基于可信度的投票法

燕继坤[1),2)] 郑 辉[1)] 王 艳[1)] 曾立君[1),3)]

[1)](西南电子电信技术研究所信号盲处理国家重点实验室 成都 610041)

[2)](信息工程大学电子技术学院 郑州 450004)

[3)](公安海警高等专科学校电子技术系 宁波 315801)

摘 要 可信度投票法不仅使用了基分类器输出的类别,还使用了输出的可信度.推导了该方法训练错误率的界以及期望错误率的界.发现为了最小化期望错误率的界,应该使用错误独立的基分类器,如果基分类器的错误率不是很高,这个界以指数级速度随着基分类器错误率的降低而降低,而且这个界随着投票次数的增加也会下降.在最小化训练错误率的界的意义下,得到了一种权值分配方法.把这个方法应用于一种 Bagging 算法:AB,得到了综合分类算法 CAB.使用 UCI 机器学习数据集中的数据,通过实验验证了 CAB 的有效性.

关键词 机器学习;综合分类;可信度投票法;错误率的界;Bagging

中图法分类号 TP18

Voting by Confidence

YAN Ji-Kun[1),2)] ZHENG Hui[1)] WANG Yan[1)] ZENG Li-Jun[1),3)]

[1)](*State Key Laboratory of Blind Signal Processing, South-West Institute of Electronics & Telecommunication Technology, Chengdu* 610041)

[2)](*Institute of Electronics, Information Engineering University, Zhengzhou* 450004)

[3)](*Department of Electronics, Public Security Marine Police Academy, Ningbo* 315801)

Abstract The method of voting by confidence exploits not only the label but also the confidence outputted by base classifiers. The bound of training error rate is drawn as well as that of expected error rate. It is shown that error-independent base classifiers should been employed in order to minimize the bound of expected error rate. And voting by confidence can achieve expected error rate that decreases exponentially fast with the decreasing of base classifiers' error rate. Also, the bound will decrease with the increasing of voting times if the base classifiers' error rate is not too high. For minimizing the bound of training error rate, a scheme that assigns weights to base classifiers is introduced. An ensemble algorithm named CAB is obtained when applying the scheme to AB, namely a kind of Bagging. Experiments with UCI datasets show the validity of CAB.

Keywords machine learning; ensemble classification; voting by confidence; bound of error rates; Bagging

1 引 言

综合分类(ensemble classification)指利用多个基分类器的输出以得到更好的分类器,这些基分类是为同一个任务而训练出的,每一个都能单独完成分类任务.综合分类在很多应用中都表现出很好的性能,是目前机器学习最重要的研究方向之一[1~3].

收稿日期:2004-02-10;修改稿收到日期:2005-05-11.本课题得到国家预研基金(514950307)资助.燕继坤,男,1973 年生,博士,讲师,研究方向为模式识别与机器学习.E-mail: yan_jk@sina.com.郑 辉,男,1957 年生,教授,博士生导师,研究方向为信号与信息的智能处理、信号盲处理.王 艳,女,1978 年生,工程师,研究方向为模式识别及语音处理.曾立君,男,1973 年生,硕士研究生,讲师,研究方向为机器学习在网络安全中的应用.

综合多个基分类器最简单的方法是对它们的输出进行线性加权，这称为投票法. 投票法给每个基分类器分配权值，权值反映各基分类器对最终结果的影响程度. 如果各基分类器的权值相同，则称为简单投票法.

基分类器的性能体现在两个方面：一是可靠性；二是稳定性. 大部分权值分配方法都试图给可靠性高的基分类器分配较大的权值，例如，Xu 用从训练集中计算出的信任度(belief)给各基分类器分配权值[4]，孙怀江用相关证据模型给基分类器赋权值[5]. Perrone 认为当基分类器无偏且互不相关时，在均分误差最小的意义下，权值应与基分类器的方差成反比[6]，方差反映分类器的稳定性，这个方法实际上是给稳定的基分类器以更大的权值.

在分类时，有些基分类器只能输出类别，例如 C4.5，但多数基分类器除了输出类别，还可以输出该类别的可信度，例如 Bayesian、kNN、基于距离的分类器、SVM、ANN. 如果只利用类别进行投票在直观上是不合理的：如果两个基分类器输出相同的类别，但一个可信度很高，另一个可信度很低，表决时它们就不应起到相同的作用. 所以实际上有很多投票法不仅使用类别，还使用类别的可信度[1,2,4]，我们把这种投票法称为可信度投票法.

