基于近似支持向量机的 Web 文本分类研究*)

钟 将[1] 温罗生[1] 冯 永[1] 叶春晓[1] 李志国[1,2]

(重庆大学计算机学院 重庆 400044)[1] (上海宝信软件西南研发中心 重庆 400041)[2]

摘 要 文本分类技术是知识管理系统实现知识有效组织、存储和检索的重要手段。本文提出了一种新的基于近似支持向量机的分类算法,并将该分类算法应用于文本分类分析。实验过程中与现有的分类方法比较,新的分类方法具有训练速度快、分类精度比较高的优点。

关键词 文本分类,近似支持向量机,二次规化,降维算法

Study on the Web Classification Based on Proximal Support Vector Machine

ZHONG Jiang[1] WEN Luo-Sheng[1] FENG Yong[1] YE Chun-Xiao[1] LI Zhi-Guo[1,2]

(College of Computer Science, Chongqing University, Chongqing 400044)[1]

(Southwest Research Center of Shanghai Baosight Software Corporation, Chongqing 400041)[2]

Abstract Classification of the documents is a very important task. Based on the proximal support vector machines (PSVM) classification method could solve classification problem with small training set, without too much loss of classification accuracy. This paper describes a new PSVM training algorithm based on descending dimension methods, which has faster training speed and smaller memory requirements advantages. Finally we apply the method to solve text classification problem. Experiments results show that the new classification algorithm has better classification performance.

Keywords Text classification, Proximal support vector machines, Quadratic programming, Descending dimension algorithm

随着网络技术的迅速发展,企业内部的数据规模呈指数增长,如何从海量信息中搜索、过滤、管理这些数据资源成为十分重要的问题。其中以知识管理系统为中心的数据组织和管理的模式逐渐成为企业的首要解决方案[1]。分类方法是有效组织和管理海量的文本信息资源的有效途径,特别是采用计算机实现各类文本信息自动归类是实现信息和知识自动获取、分类和智能检索的基础,是有效利用海量知识的前提,因而受到广大研究者的重视。

文本分类(Text Classification)是指由计算机自动提取文本的特征,依据一定的算法,将文本按内容或属性归到一个或多个类别的过程。因此文本分类技术有助于知识的组织和管理,进而建立合理的知识分类库,对于提高知识检索效率十分有效。目前已有许多机器学习方法应用到文本分类中,如 Vapnik 提出的支持向量机(SVM)[2]、K 近邻(KNN)分类器[3],Generalized Instance Set 的方法[4]等。这些分类算法基于文档的向量空间表示模型,然后在每个类别的训练文本集合的基础上训练出一个分类器,最后通过分类器将文本分类。

近年来,近似支持向量机(proximal SVM, PSVM)[3]由于具备较快的训练速度,特别在高维的文本分类算法中能够有效避免过学习的问题,具备较好的鲁棒性和推广性能,因此近年来出现了多种基于近似支持向量机的文本分类算法。

1 近似支持向量机

近似支持向量机与标准支持向量机的主要区别在于它们对应的优化问题约束条件不同,即支持向量机将问题归结成为线性不等式约束二次规划问题,而近似支持向量机将问题归结成为仅含线性等式约束二次规划问题。

标准 SVM 使用一个 n 维向量空间中的超平面 $w\cdot x+b=0$ 来分割正类和负类,其分类函数可写为

$$c(x)=\begin{cases}+1, \text{if } w\cdot x+b\geqslant 0\\ -1, \text{if } w\cdot x+b<0\end{cases} \tag{1}$$

其中的分割超平面是由两个参数 w 和 b 决定的。SVM 算法的目标就是根据训练样本来确定 w 和 b。标准 SVM 通过求解下面的优化问题来确定 w 和 b:

$$\begin{cases}\min & \frac{1}{2}\|w\|^2+\sum_{i=1}^{m}\xi_i\\ \text{s.t.} & y_i(w\cdot x_i+b)+\xi_i\geqslant 1\\ & \xi_i\geqslant 0\end{cases} \tag{2}$$

近似支持向量机 PSVM 也使用一个超平面 $w\cdot x+b=0$ 来分割正类和负类,但其参数 w 和 b 是通过求解如下优化问题决定的:

$$\begin{cases}\min & \frac{1}{2}(\|w\|^2+b^2)+C\sum_{i=1}^{m}\xi_i^2\\ \text{s.t.} & y_i(w\cdot x_i+b)+\xi_i=1\end{cases} \tag{3}$$

