方法[4][5]。前者通常需要以高覆盖度的情感词典作为支撑，但是（以较少人工代价）构建高覆盖度的情感词典，即情感词典的自动扩展本身就是一件具有挑战性的任务[6]。基于自动分类模型的方法则将音乐情感分类问题看作文本分类问题，即将每首歌曲的歌词看作是一个文本。理论上任何分类模型，如最大熵（Maximum Entropy，ME）、支持向量机（Support Vector Machines，SVMs）都可以用来完成情感分类任务。但为了获得较高的分类性能，基于自动分类模型的方法需要人工标注一定规模的训练数据，而人工构建数据集与构建情感词典一样是一个耗时耗力的过程。如何在保证性能的同时减少有标注数据的使用是分类任务的一个重要问题。

我们发现，虽然音乐情感分类任务可以形式化为文本分类问题，但是该任务具有自己的鲜明特点。具体地说，目前音乐情感分类任务尚未有广为接受的类别体系。类别体系的缺失导致“标准”情感分类语料库的缺失，从而使得以离线（off-line）方式对所有音乐进行分类变得困难。实际上用户只希望获得自己感兴趣的歌曲，例如表达“快乐”情感的歌曲，而对不感兴趣的歌曲，用户并不关心其应该如何分类。我们可以将用户指定的（感兴趣的）音乐类别作为正类而将用户不感兴趣的其他所有音乐类别归为负类。这一处理方法与传统的文本分类方法的不同之处在于：首先，它回避了构建训练数据时必要的定义类别体系的过程，只需要用户指定感兴趣的歌曲并利用这些歌曲实例建立一个抽象的类别概念；其次，它使得部分监督学习方法（**partially supervised learning method**）[7][8]在音乐情感分析领域中的研究成为可行：只需要利用正类数据和大量不带标数据（其中可能包含正类样本）进行音乐情感分析，从而减少标注训练数据的代价。据我们所知，在情感分类任务，尤其是音乐情感分类任务上，部分监督学习方法的研究处于刚起步阶段。

本文拟采用两种部分监督学习方法进行音乐情感分类，即启发式的二元分割方法和 Spy-SVM 方法[8]。所有的实验在公开数据“5SONGS 歌词标注语料库”上进行。实验结果表明二元分割方法与 Spy-SVM 方法性能上具有可比性，但在正类数据和无标注数据的分布发生较大变化时，Spy-SVM 能够保持性能相对稳定。

2 问题描述

分类问题定义为：对样本集合 X 中任意元素 x，赋予其预定义类别集合 Y 中的某个类别 y。分类模型的训练过程可以形式化为寻找由 X 到 Y 的映射函数 $f(x)$的过程，其中 X 表示待分类对象集合，Y 表示可能的类别集合。本文假设任何 x 对应且只对应一个 y，即只研究单标签（single label）分类问题。不失一般性，我们进一步假设 $Y=\{0,1\}$，即二类分类问题。根据用于学习 $f(x)$的训练数据是否带标注，分类问题可以进一步分成有监督（**supervised**）和无监督（**unsupervised**）两种情况。前者要求$<X,Y>$作为训练数据，即对任意 x，其对应的 y 已知；后者训练数据中只有 X，而 Y 未知。无监督学习也被称作为聚类（**clustering**）问题。

本文研究的部分监督学习问题介于有监督学习和无监督学习之间，即部分训练数据的类别标记已知。具体地说，训练数据分为两部分：P 和 U，其中 P 表示正类样本集合，U 则表示类别未知的样本集合（无标注的样本的集合，简称为无标注集合）。需要强调的是，U 不一定只包含负类样本，相反 U 中同样可以包含正类样本。因此 U 也可以被称作混合数据集合。对音乐情感分类来说，P 表示用户感兴趣的歌曲（歌词）集合，由用户以较小的标注代价得到；U 表示任意歌曲（用户感兴趣或者不感兴趣的歌曲）的集合。本文研究如何从正类样本集合 P 和无标注样本集合 U 中学习分类函数 $f(x)$以便于计算机自动寻找用户感兴趣的歌曲集合。