Bagging 和 AdaBoost 是两种最重要的综合分类方法，都使用了投票法. Bagging 的核心是通过对训练实例的重抽样形成多个基分类器，用简单投票法综合这些基分类器[7]. AdaBoost 给每个训练实例维持着一个权值(不是投票时使用的权值)，以级联的方式训练多轮. 每一轮中由训练实例及相应的权值训练出一个基分类器，再根据该基分类器对每个训练实例分类的结果调整其权值，增加错分实例的权值，这样迫使下一轮训练出的基分类器更关注于当前基分类器失败的实例. AdaBoost 按照基分类器在训练集上的准确率分配权值，用投票法进行综合[8].

Matan 给出了简单投票法期望错误率的界[9]，但这个结论并不能适用于一般的可信度投票法. 我们推导出可信度投票法错误率的界，并在最小化训练错误率的界的意义下给出了一种新的权值分配方法. AB(属性 Bagging；Attribute Bagging)是一种新的 Bagging 算法，用属性重抽样代替实例重抽样[10]. 作为可信度投票法的一个应用，把我们的权值分配方法应用于 AB 得到一种称为 CAB(Confidence-based AB)的综合分类算法. 通过 UCI 机器学习数据集中的一些数据用实验比较了 CAB 与 AB 的性能.

2　可信度投票法

我们只研究最简单的两分类问题. 设实例空间为 $\aleph$，类别空间为 $\Re=\{-1,1\}$. 训练集为 $S=\{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}$，其中 $x_i\in\aleph$，$y_i\in\Re$，S 中的元素按照某个分布 D 独立地抽取得到. 用 T 个基分类器进行投票，第 t 个基分类器为 $h_t(x)$，它的符号表示类别，绝对值表示可信度. 分配给 $h_t(x)$ 的权值为 α_t，投票函数为 $H(x)=\mathrm{sign}(f(x))$，其中 $f(x)=\sum_{t=1}^{T}\alpha_t h_t(x)$，$\sum_{t=1}^{T}\alpha_t=1$. 我们首先给出可信度投票法训练错误率的界.

2.1　可信度投票法训练错误率的界

引理 1.　可信度投票法对训练集 S 的训练错误率满足下面的界：

$$\frac{1}{m}\left|\{i:H(x_i)\neq y_i\}\right|\leq\frac{\prod_{t=1}^{T}Z_t}{m}\qquad(1)$$

其中 $Z_t=\sum_{i=1}^{m}\mathrm{e}(-\alpha_t y_i h_t(x_i))$，$m$ 为 S 中训练样本的个数.

引理 1 的证明见附录 A. 式(1)左边为训练错误率，右边的 Z_t 衡量了基分类器 h_t 由 α_t 加权的训练错误率的大小，h_t 的错误率越小，Z_t 也越小.

我们还给出另外一个较松的界，但这个界可以用于估计简单投票法的期望错误率，而且可进一步由它导出可信度投票法期望错误率的界.

引理 2.　可信度投票法对 S 的训练错误率还满足下面的界：

$$\frac{1}{m}\left|\{i:H(x_i)\neq y_i\}\right|\leq\prod_{t=1}^{T}G_t\qquad(2)$$

其中 $G_t=\left(\frac{1}{m}\sum_{i=1}^{m}\exp(-y_i h_t(x_i))\right)^{\alpha_t}$.

引理 2 的证明见附录 B. 式(2)中的 G_t 同样也反映了基分类器 h_t 错误率的大小，但形式与 Z_t 不同.

2.2　可信度投票法的期望错误率

如果已知基分类器的错误率，可以用引理 2 粗略地估计出简单投票法期望错误率的界. 设 $h_t(x_i)\in\{-1,1\}$，G_t 中的 $\sum_{i=1}^{m}\exp(-y_i h_t(x_i))$ 可分解为两项：

$$\sum_{i=1}^{m}\mathrm{e}(-y_i h_t(x_i))=\sum_{i:\,y_i\cdot f(x_i)=1}\mathrm{e}(-y_i h_t(x_i))+\sum_{i:\,y_i\cdot f(x_i)=-1}\mathrm{e}(-y_i h_t(x_i))$$

$$= \sum_{i;\, y_i \cdot f(x_i)=1} e^{-1} + \sum_{i;\, y_i \cdot f(x_i)=-1} e \tag{3}$$

