一种新的支持向量机主动学习策略及其在文本分类中的应用

刘 宏[1] 屠铁清[2] 黄上腾[1]

(上海交通大学电子信息学院计算机科学与工程系 200030)[1] (澳大利亚 Deakin 大学计算与数学系)[2]

A New Support Vector Machines Active Learning Approach and its Application in Text Classification

LIU Hong[1] TU Zhi-Qing[2] HUANG Shang-Teng[1]

(Department of Computer Science and Engineering, College of Electron Information of Shanghai Jiao Tong University, Shanghai 20030)[1]

(Department of Computing and Mathematics, Deakin University of Australia)[2]

Abstract There are two well-known characteristics about text classification. One is that the dimension of the sample space is very high, while the number of examples available usually is very small. The other is that the example vectors are sparse. Meanwhile, we find existing support vector machines active learning approaches are subject to the influence of outliers. Based on these observations, this paper presents a new hybrid active learning approach. In this approach, to select the unlabelled example(s) to query, the learner takes into account both sparseness and high-dimension characteristics of examples as well as its uncertainty about the examples' categorization. This way, the active learner needs less labeled examples, but still can get a good generalization performance more quickly than competing methods. Our empirical results indicate that this new approach is effective.

Keywords Active learning, Text classification, Orthogonalization, Support vector machines

1. 引言

随着电子文档数量的剧增，文本自动分类[1]已成为信息化建设中的一项重要任务。支持向量机[2]作为一种比较新的机器学习方法，已在文本分类领域取得了良好的应用效果；大量的实验表明，在大多数情况下，支持向量机方法优于其它(如，朴素贝叶斯分类、决策树、神经网络等)方法。

以往，人们应用支持向量机进行文本分类通常采用监督学习模型。在该模型中，为了组织训练样本，人们需要标识大量的样本(例如，在二元分类问题中，人们为正样本加标记1，而为负样本加标记－1)，但是，标识样本是一项耗时且代价昂贵的工作，特别是对于文本分类来说更是如此。

最近，人们开始研究如何将支持向量机方法引入基于池(pool)的主动学习模型[3]。这一学习模型是由 Lewis 和 Gale 在1994年提出的。在该模型中，由所有未加标记的样本组成一个样本池，而学习器(如支持向量机)在学习过程中可以访问这个池，并可以询问池中样本的实际标记。主动学习器从很少的标记样本开始学习，然后按照某种启发式规则选择另外的很少的一些未标记样本，询问它们的实际标记，将它们加入到原来的训练样本集合中，通过对这个新的训练样本集合的学习，更新原来的知识，接着再选择一些未标记的样本，继续学习。按照这种学习方式，学习器通过对少量标记样本的学习就能得到很好的推广能力。

到目前为止，支持向量机在主动学习过程中采用的启发式规则基本上都是基于如何减小版本空间(Version Space)的[4,5]；这种策略计算起来比较简单(当评价每个未标号样本时，只需要一次内积运算)，但是，它有一个明显的缺点，即学习器容易受到孤立点的影响。为了克服这个缺点，有人提出，在选择要询问的未标记样本时，优先考虑能在最大程度上减小期望风险估计的样本[6]。这种策略虽然在一定程度上降低了孤立点的影响，使得学习器询问更少的样本就能达到同样的推广能力，但是，该策略是建立在大量计算的基础之上的，在实际应用中难以采纳。

在文本分类中，人们通常采用向量空间模型(Vector Space Model)来表示文本，即将每个文本表示成一个特征向量，人们已提出多种文本特征选取方法[1]，这里不再赘述。我们看到，文本分类问题具有两个典型的特征。第一，样本空间的维数特别高，并常常大于人们拥有的训练样本数；即使经过降维处理，最后得到的样本向量仍然高达几千，甚至上万维，比如，在我们的实验中，经过预处理后，每个样本向量的维数是8992；特别地，在主动学习模型中，人们企图使用比监督模型更少的标记样本来训练学习器，这就使得样本维数大而样本数量少这一矛盾更加突出。第二，样本向量特别稀疏，即，每个文本对应的向量中只有很少的项是非零项；特别地，在主动学习模型中，学习器通常从很少(比如几个)的标记样本开始学习，这样，尤其是在学习过程的前期阶段，已标记样本向量所张成的空间只占了整个样本空间的一少部分。

