基于类权重的模糊不平衡数据分类方法*)

薛贞霞[1,2] 张素玲[3] 刘三阳[1]
(西安电子科技大学应用数学系 西安 710071)[1] (河南科技大学数学系 洛阳 471003)[2]
(焦作大学基础部 焦作 454003)[3]

摘 要 针对现有分类算法通常对不平衡数据挖掘表现出有偏性,即正类样本(通常是更重要的一类)的分类和预测性能差于负类样本的分类和预测性能,提出一种不平衡数据分类方法。该方法通过一个超球面将两类数据以最大分离比率分离,并且引入类权重因子和样本模糊隶属度,同时考虑了不同类的重要性和不同样本对该类的不同贡献,从而提高了不平衡数据中正类的分类和预测的性能以及整体的推广能力。分别在人造数据和 UCI 真实数据上进行了实验,结果验证了该方法的有效性。

关键词 不平衡数据,类加权,模糊隶属度,分类算法

Weighted-class Based Fuzzy Classification Method for Class-imbalanced Data

XUE Zhen-xia[1,2] ZHANG Su-ling[3] LIU San-yang[1]
(Department of Applied Mathematics, Xidian University, Xi'an 710071, China)[1]
(Department of Mathematics, Henan Science and Technology University, Luoyang 471003, China)[2]
(Department of Basic Course, Jiaozuo University, Jiaozuo 454003, China)[3]

Abstract Using data sets that contain very few instances of the positive class usually produces biased classifiers and has a lower predictive accuracy over the positive class (usually the more important class) than over the negative class. Proposed a classification method for imbalance problem. This approach obtains maximum separation ratio to separate two class instances with a single sphere. Moreover, this method applies a fuzzy membership to each input point such that different input points can make different contributions to the learning of decision surface, as well as imposes distinct weight factors on each class. By this way the method can improve the predictive accuracy over the positive class, and has more generalization ability on entireness. Experiment results on artificial data sets and UCI data sets show the method's effectiveness.

Keywords Imbalanced data set, Weighted-class, Fuzzy membership, Classification algorithm

1 引言

在机器学习和数据挖掘研究中,非平衡数据通常是指两类问题中的负类样本个数远大于正类样本个数,并且正类样本往往是分类问题的关注所在。在不平衡的情况下,训练出的分类器性能下降,甚至很差。因此,对不平衡数据的学习已成为机器学习目前面临的一个挑战。由 Vapnik 等人创立的支持向量机[1,2](SVM)已经被证实是一种很有效的学习机,已得到广泛的应用[3-5],但是 SVM 对噪点非常敏感。由 Tax 等人在 SVM 的基础上提出了支持向量域描述(SVDD)[6],主要思想是通过计算包含一组数据的最小超球形边界来对该组数据进行描述,它可以对一类数据进行描述和剔除噪点或奇异点。文献[7]提出将 SVM 和 SVDD 的优点结合起来,通过求取一个超球面将两类数据以最大分离比率分离,本文称这种方法为 SSPC。该方法将剔除噪点和分类同时进行,而且得到的支持向量实际上是每一个类别中边界上的点,分类性能得到了很大提高,因此本文研究它在非平衡数据集上的分类性能。基于 SSPC 和模糊支持向量机的思想[13],提出一种针对不平衡数据分类的方法——W-FSSPC(Wighted Fuzzy SSPC)。该方法在 SSPC 中引入类权重因子和样本模糊隶属度,既考虑不同类的重要性,又考虑不同样本点对该类的不同贡献,从而在减少噪点影响的情况下提高不平衡数据的分类性能。

本文第 2 节对 SSPC 进行简单介绍;第 3 节介绍 W-FSSPC;第 4 节给出分类器性能评价的标准以及 SSPC 和 W-FSSPC 方法的对比实验结果和分析,最后在分析的基础上得出结论。

