文本分类中用于协同的特征集分割

张博锋 苏金树
(国防科学技术大学计算机学院 长沙 410073)

摘 要 用于文本分类领域的协同训练往往需要特征集的一个自然独立分割,但对大多数语料而言这种分割都很难获取或不存在。给出了特征子集间在类别下条件独立性的定量描述,并在此意义下提出了局部特征集分割的策略,以及两种分别基于样本聚类和图分块的以独立性为前提的特征集分割算法。在两个语料库上的分类实验证明:在该方法所获得的特征集分割下,协同训练方法能有效利用未标注样本提高分类器的综合效果,从而有效扩展了协同训练的可用性。

关键词 文本分类,协同训练,特征集分割,局部自适应聚类,图分块
中图法分类号 TP181 **文献标识码** A

Feature Set Splitting for Co-training in Text Categorization

ZHANG Bo-feng SU Jin-shu
(School of Computer, National University of Defense Technology, Changsha 410073, China)

Abstract In the area of text categorization, co-training method always needs a naturally independent split of the feature set which is hardly to obtain or rarely exists for most of the corpus. This paper presented the quantitative description of the conditional independence of feature subsets given the class, and suggested a strategy for splitting feature set locally in this sense. Two algorithms respectively based on sample clustering and graph partitioning for feature set splitting in the precondition of independence were proposed. Experimentation on two corpuses shows that, in the feature divisions produced by our methods, the combined effectiveness of the co-trained classifiers is improved by applying the unlabeled samples. As a result, the applicability of the co-training method is extended.

Keywords Text categorization, Co-training, Feature set splitting, Locally adaptive clustering, Graph partitioning

协同训练(co-training)是非常有效的半监督学习方法,但必须将特征集自然分割成两个可以充足构造分类器的子集合,并且要求它们之间在给定的类别下条件相互独立[1-4]。然而在大多数实际应用中,很难甚至不可能给出特征集的自然分割(split)。文献[2]发现对某些问题,即使将特征集随机分割成两个子集,仍然可以通过协同训练获得分类效果的改善,因此可考虑给出特征集相关性尽量小的分割。目前,还未见到其他关于满足协同训练要求特征集分割的相关深入研究。

本文定义了在给定类别下条件独立性的定量标准,提出局部特征集分割的策略,给出分别基于样本局部自适应聚类及特征关联图分块的两种尽量保持两个子集条件独立性的特征集分割方法,并比较了它们用于协同训练的效果。通过特征集的分割,协同训练通过未标注样本大大改善了以 Naïve Bayes[2-4]为底层(underlying)学习方法的分类器效果,适用性得到增强。

1 协同训练

协同训练是一种通过未标注样本加强较弱的底层分类器的方法,它从两个不同的视图(view)对标注和未标注样本提供的信息进行独立学习,并建立相应的工作于不同视图的分类器,其中每个视图对应一个不同特征子集上的文本表示。从学习过程的开始,通过少数已标注样本的不同视图分别初始化两个底层分类器,在训练过程的每一轮循环中,每个底层分类器都在自己的视图中审视每个未标注样本,并从未标注集合中挑选出一个(或多个)进行标注并加入到标注集合中,所挑出的样本是当前底层分类器认为自己最可信的标注结果。这样,在每一轮循环中底层分类器都会在自己的视图上从不断增长的标注集中重新构造并得到加强,此过程不断重复,直到满足终止条件[2]。本文中采用 Naïve Bayes 作为底层方法,协同训练结束后,最终的分类器是两个底层分类器的组合,例如可将底层分类器输出的关于相同类别的后验概率相乘并规格化[3]。

