利用 AdaBoost-SVM 集成算法和语块信息的韵律短语识别

钱揖丽[1,2]，冯志茹[1]

(1.山西大学 计算机与信息技术学院，山西 太原 030006；
2.山西大学 计算智能与中文信息处理教育部重点实验室，山西 太原 030006)

摘要：该文提出一种基于汉语语块结构并利用 AdaBoost-SVM 集成学习算法的汉语韵律短语识别方法。首先，对语料进行自动分词、词性标注和初语块标注，然后利用基于结合紧密度获取的规则对初语块进行归并，得到最终的语块结构。其次，基于语块结构并利用 AdaBoost-SVM 集成算法，构建汉语韵律短语识别模型。同时，该文利用多种算法分别构建了利用语块信息和不利用语块的多个模型，对比实验结果表明，表示浅层句法信息的语块能够在韵律短语识别中做出积极有效的贡献；利用 AdaBoos-SVM 集成算法实现的模型性能更佳。

关键词：汉语语块；AdaBoost-SVM；韵律短语；识别

中图分类号：TP391　　文献标识码：A

Recognition of Chinese prosodic phrase based on AdaBoost-SVM algorithm and chunk information

QIAN Yili[1,2] , FENG Zhiru[1]

(1. School of Computer & Information Technology, Shanxi University, Taiyuan, Shanxi 030006, China;
2. Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan, Shanxi 030006, China)

Abstract: In this paper, a new method for Chinese prosodic phrase recognition based on Chunk and AdaBoost-SVM algorithm is proposed. Firstly, the initial chunks are marked on the corpus of automatic word segmentation and part of speech tagging, and then they are merged using the rules based on the closeness between initial Chunks. Secondly, based on the block structure and the AdaBoost-SVM integrated algorithm, a Chinese prosodic phrase recognition model is constructed. Meanwhile this paper utilized various algorithms to build different models which use or not use Chunk information, the comparative experimental results show that the shallow syntactic information Chunk has made a positive and effective contribution to Chinese prosodic phrase recognition, and the performance of the AdaBoost-SVM model is better.

Key words: Chinese chunk; AdaBoost-SVM; prosodic phrase; recognition

1 引言

语音合成是制造语音的技术。它涉及声学、语言学、数字信号处理、计算机科学等多个学科技术，是中文信息处理领域的一项前沿技术。目前机器合成的语音与人讲的话之间还有明显的差距，其自然度还有待进一步的提高。韵律的差距是影响语音自然度的重要因素之一，合成的语音单调枯燥，且在节奏、轻重、停顿等方面的处理不当使其听起来非常别扭。充分掌握和运用自然语言的韵律信息，是提高合成语音自然度的关键。

人在说话时往往会按照话语表达的核心、语义和发音的生理机能等，自然地在话语中添加必要的停歇。停歇的位置、时长等对于语义表达、语流的生动性和自然度等有着很大的影响。

基金项目：国家自然科学基金资助项目（61175067），国家自然科学青年基金资助项目（61005053，61100138），山西省科技基础条件平台建设项目（2015091001-0102），山西省青年科技研究基金资助项目（2012021012-1），山西省回国留学人员科研资助项目（2013-022）。

作者简介：钱揖丽（1977—），女，博士，副教授，硕士生导师，主要研究方向为自然语言处理；冯志茹（1988—），女，硕士，主要研究方向为自然语言处理。

语音上的停歇与文本的韵律结构紧密相关。目前比较公认的是将韵律结构从下到上分为三个级别，即：韵律词、韵律短语和语调短语。在韵律结构边界会出现长短不同的停歇，韵律层次越高，停歇的时间就越长。由于韵律词往往与语法词相对应，而语调短语则通常是一个完整的分句，因此，韵律短语是人们研究的重点。针对韵律短语识别研究，已有的工作有基于语言学规则的方法[1]，这类方法复用度低且很容易受到人为因素的限制；有基于统计的方法，如基于二叉树[2,3]、马尔科夫模型[4]、最大熵模型[5]、决策树[6]等等，这些方法使用的特征大多为词、词性等词法特征，或者使用依赖人工标注的语法特征；还有规则和统计相结合的方法等，这些工作使得韵律结构划分问题取得了一定的进展。