设所有 $h_t(x)$ 的错误率都为 p_e，用 $|\cdot|$ 表示集合中元素的数目，当 m 比较大时有

$$\begin{cases} |\{i: y_i \cdot h_t(x) = 1\}| = (1-p_e)\cdot m \\ |\{i: y_i \cdot h_t(x) = -1\}| = p_e \cdot m \end{cases} \tag{4}$$

把式(4)代入式(3)，考虑到式(2)中 G_t 的定义及 $\sum_{t=1}^{T}\alpha_t = 1$，得到

$$\prod_{t=1}^{T} G_t = \prod_{t=1}^{T} \left(\frac{m(1-p_e)e^{-1} + mp_e e}{m}\right)^{\alpha_t} = e^{-1}(1-p_e) + p_e e \tag{5}$$

考虑到引理 2，我们有

$$\frac{1}{m}|\{i: H(x_i) \neq y_i\}| \leq e^{-1} \cdot (1-p_e) + e \cdot p_e \tag{6}$$

如果 S 是任取的，上面的推导仍然成立，所以式(6)粗略地给出了简单投票法期望错误率的上界. 这个界并不紧，当 $p_e = 0$ 时，这个界为 e^{-1}，还不能取到 0. 但它反映了一个直观的事实：基分类器错误率越低，简单投票法的错误率也越低.

记 $e_S = \frac{1}{m}|\{i: H(x_i) \neq y_i\}|$，下标 S 表示 e_S 是由训练集 S 求出的. 对式(2)两边求期望可得到可信度投票法期望错误率的界：

$$P_{x\sim D}[H(x) \neq y] = E_{S\sim D}(e_S) \leq E\left(\prod_{t=1}^{T} G_t\right) \tag{7}$$

如果基分类器 $h_t(x)$ 的错误互相独立，则 G_t 互相独立，式(7)的右端取到最小值 $\prod_{t=1}^{T} E(G_t)$. 这说明为了得到更好的投票效果应该使各个基分类器的错误互相独立. G_t 中反映基分类器错误率的 $-y_i h_t(x_i)$ 出现在指数部分，所以可信度投票法期望错误率的界随着基分类器错误率的降低以指数级速率降低. 如果基分类器的错误率低到足以使 $G_t < 1$（对基分类器的这个要求并不高），那么投票次数越多（T 越大），这个界也越低.

Breiman 指出 Bagging 只对不稳定的分类算法有效[7]. 使用重抽样时，不稳定的分类算法训练出不同的基分类器，各基分类器的错误不相关. 所以重抽样只是获得错误独立的基分类器的一种方法，对 Bagging 而言重抽样并不是本质的，对错误独立的基分类器的投票才是本质的. Chawla 通过实验也指出了这一点[11].

2.3 可信度投票法训练实例权值的选取方法

我们通过最小化引理 1 中训练错误率的界来确定基分类器 $h_t(x)$ 的权值 α_t. 用 Schapire 发展的技术可以证明定理 1[8].

定理 1. 如果 S 是训练集，当 $\alpha_t = \frac{1}{2}\ln\left(\frac{1+r_t}{1-r_t}\right)$，$t = 1, 2, \cdots, T$ 时，可信度投票法的训练错误率不大于

$$\frac{\prod_{t=1}^{T}\sqrt{1-r_t^2}}{m} \tag{8}$$

其中 $r_t = \sum_{i=1}^{m} y_i h_t(x_i)$.

证明. 下面的叙述中在不引起误解的情况下，有时略去下标 t. 我们只需最小化式(1)中的 Z 即可. 令 $u_i = y_i h_t(x_i)$，$\exp(-\alpha u_i)$ 是凸函数，所以有

$$Z = \sum_{i=1}^{m}\exp(-\alpha u_i) \leq \sum_{i=1}^{m}\left(\frac{1+u_i}{2}e^{-\alpha} + \frac{1-u_i}{2}e^{\alpha}\right) \tag{9}$$

式(9)右端对 α 求导，取导数为 0，可得到当 $\alpha = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$ 时，式(9)右端取到最小值 $\sqrt{1-r^2}$，代入式(1)的右端，即可证明式(8). 证毕.

定理 1 没有考虑 $\sum_{t=1}^{T}\alpha_t = 1$ 的限制，实际上只要做简单的归一化即可. 因为投票函数 $H(x)$ 的符号表示类别，并不关心绝对值，所以归一化可以省略.