其中为了便于下文讨论,作以下符号约定和说明:设标签(类别)集合为 $C=\{c_1,c_2,\cdots,c_{|C|}\}$,已标注样本集合为 $L=\{(d_1,l_1),(d_2,l_2),\cdots,(d_{|L|},l_{|L|})\}$,其中 $l_i\in C$ 称为文本 d_i 的标签(或类别),未标注样本集为 $U=\{d_{|L|+1},d_{|L|+2},\cdots,d_{|L|+|U|}\}$,$D=L\cup U$ 为样本集合,经过特征选择所获得的特征集合为 $V=\{v_1,v_2,\cdots,v_{|V|}\}$。称 $\langle V_1,V_2\rangle$ 是某特征子集 F

到稿日期:2008-04-28 本文受国家自然科学基金(90604006)资助。
张博锋(1978—),男,博士,助理研究员,主要研究方向为信息安全与数据挖掘,E-mail:bfzhang@nudt.edu.cn;苏金树(1962—),男,博士,教授,博士生导师,主要研究方向为网络通信与信息安全。

$\subseteq V$ 的一个分割(split)，如果有 $V_1, V_2 \subseteq F$ 均非空，满足 $V_1 \cap V_2 = \emptyset, V_1 \cup V_2 = F$。当文本 d 或样本子集 S 仅考虑在特征集子 F 上的表示时，分别记作 $d|F$ 和 $S|F$，称为在 F 上的视图或投影。

文献[3]中给出了协同训练的理论依据，即如果特征集的分割在给定类别的条件下相互独立，并且目标函数在 PAC (probably approximately correct)模型下是可学习的，则底层分类器可以通过未标注样本得到加强。例如图 1 所示的两类 Email 分类[4]，其中每一整行表示文本在整个特征集上的完整表示，深色的部分表示特征出现的频率比较高，每个文本可以被分割成在不同特征子集上的投影。首先，设定邮件头与邮件内容中出现的相同的词是不同的特征，作者认为同类的头特征与内容中的特征便有了很自然的本类下的条件独立性(如图中 c_1 的头特征与 c_1 的内容特征)；其次，不同类别间的特征子集两两之间在任何给定类下具有条件独立性(如图中 c_2 的头特征与 c_1 的内容特征)，因而图中的特征集自然分割，即目标函数(分类器)h_1 与 h_2 的视图在任何给定类下相互具有独立性。同时，目标函数 h_1 和 h_2 的可学习性一般可以通过特征集的冗余性得到满足，即尽量保证 h_1 的视图和 h_2 的视图中的特征足够多。

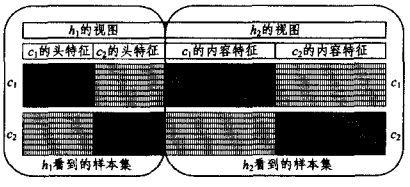

图 1　Email 分类问题的特征集分割

然而在实际应用中，大部分非结构化文本的分类问题很难在数据中给出特征集关于类别的显式独立分割，因此协同训练的应用仅局限于少数结构化的网页(文本内容特征和指入链接的 anchor words 分别看成两个分割[3])或 Email(邮件头和内容的特征分别看成两个分割[4])等问题的分类。然而在文献[2]的实验中发现，在不验证特征集分割独立性的情况下，协同训练依然会带来分类器效果的提升，即使在特征集随机分割的情况下，也将使用 co-EM 算法的 Naïve Bayes 分类器的错误率降低 10%左右，但改善效果弱于自然分割。因此，在没有特征集自然分割的情形下，给出特征集某种相关性尽量小的分割也许是比较好的提高协同训练适用性的方法。

2　以独立性为前提的局部特征集分割

实际问题中使得两个特征子集在某个类条件下完全不相关的特征集分割是极难获得或根本不存在的，因此可寻找在每个类的条件下都能使两个特征子集在某种意义下保持尽量强独立性的分割，称为在此类下以独立性为前提的分割(SPI, split in the precondition of independence)。可用特征之间的同现(co-occurrence)关系刻画特征集间的独立性；然后在局部特征选择的基础上，对每个类的局部特征子集进行在本类条件下以独立性为前提的分割。