通过对大量语料的分析可知，韵律结构和句法结构之间存在着一定的联系。韵律结构是以句法结构为基础的，在句法上不能够出现停顿的地方（如词内音节之间），韵律上也不允许出现停顿；而在句法上的高层结构之间、特别是标点符号出现的地方，韵律上一定会出现停顿[7]。但是由于汉语句子和句法结构的复杂性和灵活多变性，往往存在着一定的嵌套关系，且句法分析器的生成较为复杂，对随机的句子进行分析得到的结果还不甚理想。为了降低句法分析难度，语块在 CoNLL-2000 被提出。语块分析能够对句法分析起到很好的中介作用，并为后续的句法分析提供依据。另外，通过观察和统计发现，人们在朗读或说话的时候往往会自然地将句子切分成一定长度的语块流，语块的切分还会把句法上相关的词进行整合，对韵律短语的识别起到积极作用。所以，本文在汉语语块识别的基础上，提出将语块结构这种非递归嵌套的浅层句法结构应用于韵律短语的识别。

另外，要实现韵律短语的自动识别，就需要构造一个具有较高泛化能力的高精度学习机。但由于寻找一种较强的分类算法用于韵律短语识别较为困难，基于强、弱学习算法的等价性问题，利用集成学习方法能够使多个准确率略高于随机猜测的弱分类器进行加权融合，形成一个强学习算法，达到比强分类器更好的分类效果。所以，本文使用 AdaBoost 集成学习算法，用 SVM 方法训练生成多个基分类器，再将多个基分类器用加权投票的方法集成，形成一个新的强分类器完成对韵律短语的预测。多项对比实验结果显示，基于语块结构并利用 AdaBoost-SVM 集成学习算法构建的模型性能更佳。

2 AdaBoost-SVM 集成算法

实现韵律短语的自动识别，需要构造出一个具有较高泛化能力的高精度学习机。而领域知识和学习数据集本身及其分布对泛化能力的制约较大。传统的数理统计与模式识别的方法需要尽可能精确地找到预测的规则，故构造精度高的学习机很难；而集成学习的思想大大改变了以往研究的思路。

2.1 Boosting 算法

集成学习是一种机器学习方法，对于分类问题其主要思想是：使用一些分类效率只需略高于随机猜测的弱分类学习算法学习生成多个不同的基分类学习机，然后将多个基分类学习机组合成强分类学习机[8]，这个新形成的分类学习机具有较强的泛化能力。

从 Schapire[9]证明一个强分类学习机可以被多个弱分类学习机通过某些方法得到开始，Boosting 算法便得以出现。此后，Freund[10]提出了一种更有效的 Boost-by-majority 算法。但是，这两种算法在解决实际问题时就会有许多问题产生。在使用弱分类学习算法前，必须先知道其最差正确率。1997 年，Schapire 和 Freund[11]提出的 AdaBoost 算法解决了这一问题，且其算法效率与 Boosting-by-majority 相当，而且极易应用于实际问题中。之后，又提出了可以控制投票机制的 AdaBoost.M1、AdaBoost.M2 和 AdaBoost.R 算法。

2.2 基于 AdaBoost 的 SVM 集成算法

虽然 AdaBoost 方法自适应能力强且实现简单，可以提高任意一种弱分类器的分类精度。但却特别容易受到噪声数据的影响[13]。这是由于 AdaBoost 算法强调分类错误的数据更为重

要，所以在每次训练结束后会对训练错误的数据赋予更大的权重。这种现象在迭代多次后更为明显，因此导致最终的集成分类器效果下降。所以，为了保证和提高算法效果，本文在使用 AdaBoost 算法训练时对数据权重的赋值加入了一个参数进行调节。