基于以上认识，我们提出一个混合的支持向量机主动学习策略，使学习器在选择要询问的未标记样本时能综合更多的样本信息，从而进一步减少所需的标记样本数目，并在一定程度上降低孤立点带来的影响。

2. 支持向量机主动学习模型

支持向量机(SVM)是由 Vapnik 提出的一种基于结构风险最小化原理的机器学习方法。在最简单的情形中，线性 SVM 通过学习得到一个超平面，该超平面以最大分类间隔将

刘 宏 博士生，主要研究方向为机器学习、文本分类。屠铁清 博士生，主要研究方向为模式识别、数字图像处理。黄上腾 教授，博士生导师，主要研究方向为面向对象数据库、机器学习。

正样本集合与负样本集合分离开，此处的间隔(Margin)是指超平面与距离它最近的正样本和负样本之间的距离，如图1所示。

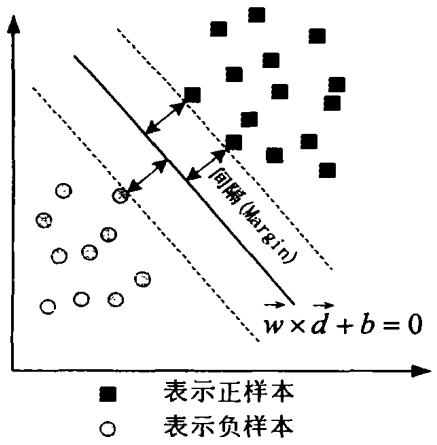

图1 二维情形中的最优超平面和间隔(Maregin)

对于线性可分的情形，分类间隔最大化可以表示为如下的优化问题：

$$\underset{\vec{w},b}{Minimize}\quad \frac{1}{2}\|\vec{w}\|^2$$

$$s.t.\quad c_j(\vec{w}\cdot\vec{d}_j+b)\geqslant 1\quad j=1,\cdots,n$$

其中，$\vec{d}_j$ 是第 j 个训练样本，c_j 是 $\vec{d}_j$ 的实际标记。在主动学习模型中，支持向量机通过询问获得 $\vec{d}_j$ 实际标记。对于线性不可分的情形，我们可以引入非负松弛因子，以便在最少错分样本和最大分类间隔之间做出权衡。

从数值计算的角度考虑，上述寻优问题解决起来比较困难。通常的做法是通过引入 Lagrange 乘子将原问题转化为其对偶问题：

$$Minimize\ \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j c_i c_j\,\vec{d}_i\cdot\vec{d}_j-\sum_{i=1}^{n}\alpha_i$$

$$s.t.\ \sum_{i=1}^{n}\alpha_i c_i=0\ \forall\ i:\alpha_i\geqslant 0$$

关于支持向量机的详细介绍及最新进展请参见文[2,9]。

在学习过程中，学习器依据什么策略从未标记样本池中选择下一个要询问的样本是支持向量机主动学习中的一个关键问题。如引言中所述，现有的做法大都是考虑如何快速地减小版本空间，即支持向量机选择它对其类别最不确定的一个样本，换句话说，就是选择距离当前得到的超平面距离最近的一个样本，该距离的计算方法如式 $|\vec{w}\cdot\vec{d}_i+b|$，记为 p(如果样本向量都是单位向量，则有 $p\in[0,1]$)。这种启发式规则简单直观，但是，如果把样本类别的不确定性作为选择的唯一依据，则学习器很容易受到孤立点的影响。具体地说，学习器在选择样本点的时候无法区分孤立点和非孤立点，如果学习器在学习过程中选择了孤立点，那么在接下来的参数调整过程中，它将要去除的版本空间区域其实多是一些对学习器推广能力没有实际影响的部分，从而使学习过程需要更多的标记样本。