有必要强调部分监督学习和半监督学习（**semi-supervised**）之间的区别。半监督学习同样是介于有监督学习和无监督学习之间的学习模式，采用（小规模）带标注数据和（大规模）无标注数据学习分类函数 $f(x)$。与部分监督学习相比，半监督学习要求带标注数据中出现的类别必须覆盖整个分类体系。以二类分类问题为例，半监督学习要求带标注数据中必须同时包含正类和负类数据，而部分监督学习的带标注数据中只有正类数据。

3 部分监督分类方法

部分监督分类方法在给定正类数据和无标注数据的前提下学习二类分类模型。如果无标注数据中的样本标记已知，则这一问题转化为常规的有监督学习。因此解决部分监督学习问题的最直接的思路是给无标注数据中的样本自动地赋予类别标记。本文采用两种方法解决这个问题：启发式的二元分割方法以及基于统计学习的 Spy-SVM 方法。

3.1 启发式的二元分割方法

二元分割方法基于下面所述的启发式规则：无标注样本中，距离正类样本距离近的倾向于被标为正类，而距离正类样本距离远的倾向于被标为负类。因此本文提出如下的两阶段方法：

- 根据正类数据求解正类数据集的中心向量，并且计算所有的正类样本距离该中心向量的最大距离 dmax。本文采用正类样本的平均值作为正类样本集的中心，同时采用 *Cosin* 相似度的倒数作为距离。计算公式如下所示：

$$P_{Center} = \frac{\sum_{i=1}^{N} d_i}{N} \tag{1}$$

$$dist = \frac{|d_i||d_j|}{d_i \times d_j} \tag{2}$$

- 将无标注数据中距离中心点距离小于 dmax 的样本标为正类，而将距离大于 dmax 的样本标为负类。距离中心点的距离约等于 dmax 的样本可以被认为是正类与负类的模糊样本。为了进一步提高上述方法的准确性，在实际实现时可以设置某个阈值，任何离中心点的距离与 dmax 的差值小于该阈值的样本可以被认为是模糊点而被忽略。

虽然上述方法简单而直接，4.2 节的实验却表明该方法可以充当一个较强的基准系统。

3.2 Spy-SVM 方法

部分监督学习方法通常可以描述为一个两阶段过程：1）识别无标注数据中可靠的负类样本（Reliable Negative，RN）；2）利用正类样本和可靠的负类样本构建分类模型，并利用该模型去自动识别无标注数据中的剩余样本（除 RN 以外的样本）的类别。在上述算法框架下，Liu 等[1]比较分析了多种学习方法，其中 Spy-SVM 取得了最佳性能，即第一阶段采用 Spy 方法而第二阶段采用 SVM 分类模型。本文将研究该方法在音乐情感分析任务中的应用。

3.2.1 识别可靠的负类样本

Spy 方法用于识别无标注数据中的可靠的样本。其基本思想是应用朴素贝叶斯模型（Naïve Bayes，NB）对无标注数据进行自动分类，然后在分类结果中选取最可靠的负类样本。具体过程如表 1 所示。

算法中另一个问题在于如何利用样本子集 S 确定阈值 t（步骤 6）。最直接的方法是选择 S 中样本的最大后验概率作为阈值，即选择阈值 t 使 S 中的样本不出现在 RN 中。但是实际情况中，S

当中可能存在噪音数据，其后验概率可能是接近 1 的一个数值。如果选择这样的数值作为阈值，将导致大量的可靠的负类样本被排除在 RN 之外。为了避免孤立点的影响，我们可以选择 t 使其允许 S 中的 l 个样本被归到 RN 当中。实际上，不同的 l 值（例如选择 5 或者 10）对结果影响不大[7]。因此本文工作将 l 设置成 5。

表 1　Spy 方法描述

输入：　正类数据集 P 和无标注数据集 U
输出：　可靠的负类样本集合 RN
1. RN=NULL
2. 从 P 中抽取子集 S
3. Ps=P-S，Us=U+S // 重新构建数据集合
4. Ps 中的样本当作正类，Us 中的样本当作负类，训练分类器 NB-C
5. 利用 NB-C 对 Us 进行自动标注，对 US 中的每个样本得到概率 $P(c=-1\|d)$，即样本属于负类的后验概率
6. 利用 S 确定某个阈值 t。Us 中负类概率大于 t 的样本将被看作可靠负类
7. 返回 RN