AdaBoost-SVM 集成算法的主要思想是：选用 SVM 作为基分类器，再用 AdaBoost 算法进行迭代生成 T 个子 SVM 分类器，在迭代的过程中为保证每次生成的子 SVM 分类器之间的差异性，对每个子分类器输入大小相同但内容包含前面分类器给出的错分样本的子训练集。这样能使得算法更关注错分样本，并不像 AdaBoost 算法使用的是原始训练数据集。最后将这些子 SVM 分类器按照加权投票的方法组合生成最终的集成分类器。

本文中的 AdaBoost-SVM 算法描述为：

输人：训练样本集 $\mathrm{L}=\{(\mathrm{x}_1,y_1),(\mathrm{x}_2,y_2),\ldots,(\mathrm{x}_i,y_i),\ldots,(\mathrm{x}_N,y_N)\}$，其中 $x_i\in R^n$，$y_i=\{1,-1\}$，迭代次数 T，基分类算法 SVM。

输出：用于韵律短语识别的集成分类器 $H(x)$。

初始化训练集样本权重 $\varphi_1(x_i)=1/N$，$i=1,2,\ldots,N$，迭代次数 t=1。

For t=1,…，T：

① 根据分布 φ_t，从原始训练样本集 L 中有放回的随机抽取 M（$M<N$）个训练样本，得到新的训练集 $L_t=\{(x_i^{(t)},y_i^{(t)})\}_{i=1}^{N}$ $(L_t\in L)$。

② 在得到的训练集 L_t 上利用 SVM 分类算法训练生成一个基分类器 $h_t: x\to\{-1,1\}$，并计算分类器在整个训练集 L 上的分类误差：

$$\varepsilon_t=\sum_{i=1}^{N} I(h_t(x_i)\neq y_i)\varphi_t(x_i)$$

③ 若：$\varepsilon_t>0.5$，则令 $T=t-1$ 退出循环；若：$\varepsilon_t(x_i)=0$，则令 $\varepsilon_t(x_i)=10^{-10}$。

④ 令 $\alpha_t=\frac{1}{2}\ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$

⑤ 更新样本权重：$\varphi_{t+1}(x_i)=\frac{\varphi_t(x_i)}{Z_t}\times\begin{cases}e^{-\alpha_t/\beta} & if \quad h_t(x_i)=y_i\\ e^{\alpha_t/\beta} & if \quad h_t(x_i)\neq y_i\end{cases}$

$$=\frac{\varphi_t(x_i)\exp(-(\alpha_t(x_i)/\beta)y_i h_t(x_i))}{Z_t}$$

其中，Z_t 为归一化因子。

End For

输出最终集成分类器：

$$H(x)=sign\left(\sum_{t=1}^{T}\alpha_t h_t(x)\right)$$

3 语块结构及其处理

3.1 语块的分类

语块是指介于词汇和句子之间的模式化的短语。语块的识别和分析属于浅层句法分析的范畴。目前中文语块的定义主要有两大类：一类是从进行了句法标记的句法树库中直接抽取出句法树的非终结点作为语块[12,13]，另一类是根据具体的中文语法现象对句子进行分析，构造出具有独立性和完整性的语块定义[14]。

本文建立的语块属于第二类，共分 8 种类型，分别是：名词语块（NC）、动词语块（VC）、形容词语块（JC）、副词语块（AC）、介词语块（PC）、连词语块（CC）、数量词语块（QC）、方位语块（LC）。它们具备两个特征：一是语块之间无重叠，句子中的任一词都只能属于一个语块，且语块之间无嵌套，若有歧义则按照最长匹配的原则进行划分[10]；二是句子中的每个词都必须进行语块标注，且语块内部不再进行细分。

3.2 语块的加工处理

3.2.1 初始语块的标注

初始语块的标注方法为：首先根据汉语的句法特征总结归纳出各类语块的具体特征，如：助词“的”往往依附于其前面的成分，数词和量词往往是一个整体等；然后利用正则文法，设置不同的子文法限制，各子文法结合有限状态自动机嵌套递归对文本中的句子进行正则匹配，从而完成初始语块的标注。

例如，经过分词和词性标注的句子为：

我们/r 从/p 实际/n 出发/v，大力/d 种植/v 石榴/n，摸索/v 出/v 了/u 一/m 条/q 治理/v 水土/n 和/c 治穷/v 致富/v 相/d 结合/v 的/u 成功/a 之/u 路/n