可以看到,式(2)和式(3)的主要区别在于优化问题的约束条件不同。式(2)中使用的是不等式约束,而式(3)中使用的是等式约束。这意味着在标准 SVM 中,只有位于两个超平面 $w\cdot x+b=1$ 和 $w\cdot x+b=-1$ 之间的点才会产生训练

*)浦东新区科技发展基金(PKK2005-07);国家发改委科学研究计划项目(CNGI-04-6-2T)。钟 将 博士,主研方向:知识管理、知识发现;温罗生,冯 永,李志国 博士;叶春晓 副教授。

误差，而PSVM中位于这两个超平面之内和之外的点都会产生分类误差。在这种情况下，训练误差 ξ_i 可能为正也可能为负，所以在式(3)的目标函数中使用 $\sum_{i=1}^{n}\xi_i^2$ 作为损失函数。PSVM分类器的目标可总结为：使正类尽量靠近 $w\cdot x+b=1$，使负类尽量靠近 $w\cdot x+b=-1$，使两个超平面 $w\cdot x+b=1$ 和 $w\cdot x+b=-1$ 之间的间隔尽量大[4,5]。

2 基于降维的近似支持向量机学习算法

2.1 等式约束问题降维形式的K-T条件

根据式(3)，近似支持向量机的训练和学习过程可以看成一个线性等式约束二次规划问题。本文将设计一种基于降维和分块矩阵的文本分类算法。首先介绍一般的等式约束问题的降维K-T条件：

$$(ECP)\begin{cases}\min & f(x)\\ \text{s.t.} & h(x)=0\end{cases} \tag{4}$$

其中，$f:R^n\to R, h:R^n\to R^m, m\leqslant n$，记 $p=n-m$，

$$P(x)=(\frac{\partial f(x)}{\partial x_1},\frac{\partial f(x)}{\partial x_2},\cdots,\frac{\partial f(x)}{\partial x_p})^T$$

$$Q(x)=(\frac{\partial f1(x)}{\partial x_{p+1}},\frac{\partial f(x)}{\partial x_{p+2}},\cdots,\frac{\partial f(x)}{\partial x_n})^T$$

$$N(x)=\begin{bmatrix}\frac{\partial h_1(x)}{\partial x_1} & \frac{\partial h_1(x)}{\partial x_2} & \cdots & \frac{\partial h_1(x)}{\partial x_p}\\ \frac{\partial h_2(x)}{\partial x_1} & \frac{\partial h_2(x)}{\partial x_2} & \cdots & \frac{\partial h_2(x)}{\partial x_p}\\ \vdots & \vdots & \vdots & \vdots\\ \frac{\partial h_m(x)}{\partial x_1} & \frac{\partial h_m(x)}{\partial x_2} & \cdots & \frac{\partial h_m(x)}{\partial x_p}\end{bmatrix}$$

$$M(x)=\begin{bmatrix}\frac{\partial h_1(x)}{\partial x_{p+1}} & \frac{\partial h_1(x)}{\partial x_{p+2}} & \cdots & \frac{\partial h_1(x)}{\partial x_n}\\ \frac{\partial h_2(x)}{\partial x_{p+1}} & \frac{\partial h_2(x)}{\partial x_{p+2}} & \cdots & \frac{\partial h_2(x)}{\partial x_n}\\ \vdots & \vdots & \vdots & \vdots\\ \frac{\partial h_m(x)}{\partial x_{p+1}} & \frac{\partial h_m(x)}{\partial x_{p+2}} & \cdots & \frac{\partial h_m(x)}{\partial x_n}\end{bmatrix}$$

定理1[4] 设 $x^0\in R^n$ 是(*ECP*)的最优解，f 和 h 连续可微；若矩阵 $M(x^0)$ 是非奇异，则有

$$P(x^0)=[M(x^0)^{-1}N(x^0)]^TQ(x^0) \tag{5}$$

定理2[5] 假定有

ⅰ) $h(x^0)=0$；

ⅱ) 矩阵 $M(x^0)$ 是非奇异；

ⅲ) $P(x^0)=[M(x^0)^{-1}N(x^0)]^TQ(x^0)$；则 x^0 是(*ECP*)的一个K-T点。

推论1 如果 $x^0\in R^n$ 是方程组

$$\begin{cases}P(x)=[M(x)^{-1}N(x)]^TQ(x)\\ h(x)=0\end{cases} \tag{6}$$

的解，使得 $M(x^0)$ 非奇异，则 x^0 是(*ECP*)的一个K-T点。

相对于经典的K-T条件，由于表达式(6)不含对应的Lagrange乘子，方程的维数得以降低 m 维(等式约束个数)，所以我们称条件(6)为等式约束问题(*ECP*)的降维K-T条件。