2 SSPC 简介

SSPC[7]是结合 SVM 和 SVDD 的优点,通过求取两个同心的超球面(里面的超球面对一类数据进行描述,外面的超球面对另一类数据进行描述,但是它是将该组数据排除在该超球面外面,当然两个超球面都可以剔除噪点)的最大间隔将两类数据分离。这种方法将剔除噪点和分类同时进行,分类性能确实得到了提高。下面介绍 SSPC 算法。

设给定样本集$(x_i, y_i) \in R^{d+1}$,其中 $x_i \in R^d$,$y_i \in \{-1, 1\}$为相应的类标,$i=1,2,\cdots,n$,SSPC 就是找一个球,设该球为 $S(a,R)$,其中 a 和 R 分别是球心和半径,该球将几乎全部

*)国家自然科学基金(60574075),国家自然科学基金项目(60703118)。薛贞霞 博士研究生,从事机器学习、模式识别和最优化理论方法及应用的研究。

的正类样本点包在内部，将几乎全部的负类点排除在该球球外面，且设该球将两类以间隔 $2d$ 进行分离，再引入松弛变量和非线性映射(核函数)，其优化模型可写为：

$$\min_{a,R,d,\xi_i} R^2-Kd^2+C\sum_{i=1}^{n}\xi_i \tag{1}$$

$$\text{s.t.}: y_i(\|\phi(x_i)-a\|^2-R^2)\leqslant -d^2+\xi_i \tag{2}$$

$$d^2\geqslant 0 \tag{3}$$

其中 $K\geqslant 1$，上述问题的对偶问题为

$$\min_{\alpha_i}\sum_{i,j}\alpha_i\alpha_j y_i y_j k(x_i,x_j)-\sum_i \alpha_i y_i k(x_i,x_i) \tag{4}$$

$$\text{s.t.}\ \sum_i \alpha_i y_i=1 \tag{5}$$

$$\sum_{i=1}^{n}\alpha_i=K \tag{6}$$

$$0\leqslant\alpha_i\leqslant C\quad i\in 1,\cdots,n \tag{7}$$

其中惩罚参数 C 是用来折中最大化间隔和最小化训练错误率的，K 为分离比率参数。这个对偶问题是个二次规划问题，其中 C 为惩罚参数，K 为分离比率，可以用标准的二次规划算法求解，也可以利用解的稀疏性和 KKT 条件用类似于 SVM 的 SMO 算法和其他的分解算法来加速求解过程。其中 $0<\alpha_i\leqslant C$ 对应的向量称为支持向量(SV)，球心可表示为 $a=\sum_{i=1}^{n}\alpha_i y_i\phi(x_i)$，半径 R 和决策函数可以通过以下两式求得：

$$R^2=\frac{\langle\phi(x_i)-a,\phi(x_i)-a\rangle|_{0<\alpha_i<C,y_i=-1}+\langle\phi(x_i)-a,\phi(x_i)-a\rangle|_{0<\alpha_i<C,y_i=1}}{2} \tag{8}$$

$$f(x)=\text{sign}(R^2-k(x,x)-\sum_{i,j}\alpha_i\alpha_j y_i y_j k(x_i,x_j)+2\sum_i \alpha_i y_i k(x_i,x)) \tag{9}$$

3 W-FSSPC 方法

本节提出一种针对不平衡数据分类的方法，即 W-FSSPC (Wighted Fuzzy SSPC)，该方法在 SSPC 中引入类权重因子和样本模糊隶属度，同时考虑了不同类的重要性和不同样本隶属度的差别，既调整正类和负类的权重，又给每个样本赋予不同的隶属度，从而既提高了不平衡数据对正类分类和预测的性能，又具有像模糊支持向量机那样可以减少噪点的影响，进而也提高了整体的分类性能。