上述例句的初始语块标注结果为：

【NC 我们/r】【PC 从/p】【NC 实际/n】【VC 出发/v】,【AC 大力/d】【VC 种植/v】【NC 石榴/n】,【VC 摸索/v】【VC 出/v 了/u】【QC 一/m 条/q】【VC 治理/v】【NC 水土/n】【CC 和/c】【VC 治穷/v】【VC 致富/v】【AC 相/d】【VC 结合/v 的/u】【JC 成功/a 之/u】【NC 路/n】

其中，位于每个“【】”之间的部分就是语块。

3.2.2 基于结合紧密度的初始语块归并

将各类语块间的结合紧密度定义为：

$$P(\overline{B} \mid C_i C_j) = \frac{C(\overline{B} \mid C_i C_j)}{C(C_i C_j)} (1 \le i \le 8,\ 1 \le j \le 8)$$

其中：$C_i C_j$ 表示两个相邻语块分别为 C_i 类和 C_j 类；$\overline{B}$ 表示两个相邻语块间未出现韵律短语边界；$C(C_i C_j)$ 表示 C_i 类、C_j 类语块相邻共现的总次数；$C(\overline{B} \mid C_i C_j)$ 表示 C_i 类、C_j 类语块相邻共现且其分界处不作为韵律短语边界的次数。

两类语块的结合紧密度 $P(\overline{B} \mid C_i C_j)$ 的值越大，表明它们结合得越紧密，即当此两类语块邻接时，其中间不出现韵律短语边界的概率也越大。通过挖掘初始语块边界与韵律短语边界之间的潜在关系，依据任意两类语块的结合紧密度计算结果，总结归纳出以下 8 条初始语块的归并规则[15]。

（1） VC+NC→VC；
（2） JC+NC/VC→JC；
（3） QC+NC/JC→QC；
（4） CC+NC/VC/JC →CC；
（5） xC+LC→LC，xC 表示任意语块类型；
（6） PC+yC→PC，yC 表示除介词语块 PC 外的其余任意语块类型；
（7） AC+zC→AC，zC 表示除连词语块 CC 外的其余任意语块类型；
（8） mC + xC →mC ，mC 为以“的”结尾的任意语块类型。

利用上述规则，将初始语块进行归并，生成最终的语块结构。这样，就能筛选剔除结合

紧密的语块间的边界，缩小了待预测边界的范围，更有利于韵律结构的分析标注。

例如，3.2.1 中例句经过初始语块归并后的结果为：

【NC 我们/r】【PC 从/p 实际/n】【VC 出发/v】,【AC 大力/d 种植/v】【石榴/n】,【VC 摸索/v】【VC 出/v 了/u】【QC 一/m 条/q】【VC 治理/v 水土/n】【CC 和/c 治穷/v】【VC 致富/v】【AC 相/d 结合/v 的/u】【JC 成功/a 之/u 路/n】

在初始句子中，共有 22 个词间边界，它们都是潜在的韵律短语边界；经过语块标注和归并后，最终待预测的边界缩减至 12 个，共有 10 个结合紧密的词间边界被首先剔除。

4 利用 AdaBoost-SVM 和语块信息的韵律短语识别

4.1 模型特征及处理

考虑到 SVM 具有良好的泛化能力，且本文使用 SVM 主要用于 AdaBoost 算法的基分类算法，也就是说，只要 SVM 分类效果好于随机猜测的结果就行，所以基分类器选取的特征为：当前语块内容 c 、当前语块的类型 t 、当前语块所含词的个数 $wlen$ 、当前语块所含字的个数 $clen$ 。特征向量表示为：

$$x=(c，\ t，\ wlen，\ clen)$$

另外，为了进行对比实验，本文也实现了不利用语块信息的分离器，选用的特征为：当前词的内容 w ，当前词的词性 p ，当前词的长度 l 。特征向量表示为：

$$x=(w，\ p，\ l)$$

使用 LibSVM 工具包作为 SVM 分类器进行实验，由于 SVM 只能处理数值型的特征数据，而本文采用的特征：语块内容、语块类型、词、词性均为文本型数据，所以本文首先采用构建词袋和词性袋等方法，对数据集中的文本数据进行数值化处理，使其适用于 SVM 分类器的数据处理过程。