2.1　特征子集独立性

即使是一个自然的分割，两个特征集还是存在某种特定的关联，例如两个分割中总有一些特征有紧密的关于样本的同现关系，因此只能要求特征集之间的某种关联保持很低的水平。可以从特征之间的同现率出发定义两个特征之间的相关性，并进一步定义两个特征集之间的相关性。

定义 1　设 $t, s \in V$ 是两个特征，定义它们在给定类别 $c \in C$ 下的条件相关系数，简称为在 c 下的相关系数为：

$$r_c(t,s)=\begin{cases}\dfrac{N_c(t,s)^2}{N_c(t)\cdot N_c(s)}, & N_c(t), N_c(s)>0\\ 0, & \text{otherwise}\end{cases} \tag{1}$$

其中 $N_c(t,s)$ 表示在类别 c 中同时包含特征 t 和 s 的文本数，$N_c(v)$ 表示类别 c 中包含某特征 $v \in V$ 的文本数。不难得到以下性质：

性质 1　$0 \leqslant r_c(t,s) \leqslant 1$。$r_c(t,s)=1$ 当且仅当 t 和 s 在 c 的样本中总是同时出现，$r_c(t,s)=0$ 当且仅当它们没有在 c 中的任何样本中同时出现。

性质 2　$r_c(\cdot,\cdot)=1$ 是特征集 V 上的一个等价关系。

性质 3　若 $r_c(t,s)=1$，则对任何 $v \in V$，$r_c(t,v)=r_c(v,s)$。

直观上，当 $r_c(t,s)$ 接近 1 时，特征 t 与 s 具有在 c 下较大的相关性，而 $r_c(t,s)$ 接近 0 时，它们在类别 c 下的相互独立性更强。

定义 2　设 $T, S \subseteq V$ 是两个非空特征子集，定义它们在给定类别 $c \in C$ 下的条件相关系数，简称为在 c 下的相关系数为：

$$R_c(T,S)=\frac{\sum_{t\in T, s\in S} r_c(t,s)}{|T|\cdot|S|} \tag{2}$$

可以证明，特征集间条件相关系数有下面重要的性质：

性质 4　设非空特征子集 $T_i, S_j \subseteq V$ $(0<i\leqslant m, 0<j\leqslant n)$ 中任意两个互不相交，$0<\varepsilon<1$，若 $R_c(T_i,S_j) \leqslant \varepsilon$ $(0<i\leqslant m, 0<j\leqslant n)$，则 $R_c(\bigcup_{0<i\leqslant m} T_i, \bigcup_{0<j\leqslant n} S_j) \leqslant \varepsilon$，其中 $m, n>1$ 是常数。

这种性质说明两组不相交的特征子集分别合并时，组间的条件相关性总体上将不会扩大。

我们将特征集间在类别 c 下的相关系数 $R_c(T,S)$ 定义为两个特征集 T 和 S 各自所含特征在 c 下两两相关系数的平均值，从而给出了两个特征集在给定类别下条件独立性(简称为在 c 下的独立性)的一种定量刻画。当 $R_c(T,S)$ 较大时，说明 T 与 S 中有很多特征之间具有很强的相关性，成为 T 和 S 两者在 c 下的独立性弱的重要依据，反之则给出了两者的独立性较强的一种依据。

2.2　局部分割

SPI 的目标是寻找能使两个特征子集在本文所定义的相关性意义下对每个类保持尽量强条件独立性的分割。首先，对特征集作以下基本假设：

(a)V 是通过局部(local)特征选择方法[1]获得的，本文根据待选特征关于每个类的重要性排名(rank)，分别为每个类选出具有代表性的局部特征集 $V(c_i)$，$i=1,2,\cdots,|C|$，则 $V=\bigcup_i V(c_i)$；

(b)所有局部特征集 $V(c_i)$ 互不相交；

(c)每个类的局部特征集 $V(c_i)$ 两两之间在每个类下具有对较强条件独立性。

事实上在进行局部特征选择时，可限定每个待选特征最多属于一个局部特征集来保证假设(a)和(b)，而(c)的满足也是一种自然属性，例如直观上对几个表达不同主题的类别，它们代表性关键词(项)的同现关系一般情况下会比较弱。