3.1 W-FSSPC

设给定样本集 $\{(x_i,y_i,\lambda_{y_i},\mu_i)\,|\,x_i\in R^d, y_i\in\{-1,+1\}, \mu_i\in(0,1),\lambda_{y_i}\in(0,+\infty), i=1,2,\cdots,n\}$，其中 d 为样本点的特征数即维数，y_i 为相应的类标，λ_{y_i} 表示类 y_i 的权重，μ_i 表示样本点 x_i 的隶属度。我们所提出的 W-FSSPC 即为如下的优化问题：

$$\min_{a,R,d,\xi_i} R^2\text{-}Kd^2+C\sum_{i=1}^{n}\lambda_{y_i}\mu_i\xi_i \tag{10}$$

$$\text{s.t.}: y_i(\|\phi(x_i)-a\|^2-R^2)\leqslant -d^2+\xi_i \tag{11}$$

$$d^2\geqslant 0 \tag{12}$$

其中 $K\geqslant 1$，上述问题的对偶问题为

$$\min_{\alpha_i}\sum_i\sum_j\alpha_i\alpha_j y_i y_j k(x_i,x_j)-\sum_i\alpha_i y_i k(x_i,x_i) \tag{13}$$

$$\text{s.t.}\ \sum_i\alpha_i y_i=1 \tag{14}$$

$$\sum_{i=1}^{n}\alpha_i=K \tag{15}$$

$$0\leqslant\alpha_i\leqslant C\lambda_{y_i}\mu_i\quad i\in 1,\cdots,n \tag{16}$$

这个对偶问题是个二次规划问题，其中 C 和 K 与 SSPC 中意义相同。如果将 λ_{y_i} 和 s_i 设置为 1，则 SSPC 和 W-FSSPC 完全相同，$0\leqslant\alpha_i\leqslant C\lambda_{y_i}\mu_i$ 对应的向量称为支持向量(SV)，决策函数形式与(9)式相同。

3.2 模糊隶属度的确定

现在我们来确定样本点的模糊隶属度。假设我们经过 SSPC 的训练得到 $\alpha_i(i\in 1,\cdots n)$、球心 a 和半径 R，我们任取一个 $0<\alpha_i<C$ 对应的正类点可以求出 $d^2=R^2-\|\phi(x_i)-a\|^2$ 或者任取一个 $0<\alpha_j<C$ 对应的负类点也可以求出 $d^2=\|\phi(x_j)-a\|^2-R^2$。令

$$d_i^2=\|\phi(x_i)-a\|^2=k(x_i,x_i)-2\sum_{j=1}^{n}\alpha_j y_j k(x_i,x_j)+\sum_j\sum_i\alpha_i\alpha_j y_i y_j k(x_i,x_j)$$

$$L^2=\min_{y_i=-1}\|\phi(x_i)-a\|^2, r_{\min}^2=\max\{L,R^2\}\ \text{和}\ r_{\max}^2=\max_{y_i=-1}\|\phi(x_i)-a\|^2$$

则我们根据样本点的几何信息来定义隶属度，图 1 为简单的示意图。正类点的隶属度定义为

$$\mu_i^{+}=\begin{cases}1-\dfrac{d_i^2}{R^2}+\sigma^{+}, d_i^2\leqslant R^2+d^2\\ 0,\ \text{otherwise}\end{cases} \tag{17}$$

负类点的隶属度定义为

$$\mu_i^{-}=\begin{cases}1-\left(\dfrac{d_i^2-\frac{r_{\max}+r_{\min}}{2}}{\frac{r_{\max}-r_{\min}}{2}}\right)^2+\sigma^{-}, R^2-d^2\leqslant d_i^2\leqslant r_{\max}\\ 0,\ \text{otherwise}\end{cases} \tag{18}$$