3. 新的支持向量机主动学习策略

前文提到，文本特征向量比较稀疏，而维数又特别高，所以，基于主动学习模型的学习器在其学习过程中(特别是在初期阶段)使用的标记样本向量张成的空间只占整个样本空间的一小部分。由此，在选择要询问的样本时，除了未标记样本的类别不确定程度信息外，未标记样本与已标记样本向量张成的空间垂直程度也是一类十分重要的信息；未标记样本与已标记样本向量张成的空间的垂直程度越大，说明已标记样本向量组没有覆盖到的未标记样本向量中的非零维越多；因而，通过询问这样的样本，学习器就能得到已标记样本未能覆盖到的那些维的信息，从而指导自身的学习过程。

基于以上认识，我们考虑设计一种合理的方法，使得学习器在选择下一个要询问的未标记样本时不仅仅考虑未标记样本的类别不确定性，而且考虑未标记样本向量与已标记样本向量张成的空间的垂直程度。我们期望，通过综合考虑两方面的样本信息，学习器能做出更趋合理的选择，与采用现有的启发式规则相比，能进一步减少所需的标记样本数目，并在一定程度上降低孤立点带来的影响，这里，我们称这一新的策略为混合策略。

为了应用这一策略，我们需要解决两个问题，一是如何度量未标记样本向量与已标记样本向量组张成的空间的垂直程度；二是，通过什么途径来表达这一混合策略。下面具体说明我们的做法。

1. 如何度量未标记样本向量同已标记样本向量组张成的空间的垂直程度

首先需要由已标记样本向量组求得一个正交向量组，使这两个向量组张成的空间等价。为了叙述方便，这里设已标记样本向量都是 n 维列向量，并记由已标记样本向量组构成的矩阵为 $A_{n\times m}$(n 是样本向量的维数，m 是已标记样本向量的个数)，而记求得的与已标记样本向量组等价的正交向量组构成的矩阵为 $S_{n\times l}$(n 仍是样本向量的维数，l 是正交向量组中向量的个数，且一定有 $l\leqslant m$)。由于我们不知道已标记样本向量组中的这些向量是否线性相关，我们有下述几种选择：

第一种方法是应用 Gram－Schimdt 正交化过程处理 $A_{n\times m}$ 中的列向量；在处理过程中，如果碰到某个向量是前面处理过的向量的线性组合，那么，经过正交化处理后，它一定是个各项都为0的向量，我们将其去除，最后得到的所有列向量组成的矩阵就是 $S_{n\times l}$。

第二种方法是先从已标记样本向量组 A 中抽取一个线性不相关的向量组，记为 A'，然后对 A' 进行 QR 正交三角分解[8]，即

$$A'=QR \tag{1}$$

其中，Q 是我们所求的 $S_{n\times l}$。

第三种方法是直接对向量组 A 进行奇异值分解(Singular Value Decomposition)[8]，即

$$A=UDV' \tag{2}$$

其中，U 是我们所求的 $S_{n\times l}$。

在比较了上述三种方法后，我们选用了第一种方法，因为这种方法比较直观，实现起来也比较简单。

在求得与已标记样本向量组等价的正交向量组后，通过计算未标记样本向量与这个正交向量组的夹角的余弦大小，我们就能度量未标记样本向量同已标记样本向量组张成的空间的垂直程度了。为此，我们先要对该正交向量组，即 $S_{n\times l}$ 中的列向量进行标准化处理；若未标记样本向量不是单位向量，我们还要对所有的未标记样本向量进行标准化处理。记任一个长度为1的未标记样本向量为 n 维列向量 $\vec{d}$，并记它在由 $S_{n\times l}$ 中的列向量张成的空间中的投影为 n 维列向量 $\vec{y}$，则有

$$\vec{y}=S(S'\vec{d}) \tag{3}$$

因为 $\vec{y}$ 和 $\vec{d}$ 都是单位向量，所以二者夹角的余弦就是 $|\vec{y}\cdot\vec{d}|$，记为 q（显然，$q\in[0,1]$）。q 的值越小，说明 $\vec{d}$ 与 U 张成的空间的垂直程度越大；反之，则越小。