根据以上假设，如图 2 所示，只要获得每个局部特征集 $V(c_i)$ 的分割 $\langle V_1(c_i), V_2(c_i)\rangle$，称之为一个局部分割，并使得这两个子集之间在 c_i 下的相关性尽可能小，即获得在 c_i 下 $V(c_i)$ 的一个 SPI。如果令 $V_1=\bigcup_i V_1(c_i)$，$V_2=\bigcup_i V_2(c_i)$，由假设(c)和性质 4，$\langle V_1, V_2\rangle$ 就是 V 在任何类下相关性都较小的 SPI。其中，假设(a)保证了分割的完整性，假设(b)保证了子集没有交集，而假设(c)和性质 4 则保证了两个子集在每个类下的条件独立性。局部特征子集分割很大程度上保证了全局特征集分割的两个子集 V_1 和 V_2 中都分别含有一定数量的任何类别的代表性特征，从而满足对每个类别可学习性的要求。

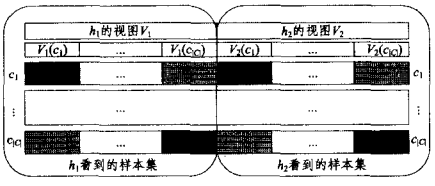

图 2　局部特征集的分割策略

3　特征集分割方法

我们采用基于文本聚类和基于边加权无向图分块的两种特征集分割方法求解类别 $c\in C$ 的局部特征子集 $V(c)$ 在本类下的局部 SPI 问题。前者是通过从 c 类样本不同的主题角度给出 $V(c)$ 的一种自然分割，而后者则是以相关性为约束，对 $V(c)$ 在 c 下的关联图进行分块的强行分割。为便于讨论不妨设 $V(c)=\{v_1, v_2, \cdots, v_{|V(c)|}\}$。

3.1　基于文本聚类的特征集分割

如果通过某种聚类方法将属于 c 的文本分为两组(cluster)，则不同的组就可能代表了本类不同侧重角度的主题，这些组间文本的生成具有一定的独立性，考虑到特征对每个组重要性的不同，就可以给出 $V(c)$ 的一个自然分割，这与相关文献中自然分割的思想类似[2-4]。当获取了两个具有主题差异的组后，可以通过特征与两个组的关联性进行特征集的分割。这里将类别 c 看作由其标注样本的构成的样本子集，令：

$c'=c|V(c)=\{\boldsymbol{x}=(x_1, x_2, \cdots, x_{|V(c)|})|\boldsymbol{x}$ 是类别 c 标注样本在 $V(c)$ 上的向量表示$\}$

采用 LAC(locally adaptive clustering)聚类方法[5]将 c 的标注样本即 c' 中的向量分为两组，其重要的特点是两间点的距离以加权 L_2 范数计算，即对向量 $\boldsymbol{x}=(x_1, x_2, \cdots, x_{|V(c)|})$ 及权重向量 $\boldsymbol{w}=(w_1, w_2, \cdots, w_{|V(c)|})$，$\boldsymbol{x}$ 的以 $\boldsymbol{w}$ 为权重的加权 L_2 范数为：

$$L_2(\boldsymbol{w},\boldsymbol{x})=\left(\sum_{i=1}^{|V(c)|} w_i x_i^2\right)^{\frac{1}{2}} \tag{3}$$

其中 $\sum_{i=1}^{|V(c)|} w_i=1$，每个 $w_i>0$ 刻画了特征 v_i 所在的第 i 维与本组所有向量关联的紧密程度，我们最终利用它们的数值对特征集进行分割。分割的方法为：

$$V_1(c)=\{v_i|w_{1i}>w_{2i}, 0<i\leqslant|V(c)|\},$$
$$V_2(c)=V_2(c)\backslash V_2(c) \tag{4}$$