其中 $0<\sigma^{-}<\sigma^{+}\leqslant 0.5$。

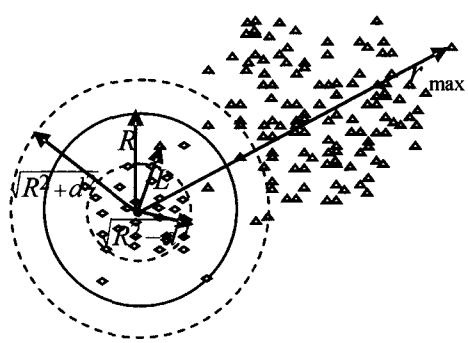

图 1 隶属度函数确定的示意图

我们这样定义隶属度有以下特性：

①对正类点来说与球心距离越小隶属度越大，反之越小；

②对负类点来说，当 d_i^2 在区间 $(r_{\min},\frac{r_{\min}+r_{\max}}{2})$ 内递增，其隶属度逐渐递增，当 d_i^2 在区间 $(\frac{r_{\min}+r_{\max}}{2},r_{\max})$ 内递增，其隶属度逐渐递减。

③两类的隶属度函数中有可以调节的自由参数 σ^{+} 和 σ^{-}，我们设置 $0<\sigma^{-}<\sigma^{+}\leqslant 0.5$，是为了使分类间隔内的(包括分类超球面上)的正类点的隶属度大于负类点的隶属度，从而提高正类点的分类精度。实际应用中可以通过交叉验证来选择参数 σ^{+} 和 σ^{-}。

4 分类器性能评价的标准及实验结果和分析

4.1 分类器性能评价的标准

分类精度(accuracy, ac)是最常用的标准，但前提是类分布是已知不变的，且每个正类样本和负类样本的误分类损失

是相等的，显然单独用它来衡量两类不平衡数据学习问题是不合适的。许多学者提出了很多针对不平衡数据的分类器评价标准，其中有以下测度：1）敏感性（sensitivity，Se）Se＝tp/(tp＋fn)；2）特效性（specificity，Sp）Sp＝tn/(tn＋fp)；3）几何均值（geometry mean，Gm）$Gm=(Se*Sp)^{1/2}$，其中 tp，fp，tn，fn 分别表示真实的正类、错误的正类、真实的负类和错误的分类的样本个数，显然 ac＝(tp＋tn)/(tp＋fp＋tn＋fn)。Se 可以表达正类的分类精度；Sp 可以表达负类的分类精度；Gm 表达了正类和负类精度的几何均值。当正类和负类的分类精度都比较高时，Gm 就比较高，这样更能反映分类器的整体性能好坏。

4.2 实验结果和分析

为了检验 SSPC 和 W-FSSPC 的性能，我们分别用人工数据和真实数据进行了实验。实验是在奔腾 1.73G，512 内存的 PC 机上安装的 Matlab7.0 软件上实现的，SSPC 和 W-FSSPC 是我们编制的 Matlab 程序。两个实验中 n_+ 和 n_- 均分别代表正类和负类的个数。样本点的隶属度是利用式(17)和(18)，C 均取 1000，K 取 100。人工数据使用的是随机生成的二维数据。正类的范围是 0.1 到 0.5，负类数据的范围是 0.4 到 0.9。使用的核函数为线性核。我们设置两类的类权重之比与样本点个数之比成反比，即 $\frac{\lambda_+}{\lambda_-}=\frac{n_-}{n_+}$，$\sigma^-=0.3$，$\sigma^+=0.5$，实验结果（20 次实验的平均值）如图 2 所示。我们又在 8 个真实数据分类问题上进行了实验，它们来自 UCI 机器学习库[14]。表 1 给出了所选数据集的特征。对每个分类问题我们将数据标度到区间[－1，1]，随机选择数据集的 2/3 来训练，1/3 来测试。SSPC 和 W-FSSPC 在这 8 个真实数据上的实验结果（20 次实验的平均值）如表 2 所示。注意：①由于是不平衡问题，主要比较前后效果，训练集及测试集一旦取出都保持不变；②选取负类权重均为 1 来估计性能；③仅把最佳参数的实验结果列出，所有实验均采用 Gauss 核函数，核参数用 σ 表示；④对多类分类问题，本文将其中一类看作正类，其他类看作负类。最好的结果用黑体标出。