AdaBoost 使用了投票机制[8]，给基分类器分配的权值 α_t 具有和定理 1 相同的形式，但 AdaBoost 的 $r_t = \sum_{i=1}^{m} w_t(i) y_i h_t(x_i)$，其中 $w_t(i)$ 是指定给训练实例 x_i 的权值. 可信度投票法相当于给每个训练实例分配了相同的权值 1. AdaBoost 使用 $w_t(i)$ 使难以正确分类的实例对 r_t 的影响更大，当训练实例没有类别噪声时有更好的性能，但当类别噪声存在时给基分类器分配的权值 α_t 就不合理，此时如果给每个训练实例分配相同的权值会有更好的效果. Dietterich 通过实验对比了几种综合分类方法，也指出了这一事实[12]. 所以与 AdaBoost 相比，可信度投票法有更好的抗类别噪声能力.

上面的 α_t 是在最小化训练错误率的界的意义下得到的，实际上更关心期望错误率. Schapire 给出了投票法的期望错误率与训练错误率的关系[8]：

定理 2(Schapire). 设 S 是含 m 个实例的训练集，其中的实例 x 按照某个分布 D 独立地选取. 设基分类器空间的 VC 维为 d，且令 $\delta > 0$. 那么任意

投票函数 f 以至少 $1-\delta$ 的概率，对于所有的 $\theta>0$ 满足

$$P_{x\sim D}[yf(x)]\leq P_{x\in S}[yf(x)<\theta]+O\left(\frac{1}{\sqrt{m}}\left(\frac{d\log^2(m/d)}{\theta^2}+\log\frac{1}{\delta}\right)^{\frac{1}{2}}\right) \quad (10)$$

式(10)的左边是期望错误率，右边第一项是训练错误率，第二项反映了实例个数、基分类器空间等的影响. 如果减少训练错误率，就减少了式(10)的右端，因而减少了期望错误率的上界. 我们的方法实际是通过减少训练错误率来得到更小的期望错误率.

3　CAB 算法

Bryll 提出一种新的 Bagging 算法 AB[10]. 设实例 x_i 用 d 个属性表示，属性集 A 为 $\{a_1,a_1,\cdots,a_d\}$. AB 的训练共进行 T 轮，在每一轮对属性集进行有放回的重抽样，这样有的属性可能出现多次，有的属性可能一次也不出现. 根据每一轮抽样出的属性集由原始训练实例得到新的训练实例，新训练实例与原始训练实例一样多，这一点也与传统 Bagging 不同. 对每轮的训练实例集用基分类算法训练出一个基分类器，对未知实例用 T 个基分类器的结果进行简单投票表决.

kNN 是一种古老而有效的分类算法. Breiman 指出 Bagging 对于稳定的学习算法无效，kNN 是稳定的，因而 Bagging 并不能增强 kNN 的性能[7]. 但 kNN 对于属性的增减很敏感，所以有研究者用 AB 来增强 kNN 的性能[13].

我们把 AB 的简单投票法改为可信度投票法，基分类器的权值 α_t 按定理 1 选取，得到的新算法称为基于可信度投票的属性 Bagging(CAB)，算法 1 给出了 CAB. CAB 执行 T 轮，每一轮中从原始属性集中随机抽取 n 个属性，原始训练集中的实例只选取这些属性得到该轮的训练集，用每轮不同的训练集训练出不同的基分类器，用可信度投票法综合多个基分类器的结果.

算法 1.　CAB.

1. 确定重抽样出的属性集大小 n，迭代次数 T 和基分类算法 h.

2. For $t=1,2,\cdots,T$ 执行以下 3 步

2.1 从 A 中重抽样(有放回) n 个属性得到属性集 A_t，S 中每个实例的属性只取 A_t 得到 S_t；

2.2 由 S_t 用 h 训练出基分类器 h_t；

2.3 计算 $\alpha_t=\frac{1}{2}\ln\left(\frac{1+r_t}{1-r_t}\right)$；

3. 对于测试样本 x，$H(x)=\operatorname{sign}\left(\sum_{t=1}^{T}\alpha_t h_t(x)\right)$.

作为一个例子，我们仍然用 kNN 作为基训练算法. 设 kNN 输出的 k 个近邻中有 n_+ 个正例，n_- 个反例，把可信度定义为 $\frac{|n_+-n_-|}{k}$.