其中 $\boldsymbol{w}_j=(w_{j1}, w_{j2}, \cdots, w_{j|V(c)|})(j=1,2)$ 是每个组的权重系数，最终可获得 $V(c)$ 的一个局部 SPI$\langle V_1(c), V_2(c)\rangle$。

3.2　基于图分块的特征集分割

设 $\langle T,S\rangle$ 是 $V(c)$ 的一个分割，在希望得到的两个子集大小基本相等的情况下，有：

$$R_c(T,S)=\frac{\sum_{t\in T,s\in S} r_c(t,s)}{|T|\cdot|S|}\approx\frac{4}{|V(c)|^2}\sum_{t\in T,s\in S} r_c(t,s)$$
$$\propto\sum_{t\in T,s\in S} r_c(t,s) \tag{5}$$

因此，要获得 $V(c)$ 的 SPI，可以使式(5)最后一项尽量小。

定义以特征集 $V(c)$ 为顶点的边加权无向图为 $G(c)=(V(c),E,w)$，其中 $E\subseteq V(c)\times V(c)$ 为边集合，$\langle t,s\rangle\in E$ 当且仅当 $t,s\in V(c)$ 且 $r_c(t,s)>0$，$w: E\to\mathbf{R}^+$ 是边上的权重，$w(\langle t,s\rangle)=w(t,s)=r_c(t,s)$，$\langle t,s\rangle\in E$。设 G 的割集合为：

$Cut(G(c))=\{\langle T,S\rangle|T,S\subseteq V$ 非空，$T\cap S=\emptyset$ 且 $T\cup S=V(c)\}$

其中集合中的每个元素称为 $G(c)$ 的一个割(cut)。

图 $G(c)$ 是 $V(c)$ 在 c 下的关联图，可以看出，根据式(5)找到一个 $V(c)$ 的 SPI 等同于图 $G(c)$ 的平衡最小割分块(balanced min-cut partitioning)问题[6]，即：

$$\begin{aligned}&\text{minimize}\quad f(V_1(c),V_2(c))\\&\text{subject to}\quad \langle V_1(c),V_2(c)\rangle\in Cut(G(c));\\&\qquad\qquad |V_i(c)|\approx|V(c)|/2, i=1,2\end{aligned} \tag{6}$$

其中 $f: 2^{V(c)}\times 2^{V(c)}\to\mathbf{R}^+$，$f(T,S)=\sum_{\langle t,s\rangle\in(T\times S)\cap E} w(t,s)$，即跨越两个分块的边的权重之和。

图的平衡最小割分块的精确求解是一个 NP 完全问题，在本文中采用了 MeTiS 软件包[6]，它实现了一个多层(multilayer)的图分块方法，具有很高分块质量和计算效率。

4　实验与比较

4.1　在 WebKB 上的实验结果

WebKB 语料库[7]包括了从美国 4 所大学的计算机系网站上收集而来的 1051 个网页的相关内容，每个样本包括了指向本网页链接的 anchor text 以及本网页本身的内容。文献[2,3]实验中的任务都是从中识别关于学术课程的网页(占总数的 22%)。WebKB 本身包含了两个链接和内容两个文本的视图，这种特征集的自然分割(NS)非常有利于文本分类的协同训练。本实验应用基于 LAC 和图分块(GP)的方法对特征集进行分割，并与 NS 下协同训练的效果进行比较。

随机挑选 263 篇(占总数 25%)样本作为测试样本，在其余的训练样本中，随机挑选 17 篇正例和 61 篇反例(分别占各自类别训练样本总数 10%)作为标注样本，剩余的为未标注样本。试验中，链接中与内容中出现的相同词作为不同的特征，从标注样本所有内容(包括链接)中剔除常见的停词(stop-words)以及出现在文本内容中小于 3 次的词。内容特征进行以 IG 为标准的局部特征选择[1]；对于链接内容虽然不作特征选择，但是根据在 IG 函数将每个词归入某个类别的局部特征集中，以便于进行局部 SPI。