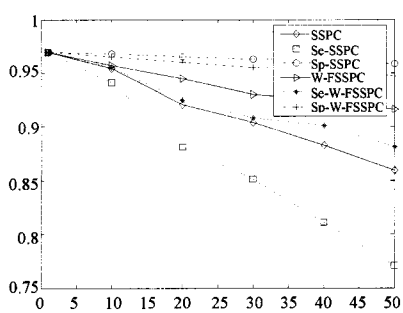

图 2 人造数据的实验结果

表 1 真实数据集的数据特征

问题	样本数	类别数	正类类别	训练集(n+，n－)	测试集(n+，n－)
Ionosphere	351	2	Good	234(84,150)	117(42,75)
Wisconsin	569	2	Malignant	380(142,238)	189(119,70)
Pima	768	2	1	513(179,334)	255(89,166)
German	1000	2	Bad	667(200,467)	333(233,100)
Wine	178	3	3	119(32,87)	59(16,43)
Glass	214	6	Ve-win	142(11,131)	72(6,66)
Bupa	345	2	1	230(77,153)	115(38,77)
Breast	683	2	Malignant	455(150,305)	228 (75,153)

表 2 真实数据上 SSPC 与 W-FSSPC 的实验对比

问题	SSPC	Se	Sp	Gm	W-FSSPC	Se	Sp	Gm
	σ				$\sigma,\lambda+,\sigma^+,\sigma^-$			
Ionosphere	0.5	0.54	0.98	0.73	0.5,5,0.5,0.3	**0.77**	0.91	**0.84**
Wisconsin	1	0.83	0.96	0.89	0.5,6,0.5,0.5	**0.93**	0.93	**0.93**
Pima	0.5	0.60	0.79	0.69	0.5,8,0.5,0.2	**0.69**	0.75	**0.72**
German	0.5	0.49	0.86	0.65	0.5,10,0.4,0.2	**0.60**	0.81	**0.70**
Wine	0.5	0.87	0.99	0.93	0.5,2,0.5,0.4	**0.91**	0.99	**0.95**
Glass	0.5	0.62	0.81	0.71	0.5,12,0.5,0.3	**0.68**	0.76	**0.72**
Bupa	1	0.52	0.61	0.56	1,2,0.4,0.3	**0.55**	0.58	**0.56**
Breast	0.5	0.87	0.95	0.91	0.5,2,0.5,0.3	**0.92**	0.94	**0.93**

由图 2 的实验结果示意图可以看出，随着不平衡程度的增大，与 SSPC 相比，W-FSSPC 虽然使负类的分类精度有所降低，但却使正类的分类精度有很大提高，从而使整体的分类性能提高了。由表 2 可以看到：①在正类的分类精度上，各个数据通过适当参数调节，本文方法均高于 SSPC，这说明本方法可以提高不平衡数据的正类分类性能；②在两类的分类精度的几何均值上除了 Bupa 数据的没有发生变化以外其余的都有所提高，几何均值可以反映整体的分类性能，这说明本文方法可以保持整体的分类性能不减，甚至有所提高。注意：①在实际应用中，如何确定类权重需要专门研究领域的知识，一个简单有效的方法就是可以用反复训练的方法确定类的权重，②本文方法需要调节类权重和两个参数 σ^+ 和 σ^-，可能会占用一些时间，但是对于不平衡数据要求正类和整体的性能都提高的目标下这是值得的。

结束语 本文提出了一种被称为 W-FSSPC 的不平衡数据分类方法。与 SSPC 相比，W-FSSPC 在保持整体分类性能不减甚至提高的情况下提高了正类的分类精度，降低了重要样本被错分的可能性。下一步我们要研究的是如何把 W-FSSPC 推广到更多的实际应用问题中以及多类不平衡数据的分类问题中。

参 考 文 献

[1] Vapnik V. Nature of Statistical Learning Theory. New York: Springer-Verlag, 2000

[2] Cristianini N, Shawe-Tayor J. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods[M]. Cambridge: Cambridge University Press, 2000