4.2 AdaBoost-SVM 算法实现

在利用 2.2 中描述的算法进行韵律短语识别时，令 $y_i=1$ 表示当前边界是韵律短语边界，$y_i=-1$ 表示当前边界不是韵律短语边界；在利用语块信息时， x_i 表示不同类型的语块；不使用语块信息时， x_i 则表示语法词。

为了使算法更精确，引入参数 β 来降低被正确分类个体上赋予权重减少的量，或被错误分类个体上赋予权重增加的量。β 的值不宜过大，随着 β 的增大算法的误差有上升趋势 [16]，所以本文将 β 设定为 5。

使用 AdaBoost 算法每生成一个子 SVM 分类器，该分类器就会在整个训练集上测试其分类效果，根据测试结果更新训练集上样本的权重，若错分则增加权重，若分类正确则降低权重，并由分类结果计算出每个分类器的权重 α_t 。若分类错误的样本较多，说明分类器的分类效果不好， α_t 的值较小；若分类错误的样本较少，则说明分类器的分类效果好， α_t 的值较大。为了保证 AdaBoost 做种生成的集成分类器的效果，往往更多地集成比较好的分类算法，所以以 α_t 作为各个基分类器 h_t 的权重。

在进行韵律短语边界预测时，对于一个测试语料集 L，输入未标注韵律结构的句子 x，

训练过程中生成的 T 个子 SVM 分类器h_t，会生成 T 个韵律短语标注结果。若$h_t(x)=y_i$（i=1，…，N），代表第 t 个子 SVM 分类器分类正确，则对子 SVM 分类器h_t投一票。最后，根据投票结果，将得票最多的分类作为 AdaBoost-SVM 对输入句子 x 的集成分类结果。

5 实验结果及分析

实验语料是来源于 1998 年《人民日报》的 3200 个句子，经过分词、词性标注以及人工韵律结构标注，平均每句含有 34.61 个词，10.36 个韵律短语。随机抽取 2800 句作为训练集，400 句用于开放测试。

5.1 语块标注与归并的影响

基于不同加工粒度的实验语料，即颗粒大小为“词”的词标注语料和以“语块”为单位的语块标注语料，分别统计和计算自然边界（词边界或语块边界）与韵律短语边界的对应关系，得到结果如下表 1 所示。

表 1 词/语块边界与韵律短语边界

语料	韵律边界占自然边界的总比率	韵律边界在词/语块间	韵律边界在词/语块内
词标注语料	19.55%	19.55%	0.00%
语块标注语料	54.69%	50.46%	4.23%

从表 1 可以看出：一方面，实验语料经过分词后，韵律边界仅占所有词边界的 19.55%；而进行语块标注和归并后，由于大量词边界被包含到语块内部自然剔除，韵律边界所占比例大幅提高到 54.69%，语块的引入剔除了大量的噪声边界，带来了积极的影响。另一方面，语块也会带来一些负面影响，有 4.23%的韵律短语边界会因被归并在语块内部而丢失，这类情况大多是多个名词、或多个动词同时出现导致的，可利用如长度约束机制等来解决。

5.2 分类器个数的影响

在生成 AdaBoost-SVM 的过程中，本文将子训练集大小设定为 N*3/4（N 为总训练集的大小）并进行迭代，直到达到训练次数或分类误差$\varepsilon_t > 0.5$为止。不同分类器个数下 AdaBoost-SVM 的韵律短语识别结果如表 2 所示。

表 2 不同分类器个数下 AdaBoost-SVM 的识别结果比较

分类器个数	正确率	召回率	F 值
5	0.5876	0.8729	0.7024
10	0.6294	0.9028	0.7417
15	0.6431	0.9162	0.7557
20	0.6829	0.9193	0.7837
25	0.7344	0.9231	0.8180
30	0.7948	0.9269	0.8558
35	0.8324	0.9376	0.8818
40	0.8341	0.9438	0.8856