表 1 是特征选择及分割后，特征集及相关子集间独立性的信息，其中 $R_{avg}(V_1,V_2)$ 是指全局特征集 V 分割的子集 V_1 和 V_2 在所有类别下相关系数的均值。可以看到在本文定义的相关性意义下，图分割的两个子集在 3 种方法中具有最强的独立性，而自然分割的独立性也较强，LAC 方法的独立性较差。同时，$V(-)$ 分割的独立性要好于 $V(+)$，这与 $V(-)$ 中的样本数和特征数较多有关，反例中有多个自然主题也是一个可能原因，后者还可能致使 LAC 方法在 $V(-)$ 上的分割独立性与其他方法最为接近。

表 1　分割后的子集独立性

	No Split	NS	LAC	GP
$\|V(+)\|$	459	459	459	459
$\|V_1(+)\|,\|V_2(+)\|$	N/A	109,350	172,287	228,229
$R+(V_1(+),V_2(+))$	N/A	0.272	0.451	0.210
$\|V(-)\|$	734	734	734	734

$\|V_1(-)\|,\|V_2(-)\|$	N/A	334,400	501,233	366,368
$R-(V_1(-),V_2(-))$	N/A	0.136	0.148	0.093
$R_{avg}(V_1,V_2)$	N/A	0.116	0.129	0.079

表 2 以错误率(error rate)的百分比数值给出了在几种特征集分割上协同训练所获得的效果，同时给出了 Naïve Bayes 方法在全部 788 个标注的训练样本上和给定的 78 个标注样本上的监督学习结果。其中错误率定义为对所有测试样本做出的错误判断数占所有测试样本的百分比。h_1 和 h_2 分别是两个底层分类器，而 $h=h_1\&h_2$ 是根据底层分类器给出的概率乘积所得的最终决策。自然分割 NS 中，链接特征集上训练的分类器为 h_1，对于监督学习，h 给出的就是监督训练最终效果。在所有分割下，协同训练都通过未标注样本提升了系统仅在少量标注样本上监督学习的分类效果，GS 的结果与 NS 的结果非常接近，是效果最好的两种方法，可能的原因是 GP 获得的两个底层分类器效果比较均衡而 NS 的基于内容特征的分类器的效果出色，又因为各种两个特征子集具有较强的独立性，从而使得两个底层分类器具有互补性，因此 h 获得较大的性能提升。

表 2 WebKB 上的错误率(%)

Splitting Method	|L|	|U|	h_1	h_2	h
No split on D	788	0	N/A	N/A	5.7
NS	78	710	13.9	**9.1**	**7.2**
LAC	78	710	13.2	14.5	8.3
GP	78	710	13.4	**11.8**	7.5
No split on L	78	0	N/A	N/A	16.9

对 NS，GP 和 LAC 分割分别所得的子集间进行多次随机的成份交换，要求这些交换只能在每个局部特征子集内进行，即不改变局部分割策略。经过交换，得到一些近似于随机的特征集分割，子集的相关系数平均值与协同训练的性能关系如图 3 所示。可以看出：GP 分割的独立性最强，而 NS 分割对分类效果的提升最好，并且在定义 1 和定义 2 的意义下，特征子集间的独立性与协同训练的效果显现出正相关性，一定程度上验证了从独立性角度寻找特征集分割的合理性。

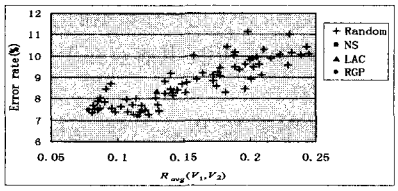

图 3 协同训练效果与独立性间的关系

4.2 在 TanCorp12 上的实验结果

TanCorp12 是中科院计算所的相关研究者发布的中文文本分类语料 TanCorpV1.0 中经过加工的单层数据集[8]，包含了 12 个类别的样本共 14150 个。随机选取 4150 个作为测试样本，在剩余的 1 万个样本中从每个类随机挑 5%，共 503 个样本作为标注样本，其余作为未标注样本。去掉停词、单字词及低频词后，采用局部 IG 标准为每个类选出 200～250 个特征(适当增加样本较多类别的特征)。