[3] 胡国胜，任震. 基于支持向量机混合模型的短期负荷预测方法[J]. 高电压技术，2006，32(4)：101-103

[4] Chuang C C, Su S F, Jeng J T, et al. Robust support vector regression networks for function approximation with outliers [J]. IEEE Trans. on Neural Network, 2002, 13(6): 1322-1330

[5] Guo G D, Li S Z. Content-based audio classification and retrieval by support vector machines[J]. IEEE Trans. on Neural Network, 2003, 14 (1): 209-215

[6] Tax D, Duin R. Support Vector Domain Description [J]. Pattern Recognition Letters, 1999, 20: 11-13

[7] Wang J, Neskovic P, Cooper L N. Pattern Classification via Single Spheres [J]. Discovery Science, 2005, 3735: 241-252

[8] Japkowicz N, Stephen S. The Class Imbalance Problem: A Systematic Study [J]. Intelligent Data Analysis, 2002, 6(5): 429-449

[9] Ling C, Li C. Data Mining for Direct Marketing Problems and Solutions[C]// Proceedings of the Fourth International Conference on Knowledge Discovery and Data Ming. New York: AAAI Press, 1998: 73-79

[10] Chawla N V, Japkowicz N, Kolcz A. Editorial: Special Issue on Learning from Imbalanced Data Sets [C]. ACM SIGKDD Explorations, 2004, 6(1): 1-6

[11] Weiss G M. Mining with Rarity - Problems and Solutions: A Unifying Framework [C]. SIGKDD Explorations, 2004, 6(1): 7-19

[12] 肖健华，吴今培. 样本数目不对称时的 SVM 模型[J]. 计算机科学，2003，30(2)：165 - 167

[13] Lin C F, Wang S D. Fuzzy support vector machines [J]. IEEE Transactions on Neural Networks, 2002, 13(2): 464-471

[14] Blake C L, Merz C J. UCI Repository of Machine Learning Database[DB/OL]. 1998. [2007-5-21]. http://www.ics.uci.edu/~mlearn/MLRepository.html

(上接第 165 页)

```
    newSkip[ch]=skip[ch]; //当 ch 在 pattern 中出现 0 次或 1 次时,newSkip[ch]等于模式串长度 m
    skip[ch]=m-i-1;// skip 数组的定义与 BMH 算法相同
    preChar[ch]=pattern[i-1];
}
i=m-1;
while(i<n) {
    k=i; j=m-1; //k 记录 text 中每次从右至左开始比较的起始位置
    while((j>=0)&&(pattern[j]==text[i])){
        i--; j--;
    }
    if(j==-1) return i+1; //在 text[i+1]处匹配成功
    if(text[k-1]! =preChar[text[k]])
        i=k+newSkip[text[k]]; //采用改进后的策略移动文本指针
    else
    i=k+skip[text[k]];
}
return -1; //匹配失败
```

根据 BMH2 算法，上文中的文本与模式串的匹配过程如表 2 所示。

表 2　改进算法匹配过程

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	
文本	a	b	h	d	g	f	d	a	b	b	d	b	d	a	b	d	b	f	d	文本指针移动距离
第1次	a	b	d	b	d															newSkip['f']=6
第2次								a	b	d	b	f	d							skip['b']=2
第3次									a	b	d	b	f	d						newSkip['a']=5
第4次															a	b	d	b	f	d

由表 2 可以看出，只需进行 4 次循环就匹配成功。BMH2 算法通过提高模式串的平均移动距离，获得了更高的匹配效率。当模式串中没有相同字符或者相同字符的间距较大时，BMH2 算法可以取得更好的匹配效率。

4　测试结果及结论

实验环境采用曙光服务器作为硬件测试平台，其中 CPU 为 AMD Opteron 2.0GHz(双核)，内存为 1.5GB，操作系统采用 Linux 2.6.9 内核，编译器为 pgcc7.0。测试分别使用 BMH 算法和 BMH2 算法进行模式匹配，比较它们的实际效率。每个算法分别执行 1000 次，运行时间取平均值。