第二个问题是在询问下一个未标记样本时如何将前述两种启发式规则结合在一起。我们考虑如式(4)所示的加权和（记为 w），

$$w=ap+(1-a)q \tag{4}$$

其中，$a\in(0,1)$。

在学习过程中，学习器在选择下一个要询问的未标记样本之前先计算样本池中所有未标记样本对应的 p 和 q，对它们进行标准化处理，然后根据(4)式计算它们对应的加权和 w，最后从所有未标记样本中选择加权和最小的一个样本作为要询问的样本。(4)式中 a 的大小反映了学习器设计者更倚重于两个因素中的哪一个。至于学习器每次可询问几个样本，并没有严格的规定，但是，不难看出，学习器每次询问的未标记样本数越多，最后学习得到的分类器的推广能力就越差，这也是符合人们的直观理解的。

4. 实验结果

为了验证本文提出的支持向量机主动学习策略的有效性，我们做了一系列文本分类实验。我们选用的实验数据是 Newsgroups 数据集，该数据集也是文本分类领域的研究人员常用的基准数据之一。这里使用了该数据集中五组标为 comp. * 的共计5000个文本。经过取词根处理，并去掉功能词和最多只在三个文本中出现的词后，我们将每个文本表示为一个8992维的向量；对于文本向量中的每个分量，我们取其 TFIDF 值，并进行标准化处理，使得每个向量的长度为1。

我们从这5000个文档中取出一半作为独立的测试集，并随机地从剩下的2500个文档中取出500个组成未标记样本池。对于这五个主题中的每一个，我们将该主题下的文本作为正样本，而将其它四个主题下的文本作为负样本。支持向量机每次从两个随机选取的标记样本学习起，一个是正样本，一个是负样本。这里使用测试精度作为对支持向量机推广能力的评价标准，并且，对于这五个主题中的每一个，我们执行10次训练－测试过程，并将每次的测试结果累计后求平均值。

表1　为达到同样的分类精度三种策略所需标记样本数目的比例关系

策略 / 比例 / 精度	混合规则	不确定性规则	随机
60.0	1	1.62	2.10
65.0	1	1.82	3.55
70.0	1	2.08	3.75
75.0	1	1.56	3.06
80.0	1	1.86	4.33
85.0	1	1.89	>4.33
平均值	1	1.80	>3.52

图2(a)中所示的曲线是将主题为 comp.os.ms-windows.misc 的文本作为正样本而将其余四个主题的文本作为负样本进行二元分类处理时得到的测试精度曲线图。图2(b)中所示的曲线是对所有五个主题的分类精度求平均值所得的精度曲线图。为了同现有的方法做比较，在实验中，支持向量机在选择要询问的未标记样本时，分别采用了三种不同的策略：一是本文提出的混合策略，实验结果对应着图2中的点划线；二是基于如何快速减小版本空间的策略，实验结果对应着图2中的点线；三是随机策略，即支持向量机随机地选取下一个未标记的样本，实验结果对应着图2中的实线。图2中的水平虚线表示当样本池中的500个未标记样本全部加上标记后所得到的分类精度。表1是从图2(b)得出的为达到相同的分类精度三种策略所需标记样本数目的比例关系。

从图2及表1我们看出，如果采用本文提出的混合策略来指导支持向量机的学习，为达到同样的分类精度，我们所需的标记样本数分别约是采用随机策略所需标记样本数的1/4和采用类别不确定性规则所需标记样本数的1/2。

这里要指出的是，通过实验我们发现，当采用混合策略时，在其它条件相同的情况下，当 a 在0.6左右取值时，为达到同样的分类精度，支持向量机所需的标记样本数目最少。图2中的点划线所对应的精度曲线正是 a=0.6时的情形。