4　实验结果

为了验证 CAB 算法，我们选取 UCI 机器学习数据集中的 5 个数据进行实验①，这 5 个数据的情况列于表 1. 表 1 中“正例、反例”一列给出了每个数据正类和反类的选择方法，“训练集、测试集”一列给出了训练集和测试集的选择方法. 实验中重抽样属性集的大小 n 选为与原始属性集相同，即表 1 中的“属性个数”一列给出的数值. kNN 的 k 取为 5. 为了对比也对 AB 在同样的数据上用同样的参数进行了实验.

表 1　UCI 5 个实验数据的情况

数据	正例、反例	训练集、测试集	属性个数
pima indinas diabetes	1 为正例，0 为反例	前 300 个实例用于训练，后 368 个实例用于测试	8
ionosphere	good 为正例，bad 为反例	前 100 个实例用于训练，后 251 个用于测试	34
sonar	mine 为正例，rock 为反例	原始数据已做了划分	60
liver disorder	1 为正例，0 为反例	奇数实例用于训练，偶数实例用于测试	6
new thyroid	normal 为正例，hyper 和 hypo 为反例	奇数实例用于训练，偶数实例用于测试	5

对每个数据集重复进行 50 次实验，求出错误率的均值和标准差，标准差反映分类器的稳定性，实验结果列于表 2. 可以看出在 5 个数据上 CAB 得到的错误率都低于 AB，除 new thyroid 外(表中用黑体表示)，CAB 的方差也低于 AB. CAB 较低的错误率和较小的方差说明可信度投票法有比简单投票更好的性能.

表 2　UCI 5 个数据集的实验结果(错误率)

数据集	CAB		AB	
	均值	标准差	均值	标准差
pima indinas diabetes	0.222	0.005	0.235	0.008
ionosphere	0.230	0.007	0.250	0.013

① Blake C. L., Merz C. J.. UCI Repository of machine learning databases[http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science, 1998

（续 表）

数据集	CAB		AB	
	均值	标准差	均值	标准差
sonar	0.167	0.009	0.179	0.011
liver disorder	0.336	0.015	0.350	0.015
new thyroid	0.060	**0.009**	0.107	0.005

5 结 论

我们给出了可信度投票法训练错误率的界以及期望错误率的界.期望错误率的界表明为了使投票法有效,必须使用错误独立的基分类器,此时可信度投票法期望错误率的界以指数级速度随着基分类器错误率的降低而降低,而且当基分类器错误率不是特别高时,随着投票次数的增多,这个界也会降低.

在最小化可信度投票法训练错误的界的意义下,我们导出了一种基分类器的权值分配公式.

把我们的可信度投票法应用于 AB 得到新的综合分类算法:CAB.以kNN作为基分类器,发现 CAB 有比 AB 更好的性能.

尽管我们用可信度投票法改造 AB,实际上可信度投票法可应用于其它使用投票机制的分类器中.

参 考 文 献

1 Comay O., Intrator N.. Ensemble training: Some recent experiments with postal zip data. In: Basri R., Schild U. J., Stein Y. eds.. Proceedings of the 10th Israeli Conference on AICV. Amsterdam: Elsevier, 1993, 201～206

2 Sharkey A. J. C., Sharkey N. E., Gerecke U., Chandroth G. O.. The test and select approach to ensemble combination. In: Kittler J., Roli F. eds.. Proceedings of the 1st International Workshop on Multiple Classifier Systems. Berlin: Springer-Verlag, 2000, 30～34

3 Dietterich T.. Machine learning research: Four current directions. Artificial Intelligence, 1997, 4(18): 97～136

4 Xu L., Krzyzak A., Suen C. Y.. Methods of combining multiple classifiers and their application to handwriting recognition. IEEE Transactions on System, Man, and Cybernetics, 1992, 22(3): 418～435

5 Sun Huai-Jiang, Hu Zhong-Shan, Yang Jing-Yu. A study on combining multiple classifiers based on evidence theory. Chinese Journal of Computers, 2001, 24(3): 231～235(in Chinese)
(孙怀江,胡钟山,杨静宇.基于证据理论的多分类器融合方法研究.计算机学报, 2001, 24(3): 231～235)

6 Perrone M. P.. Improving regression estimation: Averaging methods for variance reduction with extensions to general convex measure optimization[Ph. D. dissertation]. Department of Physics, Brown University, 1993