测试 1：选取 1M 的纯英文文本，并采用不同长度的英文短语作为模式串，得到的测试数据如表 3 所示。

表 3　纯英文文本的匹配结果对比

	模式串平均每次移动距离(位/次)		检索速率(millions chars/s)		
模式串长度	BMH 算法	BMH2 算法	BMH 算法	BMH2 算法	加速比
5	4.61	4.95	424	431	1.017
6	5.51	5.99	386	422	1.093
7	5.59	6.84	488	535	1.096
8	6.27	7.89	535	602	1.125
9	7.89	8.64	476	552	1.160
10	7.60	9.66	574	694	1.209
15	8.29	12.63	617	806	1.306
20	9.89	14.66	714	870	1.218
25	9.33	17.85	641	847	1.321

注 1：加速比＝BMH2 算法的检索速率/BMH 算法的检索速率。

测试 2：采用 NCBI(美国国家生物情报中心)发布的蛋白质序列文件(4.0M)作为文本，并从中选取不同的氨基酸序列作为模式串，得到的测试数据如表 4 所示。

表 4　蛋白质序列文件的匹配结果对比

	模式串平均每次移动距离(位/次)		检索速率(millions chars/s)		
模式串长度	BMH 算法	BMH2 算法	BMH 算法	BMH2 算法	加速比
5	4.78	4.97	433	446	1.030
10	7.64	9.94	602	676	1.123
15	11.25	14.72	775	862	1.112
20	13.25	19.35	840	935	1.113
25	15.87	22.06	893	971	1.087

在测试 1 中，当模式串长度小于等于 7 时，改进效果不太明显。这是由于模式串较短，任意字符 *ch* 在 *pattern* 中出现的概率都很小，改进前后的平均移动距离都很接近模式串长度，因此检索速率差不多。当模式串长度大于 7 时，BMH2 的模式串平均移动距离明显高于 BMH 算法，因此能够获得更高的匹配效率。从表 3 可以看出，对纯英文文本的检索，BMH2 算法的匹配速率比 BMH 算法平均提高了 15%到 30%。

在测试 2 中，BMH2 算法匹配的加速效果低于纯英文文本的加速效果。根据 BMH2 的算法思想，当模式串中没有相同字符或者相同字符的间距较大时，BMH2 算法可以取得较好的匹配效率。由于组成蛋白质的常见氨基酸只有 20 种，每个氨基酸的缩写为一个字符。氨基酸序列中出现相同字符的概率增加，相同字符间的平均距离减小，使得 BMH2 算法的加速效果减弱。从表 4 可以看出，对蛋白质序列文件的检索，BMH2 算法的匹配速率相比 BMH 算法平均提高了 10%左右。

结束语　模式匹配是当前入侵检测和情报检索系统中普遍采用的策略之一。随着网络带宽的不断增加，网络服务和应用对检索效率要求越来越高。本文对经典的模式匹配算法 BM 以其改进算法 BMH 做了简要的分析，并针对 BMH 算法提出了改进。若 BMH2 算法应用到入侵检测和情报检索等领域，将能够有效提高系统的检索效率。本文的下一步工作可以考虑将 BMH2 思想和其它模式匹配技术相结合，如将 BMH2 思想应用到多模式匹配算法中，使其实际性能更优。

参 考 文 献

[1] 闵联营，赵婷婷. BM 算法的研究与改进[J]. 武汉理工大学学报，2006，30(3)：528-530

[2] Horspool N R. Practical Fast Searching in Strings [J]. Software Practice and Experience, 1980, 10(6): 501-506

[3] Boyer R S, Moore J S. A Fast String Searching Algorithm [J]. Communications of the ACM, 1977, 20(10): 762-772

[4] 巫喜红，凌捷. 单模式匹配算法研究[J]. 网络与通信，2006，22(8)：202-204

[5] 孙克雷. IDS 中一种快速模式匹配算法[J]. 安徽理工大学学报，2006，26(3)：52-55

[6] 李雪莹，刘宝旭，等. 字符串匹配技术研究[J]. 计算机工程，2004，30(22)：24-26