从表 2 中可以看出，随着分类器个数的增加，AdaBoost-SVM 的分类效果也越来越好。基分类器个数为 5 时韵律短语识别的 F 值为 70.24%；当基分类器数增加到 40 个时，其 F 值提高到 88.56%，提升了 18.32%。但是，基分类器个数的增加也会增加时间开销，导致训练时间过长。

5.3 不同方法的实验结果比较与分析

基于词标注和语块标注两类语料，分别采用 CRFs、SVM、AdaBoost-SVM 方法构建实现了 6 个相应的韵律短语识别模型。各个模型的实验结果对比情况如下表 3 所示。

表 3 不同模型的实验结果比较

方法	正确率	召回率	F 值
CRFs	0.8094	0.7339	0.7699
CRFs+语块	0.8966	0.8342	0.8640
SVM	0.5235	0.7656	0.6218
SVM+语块	0.8629	0.5963	0.7053
Adaboost-SVM	0.6628	0.9098	0.7669
Adaboost-SVM+语块	0.9438	0.8341	0.8856

利用语块前后 CRFs、SVM、Adaboost-SVM 这 3 类模型韵律短语识别 F 值的比较如图 1 所示，同样利用语块时 SVM 算法与 Adaboost-SVM 算法的性能比较如图 2 所示。

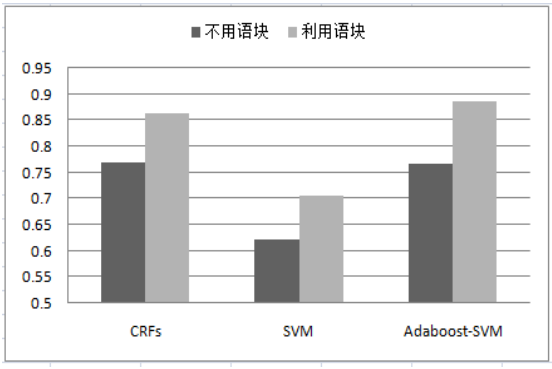

图 1 利用语块前后 3 类模型 F 值的比较

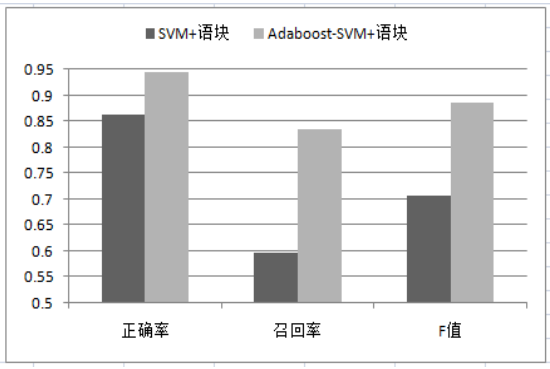

图 2 SVM 算法与 Adaboost-SVM 算法性能比较

从以上图表中可以看出：（1）对于上述 3 种方法，引入并利用语块信息之后，模型的总体性能都得到了明显的提升，CRFs 模型韵律短语识别 F 值提高了 9.41%，SVM 方法提高了 8.35%，AdaBoost-SVM 方法提高了 11.87%；（2）利用语块信息的模型，韵律短语识别的正确率都大大提高，这是通过语块标注将大量的噪声边界自然剔除的结果；（3）同样基于语块结构，与 SVM 方法相比，Adaboost-SVM 集成算法获得了更好的效果，召回率、正确率都得到了大幅的提高，其 F 值提高了约 18%。

综上所述，反应浅层句法信息的语块结构能够被应用于汉语韵律结构的分析，并做出积极有效的贡献；而且，集成学习方法的识别效果高于其他强分类器的识别效果。通过语块结构的标注和归并，实现了对语料中结合紧密语法词的整合，从而准确缩小了待识别边界的范围。另外，由于语块的粒度较大，选用语块特征相当于缩小了训练空间上的大小，模型训练的时间开销也会明显缩减，尤其在使用集成学习算法时，表现更为明显。