2.2 线性等式约束二次规划问题

再考虑线性等式约束的二次规划问题：

$$(EQP)\begin{cases}\min & f(x)=\frac{1}{2}x^TGx+g^Tx\\ \text{s.t.} & Ax=b\end{cases} \tag{7}$$

其中 G 是 n 阶对称矩阵，$g\in R^n$，A 是 $m\times n$ 矩阵，$b\in R^m$，假定秩 $A=m$；显然有

$$\begin{aligned}&\nabla f(x)=Gx+g\\ &\nabla(Ax-b)=A\end{aligned} \tag{8}$$

利用降维形式的K-T条件知道，问题(EQP)的K-T点可由方程组：

$$\begin{cases}P(x)=[M^{-1}N]^TQ(x)\\ Ax-b=0\end{cases} \tag{9}$$

的解确定。记 $p=n-m$，其中 M 是 A 中的 m 阶非奇异矩阵，用 R 表示 M 在 A 中的列号组成的集合，即 $R=\{i_1,i_2,\cdots,i_m\}$；N 是 A 中的剩余的列组成的 $m\times p$ 矩阵，S 表示 N 在 A 中的列号组成的集合，即 $S=\{j_1,j_2,\cdots,j_p\}$。相应的 x 可以分块成 $\begin{pmatrix}x_R\\ x_S\end{pmatrix}$；而

$$P(x)=\nabla_{x_S}f(x)=(\frac{\partial f(x)}{\partial x_{j_1}},\frac{\partial f(x)}{\partial x_{j_2}},\cdots,\frac{\partial f(x)}{\partial x_{j_p}})^T$$

$$Q(x)=\nabla_{x_R}f(x)=(\frac{\partial f(x)}{\partial x_{i_1}},\frac{\partial f(x)}{\partial x_{i_2}},\cdots,\frac{\partial f(x)}{\partial x_{i_m}})^T$$

满足方程组(9)的 x 即为问题(*EQP*)的K-T点。为更方便求解方程组(9)，注意到问题(*EQP*)的特殊性，有

$$\begin{cases}P(x)=G_Sx+g_S\\ Q(x)=G_Rx+g_R\end{cases} \tag{10}$$

其中 G_S 是 G 取 $j_1,j_2,\cdots,j_p$ 行形成的 $p\times n$ 矩阵，G_R 是 G 取 $i_1,i_2,\cdots,i_m$ 行形成的 $m\times n$ 矩阵，g_S 是 g 取 $j_1,j_2,\cdots,j_p$ 行形成的 p 维向量，g_R 是 g 取 $i_1,i_2,\cdots,i_m$ 行形成的 m 维向量，这样得到如下方程组：

$$\begin{cases}(G_S-(M^{-1}N)^TG_R)x=g_S-(M^{-1}N)^Tg_R\\ Ax=b\end{cases} \tag{11}$$

对上述方程组分块，得到方程组

$$\begin{pmatrix}G_{RR}-(M^{-1}N)^TG_{SR} & G_{RS}-(M^{-1}N)^TG_{SS}\\ N & M\end{pmatrix}\begin{pmatrix}x_R\\ x_S\end{pmatrix}=\begin{pmatrix}-g_R+(M^{-1}N)^Tg_S\\ b\end{pmatrix} \tag{12}$$