3.2.2　自动标注剩余样本

支持向量机是处理分类问题时使用最广泛的模型之一。其基本思想是在正类和负类数据之间寻找线性的分类超平面 $f(x)=WX+b$，使两类数据距离该超平面的距离最大。在测试阶段，应用训练得到的分类函数计算待测试样本的函数输出值并且根据此输出值的正负性确定其所属类别。关于支持向量机的更加详细的介绍可以参考[9]

利用支持向量机自动标注剩余样本（即 U-RN）的过程比较直接。首先在 P 和 RN 上训练支持向量机分类器，然后应用训练得到的分类器自动标注样本集合 U-RN。这样就可以得到 P 和 U 中所有样本的类别标记，其中 P 中由人工预先构建好且全部为正类样本。U 中样本的类别通过上述的两阶段过程自动得到。最终的分类模型将在 P 和（自动标注的）U 的并集上训练得到。

4　实验

4.1　数据及评价指标

本文实验采用“5SONGS”歌曲歌词标注语料库[4]（简称为 5SONGS）。该语料库对 2653 首中文流行歌曲进行了详细标注，包括歌名、歌手（及性别）、情感倾向等。其中情感倾向分为“轻松”（包含 1632 首歌曲）和“压抑”（包含 1021 首歌曲）两类。我们从中随机抽取了“轻松”和“压抑”各 1000 首（共 2000 首）歌曲构成本文实验的数据集。不失一般性，我们规定“轻松”类为正类，“压抑”类为负类。

我们首先对 2000 首中文歌曲进行分词，分词工具采用东北大学自然语言处理实验室的分词工具 CipSegSDK。去掉禁用词（共 611 个）之后，共包括 22528 个词类型（word type）。样本表示采用向量空间模型（Vector Space Model，VSM）[10]，即将歌曲表示为空间向量的形式。具体地说，本文采用词袋模型（Bag-of-Words），假设词与词之间相互独立并将每个词类型看作向量空间中的一维，表示样本的一个特征，以词类型在歌曲中的出现的频次（term frequency，TF）为特征值。这里我们考虑两种情况：1）采用全部 22528 个词作为特征；2）只选取 22528 个词当中的情感词作为特征。针对第二种情况，我们用一个包含 12343 词条的情感词典过滤情感词，并采用简单的方法处理了情感反转（negation）问题：将否定词与紧随其后的情感词合成一个新词。

本文实验随机抽取“轻松”和“压抑”各 200 个样本作为测试数据。正类数据和无标注数据从剩余的 1600 个样本中抽取，根据实验需要进行具体划分（参考 4.2 节的实验结果部分）。实验性

能的评价采用传统的召回率、正确率、F1 来评价分类结果。为了评价分类的整体性能，主要使用 Macro F1 进行评价。

4.2 实验结果

为了从 1600 个样本中构建正类数据集和无标注数据集，我们需要考虑三个不同变量的取值：正类数据的样本个数、无标注数据的样本个数，以及无标注数据中正类和负类数据的比例。

本文实验考虑正类样本个数为 100 和 400 两个不同取值，分别代表正类数据较少和较多两种不同情况。在固定正类样本个数的前提下，探究无标注数据的规模以及无标注数据中正类/负类比例的不同取值对性能的影响，实验结果如表 3 和表 4 所示。本文实验没有考虑上述两个变量的所有取值。我们假设无标注数据中的负类数据不少于正类数据，因此正负类数据的比例最大值取到 0.5。我们进一步假设无标注数据的规模不少于正类数据的规模，加之本文所采用的数据集的总规模的限制，实验中无标注数据的规模只取了 600，800 和 1000 三个值。表中标为 N/A 的结果表示该实验设置不成立。

表 2　以所有词为特征的实验结果（正类样本数为 100）

二元分割方法			
	600	800	1000
0.1	62.27	61.81	N/A
0.2	64.81	63.42	65.39
0.3	63.38	62.62	68.58
0.4	62.59	61.76	66.79
0.5	60.73	61.95	67.19
Spy-SVM 方法			
	600	800	1000
0.1	68.05	69.08	N/A
0.2	65.15	68.49	63.70
0.3	66.78	67.06	60.96
0.4	65.99	65.79	62.83
0.5	66.55	67.01	64.47