7 Breiman L.. Bagging predictor. Machine Learning, 1996, 26(1): 5～24

8 Schapire R., Singer Y.. Improved boosting algorithms using confidence-rated predictions. Machine Learning, 1999, 37(3): 297～336

9 Matan O.. On voting ensembles of classifiers. In: Proceedings of the Working Notes of Integrating Multiple Learned Models (IMLM-96), Portland, OR, 1996, 84～88

10 Bryll R., Gutierrez O. R., Quek F.. Attribute Bagging: Improving accuracy of classifier ensembles by using random features subsets. Pattern Recogniton Letters, 2003, 36(6): 1291～1302

11 Chawla N. V., Moore T. E., Hall L. O.. Distributed learning with Bagging-like performance. Pattern Recogniton Letters, 2003, 24(1～3): 455～471

12 Dietterich T. G.. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, Boosting, and randomization. Machine Learning, 2000, 40(2): 139～158

13 Bay S. D.. Combining nearest neightbor classifiers through multiple feature subsets. In: Proceedings of the 17th International Conference on Machine Learning, Madison, WI, 1998, 37～45

附录 A

证明. 如果 $H(x_i)\neq y_i$,则 $f(x_i)\cdot y_i\leq 0$,即$\exp(-y_i\cdot f(x_i))\geq 1$.我们有

$$1(H(x_i)\neq y_i)\leq \exp(-y_i\cdot f(x_i)) \tag{1}$$

其中 $1(\cdot)$表示这样的函数:当括号中条件成立时取 1,否则取 0.根据式(1)有

$$\begin{aligned}|\{i: H(x_i)\neq y_i\}| &\leq \sum_{i=1}^{m}\exp(-y_i f(x_i))\\ &= \sum_{i=1}^{m}\exp\Big(-y_i\sum_{t=1}^{T}\alpha_t h_t(x_i)\Big)\\ &= \sum_{i=1}^{m}\prod_{t=1}^{T}\exp(-\alpha_t y_i h_t(x_i))\\ &\leq \prod_{t=1}^{T}\sum_{i=1}^{m}\exp(-\alpha_t y_i h_t(x_i)) \end{aligned}\tag{2}$$

证毕.

附录 B

证明. 根据引理 1 的证明思路有

$$\begin{aligned}\frac{1}{m}|\{i: H(x_i)\neq y_i\}| &\leq \frac{1}{m}\sum_{i=1}^{m}\prod_{t=1}^{T}\exp(-\alpha_t y_i h_t(x_i))\\ &= \frac{1}{m}\sum_{i=1}^{m}\prod_{t=1}^{T}(\exp(-y_i h_t(x_i)))^{\alpha_t}\end{aligned}\tag{3}$$

考虑到$\sum_{t=1}^{T}\alpha_t=1$,根据赫德尔不等式有

$$\sum_{i=1}^{m}\prod_{t=1}^{T}(\exp(-y_i h_t(x_i)))^{\alpha_t}\leq\prod_{t=1}^{T}\Big(\sum_{i=1}^{m}\exp(-y_i h_t(x_i))\Big)^{\alpha_t}\tag{4}$$

把式(4)代入式(3)即得. 证毕.

YAN Ji-Kun, born in 1973, Ph. D., lecturer. His research interests include pattern recognition and machine learning.

ZHENG Hui, born in 1957, professor, Ph. D. supervisor. His main research interests include intelligent signal and information processing, blind signal processing etc.

WANG Yan, born in 1978, engineer. Her research interests include pattern recognition and speech processing.

ZENG Li-Jun, born in 1973, lecturer, master candidate. His research interest is the application of machine learning in computer network security.

Background

The project "Filtering Technique for Documents and Document Images" is intended to find machine learning algorithm to classify interested documents and document images from large quantities of uninterested ones based on their content. The work is part of this project, which focuses on ensemble learning. One of the challenge that filtering task usually faced to is imbalanced class, where the negative class is represented by a large number of examples, while the positive class is represented by only a few, and the task demands to classify important but rare positive examples.

In the authors' previous works, they have succeeded to find way to convert imbalanced class to balanced class via Bagging technique. For voting is key component of Bagging, the purpose of the work is to further improve Bagging's performance by research on voting mechanism. The voting algorithm's upper bound of training error rate is drawn as well as that of expected error rate. Furthermore, a scheme assigning weights to base classifier is introduced for voting by confidence.