其中 G_{RR} 是 G_R 取 $j_1,j_2,\cdots,j_p$ 列形成的 $p\times p$ 矩阵，G_{RS} 是 G_R 取 $i_1,i_2,\cdots,i_m$ 列形成的 $p\times m$ 矩阵，G_{SR} 是 G_S 取 $j_1,j_2,\cdots,j_p$、列形成的 $m\times p$ 矩阵，G_{SS} 是 G_S 取 $i_1,i_2,\cdots,i_m$ 列形成的 $m\times m$ 矩阵。记

$$C=C_{RR}-(M^{-1}N)^TG_{SR}$$

$$D=G_{RS}-(M^{-1}N)^TG_{SS}$$

$$b^*=-g_R+(M^{-1}N)^Tg_S$$

改写方程组(12)得到

$$\begin{pmatrix}C & D\\ N & M\end{pmatrix}\begin{pmatrix}x_R\\ x_S\end{pmatrix}=\begin{pmatrix}b^*\\ b\end{pmatrix} \tag{13}$$

并注意到 M 是非奇异矩阵，用 $\begin{pmatrix}I_p & -DM^{-1}\\ 0 & M^{-1}\end{pmatrix}$ 左乘(4)的两边，得到

$$\begin{pmatrix}C-DM^{-1}N & 0\\ M^{-1}N & I_m\end{pmatrix}\begin{pmatrix}x_R\\ x_S\end{pmatrix}=\begin{pmatrix}b^*-DM^{-1}b\\ M^{-1}b\end{pmatrix} \tag{14}$$

对方程组(14)，若能说明 $C-DM^{-1}N$ 是非奇异的，则方程组的解是容易求得且解唯一[6]。

定义1 称二次规划问题(9)满足二阶充分条件，若 $\forall x\in\{x\in R^n\mid Ax=0\}$，且 $x\neq 0$，则有 $x^TGx>0$。

定理3 若问题(*EQP*)满足二阶充分性条件，则 $C-DM^{-1}N$ 非奇异，进而 $\begin{pmatrix}(G_S-(M^{-1}N)^TG_R)\\ A\end{pmatrix}$ 也是非奇异的，

所有方程组(11)的解唯一。

证明:由于(*EQP*)满足二阶充分性条件,故

$\forall x\in\{x\in R^n\mid Ax=0\}$,且 $x\neq 0$,则有 $x^TGx>0$;

即当$(N\quad M)\begin{pmatrix}x_R\\x_S\end{pmatrix}=0$时,有

$$(x_R^T\quad x_S^T)\begin{pmatrix}G_{RR}&G_{RS}\\G_{SR}&G_{SS}\end{pmatrix}\begin{pmatrix}x_R\\x_S\end{pmatrix}>0$$

由 M 非奇异,可得到 $x_S=-(M^{-1}N)x_R$,有

$$(x_R^T\quad x_S^T)\begin{pmatrix}G_{RR}&G_{RS}\\G_{SR}&G_{SS}\end{pmatrix}\begin{pmatrix}x_R\\x_S\end{pmatrix}=x_R^T(I\quad -(M^{-1}N)^T)$$

$$\begin{pmatrix}G_{RR}&G_{RS}\\G_{SR}&G_{SS}\end{pmatrix}\begin{pmatrix}I\\-M^{-1}N\end{pmatrix}x_R=x_R^T(G_{RR}-(M^{-1}N)^TG_{SR}-(G_{RS}-(M^{-1}N)^TG_{SS})M^{-1}N))x_R=x_R^T(C-D(M^{-1}N))x_R$$

由于 x_R 是自由变量,故 $C-D(M^{-1}N)$是正定的,当然也是非奇异的。

结合式(11)~(14)显然知道

$$\begin{pmatrix}((G_S-(M^{-1}N)^TG_R)\\A\end{pmatrix}$$

也是非奇异的。显然方程组(11)的解存在唯一。证毕。

推论 2 若问题满(*EQP*)满足二阶充分性条件,则问题(*EQP*)存在唯一解。

根据以上的结果,可以得到一个基于降维方法的二次规划算法,用来寻找该问题的最优解。具体算法如下:

算法 1 基于降维的二次规划算法

Step1 对方程 $Ax=b$ 的增广矩阵利用 Gauss 列主元进行变换,得到形式为 $x_R+\bar{N}x_S=\bar{b}$ 的方程,并记下 $\bar{N}$ 所在的列号 $S=\{j_1,j_2,\cdots,j_p\}$和系数矩阵 A 中其余的列号集 $R=\{i_1,i_2,\cdots,i_m\}$;

Step2 根据 R 和 S 得到 $C=G_{SR}-\bar{N}^TG_{RR}$,$D=G_{SS}-\bar{N}^TG_{RS}$,$b^*=\bar{N}^Tg_R-g_S$;