表 3　以所有词为特征的实验结果（正类样本数为 400）

二元分割方法			
	600	800	1000
0.1	68.28	65.31	N/A
0.2	66.47	69.09	68.54
0.3	64.85	65.99	65.31
0.4	63.21	65.15	64.96
0.5	61.74	60.27	N/A
Spy-SVM 方法			
	600	800	1000
0.1	66.67	68.99	N/A
0.2	66.89	67.34	66.78
0.3	66.78	66.78	67.18
0.4	66.78	66.53	67.80
0.5	66.34	66.89	N/A

从上面的实验结果中我们可以发现如下规律。第一，二元分割方法可以被用作较强的基准系统，尤其是当无标注数据中的正类数据较少（譬如比例为 0.1 或者 0.2）的情况下，其性能与 Spy-SVM 的性能可比较，在某些情况下甚至能超过 Spy-SVM 方法。第二，随着无标注数据中正负类数据的比例趋向 0.5，二元分割方法的性能相应地下降。其原因在于随着无标注数据中的正类与负类数据趋向于平衡，对数据的自动标注将逐渐变得困难，简单方法不容易取得好的效果。第三，当正类数据的规模较小时（譬如表 3 中使用 100 个样本），二元分割方法较 Spy-SVM 方法缺少稳定性：即正类样本的规模对前者的影响更大一些。

本文进行的另外一个实验是采用情感词作为特征。实验目的是为了比较分析采用情感词为特征对两种方法的性能影响。这里，我们只考虑正类样本个数为 400 的情况。实验结果如表 5 所示。

表 4　以情感词为特征的实验结果（正类样本数为 400）

二元分割方法			
	600	800	1000
0.1	64.62	63.96	N/A
0.2	65.19	64.77	66.13
0.3	63.77	62.24	64.01
0.4	60.80	61.80	61.95
0.5	58.72	62.05	N/A
Spy-SVM 方法			
	600	800	1000
0.1	62.23	61.73	N/A
0.2	63.83	57.68	54.75
0.3	63.39	59.57	50.88
0.4	62.34	56.50	45.63
0.5	60.04	54.08	N/A

从表中结果可以看到：当采用情感词作为特征时，二元分割方法和 Spy-SVM 方法的性能都严重下降。主要原因在于情感词典的覆盖率不高。实际上，在应用 4.1 节中介绍的情感词典时，只有 2670 个情感词出现在本文实验所用的 2000 首中文歌曲中。另一方面，从表 5 的实验结果中可以看到，虽然二元分割方法同样有所下降，但是其下降的幅度要远低于 Spy-SVM 方法。

5　相关工作

20 世纪 90 年代，音乐情感分析作为音频信号处理领域中的任务之一首次被提出。初期的研究工作完全基于对音频信号的处理，最常用特征包括音色、强度和节奏等，所采用的方法大多局限于一些经典的机器学习方法。情感类别体系方面，Thayer 于 1989 年提出的二维情感模型被广泛采用[11]。这一模型认为情感由压力（valence）和能量（arousal）两个因素决定，可分为如下四类：满足（contentment）、沮丧（depression）、愉快（exuberance）和忧虑（anxious）。

Lu 等[12]在分类过程上有所创新，提出了分层情感分析过程。不同于先前工作中将歌曲直接归到四类中某一类的做法，他们分两阶段来实现音乐情感分析任务。第一阶段仅利用音乐强度特征得出其能量水平，第二阶段则利用音色和节奏特征得出压力水平，结合两阶段结果最终确定歌曲所属的情感类别。实验中，上述特征（音乐强度、音色、节奏）均从歌曲音频数据中抽取。

Chen 等[13]较早在音乐情感分析任务引入歌词信息。他们改进了 Lu 的两阶段方法，具体地说，Chen 等在第二阶段利用歌词文本来确定歌曲的压力水平。实现过程中，选用情感词来产生歌词向量，然后采用 K 最近邻居（K-Nearest Neighbor，KNN）方法确定歌词的压力水平。然而，实验结