TanCorp12 没有公认的特征集的自然独立分割，因此比较在局部策略下 LAC，GP 以及特征集的一个随机等分所获得的特征集分割下，协同训练的提升效果，同时还给出 Naïve Bayes 方法在采用全部训练样本以及仅标注样本作为训练样本的监督学习效果，如表 3 所列。在表中，两种方法的提升效果都好于随机分割，仍然是 GP 方法获得了最强的独立性，但 LAC 方法获得了最好的效果提升，我们认为这与语料库的结构有关。事实上，从 TanCorpV1.0 的说明中我们发现，每个 TanCorp12 单层类都还有很多不同的主题，因此与 WebKB 相比，在 TanCorp12 上 LAC 能够获得更加"自然"的独立性，这说明在每个类别的样本可以被分组的情况下，LAC 也许是一种比较好的特征分割方法。

表 3 TanCorp12 上的错误率

Splitting Method	|L|	|U|	R_{avg}	h_1	h_2	h
No split on D	10000	0	N/A	N/A	N/A	**7.3**
LAC	503	9497	0.219	**15.0**	15.3	**11.9**
GP	503	9497	**0.173**	**13.4**	16.1	12.2
Random	503	9497	0.251	16.3	15.2	14.7
No split on L	503	0	N/A	N/A	N/A	24.1

结束语 协同训练是一种有效的半监督学习方法，但对特征集的自然独立分割的要求限制了其应用范围。本文给出了特征及特征集间在类别条件下相关系数的定义，定量描述了特征集间在给定类别下的条件独立性，并以此为基础提出以独立性为前提的分割(SPI)。我们通过局部特征集分割的框架来满足分割的每个子集上底层算法的可学习性，而基于 LAC 样本聚类和基于图分块的特征分割方法则很大程度上保证子集间的独立性。在两个语料库上的分类实验证明我们的方法所获得的特征集分割下，协同训练能有效利用未标注样本提高分类器的综合效果，从而进一步扩展了协同训练的适用性。

参 考 文 献

[1] 苏金树，张博锋，徐昕. 基于机器学习的文本分类技术研究进展. 2006,17(9):1848-1859

[2] Nigam K, Ghani R. Analyzing the applicability and effectiveness of co-training// Proceedings of CIKM-00, 9th ACM International Conference on Information and Knowledge Management. McLean, US: ACM Press, 2000: 86-93

[3] Blum A, Mitchell T. Combining labeled and unlabeled data with co-training// Proceedings of COLT-98, 11th Conference on Computational Learning Theory. Madison, Wisconsin: ACM, 1998: 92-100

[4] Kiritchenko S, Matwin S. Email classification with co-training// Proceedings of the 2001 conference of the Centre for Advanced Studies on Collaborative research. Toronto, Ontario, Canada: IBM Press. 2001: 8-17

[5] Kang N, Domeniconi C, Barbará D. Categorization and keyword identification of unlabeled documents// Proceedings of 5th IEEE International Conference on Data Mining (ICDM-05). Houston, Texas: IEEE Computer Society, 2005: 677-680

[6] Karypis G, Kumar V. METIS-serial graph partitioning and fill-reducing matrix ordering. http://glaros. dtc. umn. edu/ gk-home/metis/metis/overview. 2007

[7] CMU world wide knowledge base (Web->KB) project. http://www. cs. cmu. edu/afs/cs/project/theo-11/www/wwkb/. 2001

[8] Tan S-B, Wang Y-F. A Chinese corpus for text categorization - TanCorpV1. 0. http://www. searchforum. org. cn/tansongbo/corpus1. php. 2005