Step3 $x_S=(D-C\bar{N})^{-1}(b^*-C\bar{b})$,$x_R=\bar{b}-\bar{N}x_S$,最优解就为$\begin{pmatrix}x_R\\x_S\end{pmatrix}$。

2.3 基于降维的近似支持向量机学习算法

对近似支持向量机的学习过程,可以看成是式(3)所对应的线性等式约束二次规划问题,该式可以转换为矩阵形式:

$$\begin{cases}\min\quad \dfrac{1}{2}(w^T,b^T,\xi^T)G(w^T,b^T,\xi^T)^T\\ \text{s.t.}\quad (A1,A2,A3)(w^T,b^T,\xi^T)^T=e\end{cases}\tag{15}$$

其中

$$G=\begin{pmatrix}E_n&O&O\\O&1&O\\O&O&CE_m\end{pmatrix},A_1=\begin{pmatrix}y_1w_1&\cdots&y_1w_n\\\cdots&\cdots&\cdots\\y_mw_1&\cdots&y_mw_n\end{pmatrix},$$

$w=\begin{pmatrix}w_1\\\cdots\\w_n\end{pmatrix}$,$\xi=\begin{pmatrix}\xi_1\\\cdots\\\xi_m\end{pmatrix}$,$A_2=\begin{pmatrix}y_1\\\vdots\\y_m\end{pmatrix}$,$A_3=E_m$,$E_m$ 表示 m 阶单位矩阵,e 表示 $m+n+1$ 维列向量,C 为式(3)中权系数。

记 $A=(A_1,A_2,A_3)$,$x=(w^T,b^T,\xi^T)^T$,那么式(15)可以转换为式(7)对应的线性等式约束的二次规划问题,故可以代入到算法 1,求解其对应的最优解 x。

由于传统的近似支持向量机采用基于 K-T 条件进行求解,计算复杂度为 $O(m+n)^3$,其中 m 表示训练样本的个数,n 表示训练数据集的属性的维数。本文中的新的训练方法计算时间包括:降维处理和 n 个 n 变量的方程组求解时间,因此计算复杂度为 $O(m^3+n^3)$。

空间复杂度相应地由传统 PSVM 算法所需的 $O(m+n)^2$ 降为 $O(m^2+n^2)$。因此在 m 和 n 近似时,计算复杂度可以降低为原来的 1/4 左右,空间复杂度减少一半。对于文本分类问题具有高维(通常大于 1000 维),但是训练样本数又较少的情况,本分类算法则对传统的近似支持向量机有优势。

3 基于近似支持向量机的文本分类算法

为了验证本文提出方法的可行性和有效性,本文使用搜狐实验室提供的已分类的语料库。本文选取了其中教育、财经、健康、体育、军事等五类文档,每类文档 1990 个,共计 9950 个文档。

实验过程中为每一类文本训练一个分类器,训练过程中任选 1000 个样本文件作为训练过程中的正例,其它类别中挑出的 4000 个样本为反例。在另外的 990 个文档中进行测试。实验过程中的分词技术采用 2-gram 方法,训练数据选自北京大学计算语言学研究所提供的人民日报标注语料库[7]。为了进一步提高分词精度,实验过程中又添加了 50 篇语料,及每一类文本中采用人工方式选择了 10 篇典型的语料。实验中的文本采用词向量空间的方式来表示。本文在实验过程中通过统计方法,在各类文档数据中各筛选 1500 个高频的名词和动词来构成文档的特征向量。

对于文本分类算法的评估主要考察算法的查全率 P_{recall} 和查准率 $P_{\text{precision}}$ 两个指标。查全率是指正确分类到某类文档的数量 K_{ri} 与该类所有文档的数量之比 K_{ti}。查准率是指正确分类到该类文档的数量与所有划分到该类的文档数量 K_{pi} 之比:

$$P_{\text{recall}}=\frac{\text{正确分类的文本数量 }K_{ri}}{\text{该类文本总数 }K_{ti}}$$

$$P_{\text{precision}}=\frac{\text{正确分类的文本数量 }K_{ri}}{\text{所有划分到该类文本的数量 }K_{pi}}$$

为了比较算法分类的性能,实验同时采用了 KNN 和传统的 SVM 分类方法和 PSVM 方法进行比较,这些方法都采用词向量空间模型来表示文本文件的特征,同时每一类文本也选择了 1000 维特征向量。