果表明只使用歌词文本无法达到一般文本情感分类的性能。Xia 等[4]分析了 Chen 等[13]工作中性能不佳的原因：如歌词中包含大量对情感贡献度很小的词，名词和动词在表达情感上存在歧义以及否定词和修饰词对歌词情感有特殊的贡献等。针对这些问题，Xia 等在向量空间模型基础上，提出了情感向量空间模型（Sentiment Vector Space Model，S-VSM），使其可以更好的处理基于歌词的情感分类任务。基于情感向量空间模型的支持向量机分类器的训练要求同时有正类和负类样本。Shu[3]同样着眼于基于歌词的音乐情感分类问题。与 Xia 等[4]不同，Shu 没有采用分类模型进行情感分析，而是采用了基于情感词典的方法，完成了一套歌词的情感挖掘系统。实验中，利用 SentiWordNet 来抽取歌词中的情感词，并以此为依据对主观倾向进行了处理。

本文同样研究利用分类方法进行基于歌词的音乐情感分类。与 Xia 等[4]的工作不同，本文的方法并不基于完全有监督学习。相反，本文工作研究如何只利用正类数据和无标注数据进行音乐情感分类。进行音乐情感分类时只有正类数据而缺少相应的负类数据更加符合任务的实际情况。对部分监督的音乐情感分类的研究可以认为是本文最重要的贡献。

6 结论及未来工作

本文针对基于歌词的音乐情感分类的具体特点，研究了应用部分监督学习方法进行情感分析的可行性。我们采用了两种不同思想的学习方法：启发式的二元分割方法和基于统计的 Spy-SVM 方法，并且在公开数据集上进行了比较实验。实验结果表明两种方法具有可比较性能，但是后者在正类数据和无标注数据的分布发生较大变化时能够保持性能相对稳定。

在将来工作中，我们首先将构建更加符合实际情况的数据集。本文实验中，我们只研究了二类情感分类问题。但在实际情况中，音乐的情感分类体系有必要做进一步的改进。另外，本文实验分别研究了以所有词作特征和以情感词为特征两种情况。在将来工作中，有必要进一步研究如何既要利用所有词又要突出情感词的重要性。最后，我们将比较研究其他的部分监督方法。

参 考 文 献

[1] Bo P, Lee L. Opinion Mining and sentiment analysis. Now Publishers Inc., Hanover, MA.

[2] 蒋旻隽, 周昌乐, 黄志刚. 音乐情感的自动识别. 厦门大学学报, 2011, 49(6): 798-803.

[3] Shu J. Opinion mining for song lyrics. Master thesis of Norwegian University.

[4] Xia Y, Wang L, Wong K, Xu M. Sentiment vector space model for lyric-based song sentiment classification.In Proc of ACL/HLT 2008. 2008: 133-136.

[5] 夏云庆, 杨莹, 张鹏洲, 刘宇飞. 基于情感向量空间模型的歌词情感分析. 中文信息学报, 2010, 24(1): 99-103.

[6] Qiu G, Liu B, Bu J, Chen C. Opinion word expansion and target extraction through double propagation. Computational linguistics, 2011, 37(1): 9-27.

[7] Liu B, Lee W, Yu P, Li X. Partially supervised classification of text documents. In Proc. of ICML 2002.

[8] Liu B, Dai Y, Li X, Lee W, Yu P. Building text classifiers using positive and unlabeled examples. In Proc. of ICDM 2003. 2003: 179-188.

[9] Cristianini N, Shawe-Taylor J. An Introduction to support vector machines. Cambridge University Press, Cambridge, UK, 2000.

[10] Sebastiani F. Machine learning in automated text categorization. ACM computing survey, 34(1): 1-47.

[11] Thayer R. The biopsychology of mood and arousal. Oxford University Press, 1990.

[12] Lu L, Liu D, Zhang H. Automatic mood detection and tracking of music audio signals, IEEE Transactions on Audio, Speech & Language Processing, 2006: 14(1), 5-18.

[13] Chen R, Xu Z, Zhang Z, Luo F, Content based music emotion analysis and recognition, in Proc. of International Workshop on Computer Music and Audio Technology, 2006, 68-75.