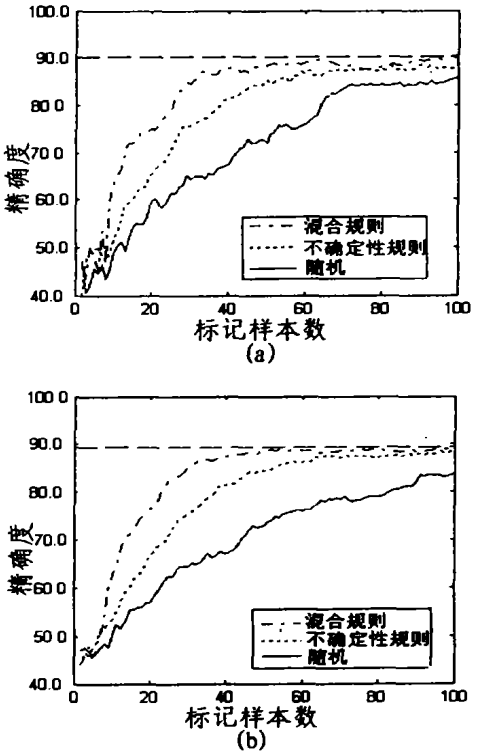

图2　(a)主题 comp.os.ms-windows.misc 对应的分类精度曲线图
(b)五个主题的平均分类精度曲线图

结论和展望　同监督学习模型相比，主动学习模型使得学习器通过对少量标记样本的学习就能得到较好的推广能力。同现有的支持向量机主动学习策略相比，本文提出的混合策略能综合更多的样本信息来确定要询问的样本，因而学习器做出的选择决定更趋合理，为达到同样的推广能力，所需的标记样本的数目更少。我们的文本分类实验表明，本文提出的这一新的支持向量机主动学习策略是十分有效的。

在实验过程中我们发现，现有的支持向量机训练算法速度仍比较慢，由于训练样本集合是在原来的基础上不断扩充，因此，如果能实现支持向量机的增量式(Incremental)学习，则会大大降低实验过程中的计算量。下一步，我们准备借鉴文[7]提出的增量式支持向量机学习算法来进一步改进我们的

（下转第135页）

Y_5),conf=0.5<supp(d)=$|Y_5|/N$(此时 $Int<1$),也即是说已知脚或翅相像的鸟类,嘴相像的可能性为0.5,要比事先我们不知任何信息,嘴相像的可能性还要小;因此这条规则实际使用价值不大;我们不妨考虑其使用价值更大的反面示例:$(b/c,Y_4)\Rightarrow(\overline{d},\overline{Y_5})$,其兴趣度 $int=N|Y_4\cap(U^2/Y_5)|/|Y_4\|U^2/Y_5|=1.25>1$,因此我们更偏向于使用这条反面规则,或者说我们对这条反面规则更感兴趣。

另外为减少噪音对规则的影响,有必要引进支持度门限 θ,若结点(B,Y)满足$|Y|\geq\theta|U^2|$,则称之为频繁结点,否则为非频繁结点;在频繁结点间提取规则,能有效地避免噪音。下面我们就给出一个基于兴趣度的在已构建好的广义概念格中提取无冗余规则的算法描述。

输入:构建好的广义概念格 L,支持度门限 minsup,可信度门限 minconf,最小兴趣度 minInt
输出:包含负面属性的无冗余规则集 R
Begin
Step1. R=∅,N=$|U^2|$;
Step2. 找到格 L 中的所有频繁结点(A,Y),即所有外延项$|Y|\geq$Nminsup 的结点;
Step3. For all frequentnode (A,Y)∈L do
{if node (A,Y) is the form (B/C,Y) then R=R∪{B⇔C,with supp=$|Y|/N$,conf=1,int=$N/|Y|$};
else if (A,Y) has directsuccessornode (B,Y') then
{sup $p=|Y|/N$;$conf$=1;int=$N/|Y'|$;
if Int≥minint then
{if $|Y|/|Y'|\geq$minconf Then
$R=R\cup\{A\Rightarrow B$,with supp,1,int $\}\cup\{B\Rightarrow A$,with supp, conf=$|Y|/|Y'|$,int};
else
$R=R\cup\{A\Rightarrow B$,with supp, 1, int};
}}
else (A,Y) have childnode (B_1,Y_1) and (B_2,Y_2) then
{sup $p=\{Y_1\cap Y_2|/N$;$conf$=$|Y_1\cap Y_2|/|Y_1|$;Int=$N|Y_1\cap Y_2|/|Y_1\|Y_2|$;
if int>minInt then
{if (conf≥minconf) then R=R∪{$B_1\Rightarrow B_2$, supp, conf, int};}
else if int ≤1 then
{S=$|Y_1\cap(U^2/Y_2)|/N$;C=$|Y_1\cap(U^2/Y_2)|/|Y_1|$;I=$NC/|U^2/Y_2|$;
if s≥minsupp and c≥minconf and I ≥minInt Then
R=R∪{$B_1\Rightarrow\overline{B}_2$,with s, c , i};
}}
}
endfor
Step4. return(R);
Step5. End.