表 1 分类算法的性能比较

分类方法	性能	教育	财经	健康	体育	军事
KNN 分类	查准率	0.637	0.681	0.730	0.749	0.792
	查全率	0.624	0.716	0.723	0.714	0.812
SVM	查准率	0.850	0.768	0.770	0.821	0.843
	查全率	0.823	0.816	0.763	0.774	0.872
PSVM	查准率	0.799	0.642	0.708	0.789	0.884
	查全率	0.734	0.751	0.691	0.765	0.854

由于本文提出的文本分类方法与传统的近似支持向量机相比较,主要在空间和时间复杂度上有一定程度的改善,在分类性能上则完全相同,因此表 1 没有列出标准 PSVM 分类方法的分类性能。

从表 1 中基于 PSVM 在分类精度上和查全率上,在大多数情况下好于 KNN 方法,但是总体性能略低于标准的 SVM 方法。但是由于 PSVM 存在唯一的最优解,而且算法的训练时间也比标准的 SVM 快,因此可广泛应用在对训练时间敏感的文本分类环境中。

结论 基于近似支持向量机的分类方法具备标准支持向量机适合小样本训练集分类问题,同时又不损失太多的分类精度。本文中基于降维近似支持向量机的文本分类算法,与传统的基于 K-T 条件的训练方法比较,具有更快的训练速度

(下转第 202 页)

改变而改变，所以不需要重新建立 INUP_Tree，只需要扫描头表的 flag 域。如果 flag 为 0，则比较该项的支持度是否大于 s'；如果大于，则构建它的条件矩阵；如果小于，则继续扫描 flag。对于 flag 为 1 的项，不用再次挖掘。

算法：增量更新算法 $INUPA_1$

输入：INUP_Tree 和 s'
输出：新的频繁项集
1. For each item X_i in Header Table
2. if X_i. flag=1 then
3. continue; //如果 flag=1 说明 X_i 已经挖掘过
4. else
5. { if X'_i count>s' then
6. flag=1 and construct its conditional matrix}
7. End for

改写上例，假设 $s'=2$。扫描头表找到新的频繁项为{f：2}。需要构建它的条件矩阵并挖掘以 f 为后缀的频繁模式。挖掘结果为{f：2，ef：2，bf：2，af：2，bef：2，aef：2，abef：2，abf：2}。

情况 2：由于篇幅所限，我们只考虑当最小支持度不变，一个事务数据集 db 添加到事务数据库 D 中，如何生成事务数据库 D∪db 中的频繁项目集。我们将其分解为以下三个子问题：(1)找出 D 中不再生效的频繁项目集；(2)找出仍然生效的频繁项目集；(3)找出 D∪db 中新的频繁项目集。

我们只需扫描 db，将新数据集插入到 INUP_Tree 中，而不需要重新建立 INUP_Tree。通过调用增量更新算法 $INUPA_2$，来解决以上的三个子问题。

算法：增量更新算法 $INUPA_2$

输入：INUP_Tree，DB，db，s
输出：事务数据库 DB∪db 中的所有频繁项目集
1. 调用函数 insert([t| T_i],T)把新数据集插入到 INUP_Tree 中
2. For each item X_i in Header Table
3. if X_i. flag=1&&X_i. item_count>s then
4. { if X_i has conditional matrix then
5. update X_i's conditional matrix
6. else
7. construct X_i's conditional matrix
8. if X_i. flag=1&&X_i. item_count<s then
9. flag=0 && delete its frequent patterns
10. End for

5 实验结果和性能分析

我们用 VC ++6.0 在内存 512M、CPU 为 Pentium4 2.4GHZ、操作系统为 Windows XP 的环境下实现了 $INUPA_1$ 和 $INUPA_2$ 算法并进行了性能测试。利用蘑菇数据库(mushroom database)来进行实验。该数据库有 8124 条记录，记录了蘑菇的 23 种属性。图 4(a)显示了在不同的最小支持度下算法的性能比较。由于 INUP_Tree 的构建只需扫描数据库一次，而且在最小支持度发生变化时不需要改变 INUP_Tree，不需要任何插入、删除操作，flag 标志的设置缩小了搜索范围，提高了发掘新频繁项集的效率。因此 $INUPA_1$ 算法的执行时间比 FP-Growth 算法要少，图 4(a)说明了这一点。