6 结论

正确划分句子的韵律结构对于提高机器合成语音的自然度具有重要的意义和作用。本文基于语块结构并利用 AdaBoost-SVM 算法实现了一个汉语韵律短语识别模型。首先，对语料进行自动分词、词性标注、初语块标注和归并处理，建立以“语块”为单位的语料。然后，基于上述语块标注语料并利用 AdaBoost-SVM 集成算法训练生成最终的分类器用于汉语韵律短语的识别。本文利用 CRFs、SVM、AdaBoost-SVM 共 3 种算法分别构建了利用语块信

息和不利用语块的 6 个韵律短语识别模型，并将测试结果进行了对比。实验结果表明，不论是上述哪种方法，引入并利用语块信息之后，其韵律短语识别效果都能得到明显的提升，反应浅层句法信息的语块能够做出积极有效的贡献。同时，利用 AdaBoos-SVM 集成算法实现的模型性能更佳，其韵律短语识别的 F 值为 88.56%，比 SVM 模型提高了 18%左右。

由于集成学习算法只要求基分类器的效果大于随机猜测的即可，故本文中 SVM 算法选用的特征仅限于当前词的内容、词性和长度，没有考虑和利用上下文语境信息。而且，在利用 LibSVM 对数据进行训练时，耗时较长，导致 AdaBoost-SVM 算法的时间复杂性仍然较高。另外，利用正则匹配的方法进行语块的识别，不可避免地会使部分韵律短语边界包含在语块结构的内部。今后的研究中会针对以上问题进行深入的研究与改进。

参考文献：

[1] 曹剑芬. 基于语法信息的汉语韵律结构预测[J]. 中文信息学报, 2003,17(3):41-46.

[2] 荀恩东,钱揖丽,郭庆,等. 应用二叉树剪枝识别韵律短语边界[J]. 中文信息学报, 2006, 20(3):1-5,28.

[3] 钱揖丽,荀恩东. 基于标点信息和统计语言模型的语音停顿预测[J]. 模式识别与人工智能, 2008,21(4):541-545.

[4] Taylor P, Black A W. Assigning phrase breaks from part-of-speech sequences[J]. Computer Speech & Language, 1998, 12(2): 99-117.

[5] 李剑锋, 胡国平, 王仁华. 基于最大熵模型的韵律短语边界预测[J]. 中文信息学报, 2004, 18(5): 56-63.

[6] 王永鑫, 蔡莲红. 语法信息与韵律结构的分析与预测[J]. 中文信息学报, 2010 (1):65-70.

[7] 曹剑芬. 汉语韵律切分的语音学和语言学线索[C]//新世纪的现代语音学—第五届全国现代语音学学术会议论文集, 北京: 清华大学出版社, 2001: 176-179.

[8] 李想. Boosting 分类算法的应用与研究[D]. 兰州: 兰州交通大学,2012.

[9] Robert E. Schapire. The strength of weak learnability[J]. Machine Learning,1990,52:197-227.

[10] Y. Freund. Boosting a Weak Learning Algorithm by Majority[J]. Information and Computation,1995,121(2):256-285.

[11] Yoav Freund,Robert E Schapire. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting[J]. Journal of Computer and System Sciences,1997,55(1):119-139.

[12] 周强, 李玉梅. 汉语块分析评测任务设计[J]. 中文信息学报, 2010, 24 (1): 123-128.

[13] 周强, 詹卫东, 任海波. 构建大规模的汉语语块库[C]//自然语言理解与机器翻译—全国第六届计算语言学联合学术会议论文集. 北京: 清华大学出版社, 2001: 102-107.

[14] 李素建, 刘群. 汉语组块的定义和获取[C]//语言计算与基于内容的文本处理—全国第七届计算语言学联合学术会议论文集, 北京: 清华大学出版社, 2003:110-115.

[15] 钱揖丽, 冯志茹. 基于语块和条件随机场(CRFs)的韵律短语识别[J]. 中文信息学报, 2014, 05:32-38.

[16] 张春霞. 集成学习中有关算法的研究[D]. 西安: 西安交通大学,2010.