若取 minsup=0.38 ,minconf=0.72, minInt=1.05时,通过该算法即可得如下:

确定性规则 $b\Leftrightarrow c$:with 0.48,1, 2.08;$de\Rightarrow e$:with 0.44,1, 1.47 ;

$de\Rightarrow$d:0.44,1,1.66 ;$a\Rightarrow e$:0.44,1,1.47;

缺省(可能)性规则 $d\Rightarrow e$:0.44,0.73,1.08;$d\Rightarrow de$:0.44,0.73,1.67

讨论 本文提出的广义概念格,优于文[9]所论述的一般概念格,因为它适应于更普遍存在的广义粗近似空间,而一般概念格只适应基于等价类的近似空间。当然,对基于等价划分的近似空间也可构造广义概念格,但通过两者所提取规则的描述方式不同,前者可描述到属性值,而后者只能描述到属性广义相似与否。从本文可看出用广义概念格产生规则较为直观有效,且一般只需 Y 的规模信息,无需 Y 全信息,这样最终可在结点外延项只记录|Y|值,从而大大减小存储空间,另外映射 h 的定义方式关系到所对应广义概念格的建立及规则提取,因此根据具体 h 的定义方式类型,研究其所对应的广义概念格模型及规则提取是很有意义的。

参 考 文 献

1 Slowinski R, et al. A generalized definition of rough approximations based on similarity. IEE Transaction on knowledge and data engineering, 2000, 12: 331～336
2 Mcdin D L, Gentner, et al. respect for similarity. Psychology Review, 1993, 100: 254～278
3 Wille R. Restructing lattice theory: an approach based on hierarchies of concept. In: Rival I ed. dordrecht: Reidel. 1982. 445～470
4 Godin R, et al. Incremental concept formation algorithms based on galois lattices. Computation intelligence, 1995, 11(2): 246～267
5 Mollestad T. A rough set approach to data mining: [PhD Thesis]. the Norwegian univ of science and technology, norway, 1997
6 Klemettinen M, et al. finding interesting rules from large sets of discovered association rules. In: proc of the 3 rd int'l conf. on information and knowledge management, USA Acm press, 1994. 401～407
7 Savasere A, et al. mining for strong negative associations in a large database of customer transactions. In: proc of the 14th int'l conf. on data engineering. USA: IEE Computer society press, 1998. 494～502
8 Oosthuizen G D. Rough sets and concept lattices, in: ziarko wped. Rough sets , and fuzzy sets and knowledge discovery. London: Springer-verlag, 1994. 24～31
9 Lwinski T B. Algebrac approach to rough sets. Bulletin of the polish academy of sciences: Mathematics, 1987, 35(9-10): 673～683

(上接第112页)

实验。另外,式中权重系数的确定问题仍未解决,这也是我们下一步的研究方向之一。

参 考 文 献

1 Sebastiani F. Machine learning in automated text categorization. ACM Computing Surveys, 2002, 34(1): 1～47
2 Vapnik V N. Statistical Learning Theory. New York: Wiley, 1998
3 Lewis D, Gale W. A sequential algorithm for training text classifiers. In: Proc. of the Seventeenth Annual Intl. ACM-SIGIR Conf. Research and Development in Information Retrieval, Springer-Verlag, 1994. 3～12
4 Tong S, Koller D. Support Vector Machine Active Learning with Applications to Text Classification. Journal of Machine Learning Research, 2001, 2: 45～66
5 Schohn G, Cohn D. Less is more: Active learning with support vector machines. In: Proc. of the Seventeenth Intl. Conf. on Machine Learning. 2000
6 Roy N, McCallum A. Toward Optimal Active Learning through Sampling Estimation of Error Reduction, ICML-2001. 441～448
7 Cauwenberghs G, Poggio T. Incremental and decremental support vector machine learning. Advances in Neural Information Processing Systems. 2001
8 程云鹏,等. 矩阵论(第二版). 西北工业大学出版社,1999
9 http://kernel-machines.org/