为了验证增量更新算法 $INUPA_2$ 的有效性，随机抽取 8000 个记录，并分为两部分：原始 DB(4000 个记录)和新增 db(4000 个记录)。最小支持度为 2.5%，从 db 中抽取不同的记录数(500，1000，2000，4000)作为 db 的不同增量情况，对更新算法 $INUPA_2$ 进行测试，结果如图 4(b)所示。由于 $INUPA_2$ 算法只需扫描新增数据 db 一次插入到 INUP_Tree 中，而不需要重新建立 INUP_Tree，并且可以充分利用已有的条件矩阵来加速本次的挖掘，因此其效率极大提高。

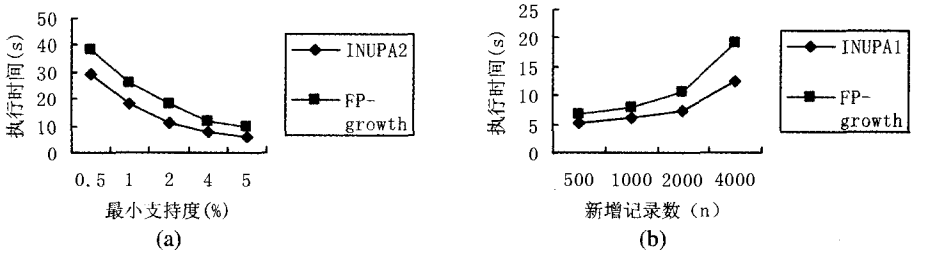

图 4 更新算法的执行时间对比

结束语 本文提出了一种新的增量更新频繁挖掘方法。使用 INUP_Tree 压缩存储数据库中的事务，它的建立只需要扫描数据库一次。同时提出了频繁模式挖掘方法 FPBM_Mine，以条件矩阵为基础进行挖掘，减少了搜索范围，加快了挖掘速度。针对不同的数据库增量更新问题，提出了 INUPA 算法，实验证明这种增量挖掘方法比 FP-growth 方法在处理增量问题时效率更高。

参 考 文 献

1 Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases. In: Proc. of the ACM SIGMOD Conference, 1993. 207～216
2 Agrawal R, Srikant R. Fast algorithms for mining association rules. In: Proc. of VLDB Conf, 1994. 487～499
3 Cheung D W, Han J, Ng V T, el al. Maintenance of discovered association rules in large databases: An incremental updating approach. In: The 12th IEEE International Conference on Data Engineering, 1996. 106～114
4 Agarwal R, Aggarwal C, Prassad V V V. A tree projection algorithm for generation of frequent itemsets. Journal of Parallel and Distributed Computing, 2001. 61: 350～371
5 Han J, Pei J, Yin Y. Mining frequent patterns without candidate generation. In: Proceeding of the ACM SIGMOD Conference, 2000. 1～12

(上接第 169 页)

和更小的内存需求的优点。特别是对于高维小样本集的文本分类问题，基于近似支持向量机能够适用于需要实时重构文本分类器的应用环境。

参 考 文 献

1 Fischer G , Otswald J. Knowledge management: problems, promises, realities, and challenges[J]. IEEE Intelligent Systems and Their Applications, 2001,16(1) :60 ～ 72
2 Kim S-B. Some Effective Techniques for Naive Bayes Text Classification[J]. IEEE Transactions on Knowledge and Data Engineering, 2006, 18(11) :1457 ～ 1466
3 Vapnik V. The Nature of Statistical Learning Theory[M]. New York: Springer, 2000
4 庄东，陈英. 基于加权近似支持向量机的文本分类. 清华大学学报(自然科学版)，2005，45 (S1)：1787～1790
5 Lewis D D, Yang Y, Rose T G, et al. RCV 1: A new benchmark collection for text categorization research [J]. Journal of Machine Learning Research, 2004, 5(2): 361～397
6 Fung G, Mangasarian O L. Proximal support vector machine classifiers [A]. In: Proc. 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining [C]. San Francisco, CA , USA: ACM , 2001
7 温罗生，李泽民. 含有线性和非线性等式约束非线性规划问题的一种降维乘子法. 见：第七届中国运筹学大会论文集. 2005
8 http://www.icl.pku.edu.cn/icl_res/