权重自适应调整的多分类器集成判决及其在文本分类中的应用*)

唐春生　金以慧

(清华大学自动化系　北京100084)

A Multiple Classifiers Integration Method Based on Adaptive Weigh Adjusting and it's Application on Text Classification

TANG Chun-Sheng　JIN Yi-Hui

(Department of Automation, Tsinghua University, Beijing 100084)

Abstract Multiple classifier systems based on the combination of a set of different classifiers are adopted to achieve high pattern-recognition performances. A multiple classifiers integration method based on adaptive weight adjusting is presented in this paper. The useful neighbors are selected from training set by analyzing the pending pattern's character, then each classifier's weight can be determined automatically by analyzing the performance of the classifier on the useful neighborhood set. The final output of the multiple classifiers systems is the effective integration of each calssifier's result. The effectiveness of the method is proved by the text classification experiments of the Reuters-21578 text sets.

Keywords Combination of multiple classifiers, Multiple classifiers integration, Text classification

1　引言

近年来，多分类器的组合方法已成为模式识别研究的热点问题，并已在模式识别的多个应用方面，如字符识别、目标识别、文本分类等领域获得了较好的应用效果[1~3]。多分类器组合方法的基本假设是：对一个需要专家进行的任务，k 个专家个人判断的有效组合应该优于个人的判断[3]。利用具有不同特性和性能的多分类器，通过进行有效的组合可以获得更高的模式识别性能。

多分类器的组合方式可以分为级联和并联两种方式。采用级联形式时前一级分类器为后一级分类器提供分类信息，指导下一级分类器的训练和分类过程。Boosting 方法是级联形式的多分类器组合方法中的代表；并联形式中，各分类器是独立设计的，其核心是找到一种合适的组合准则来将各分类器的输出综合起来形成最终的结果。并联方式的组合方法包括传统的择多判决法(如投票表决法、计分法等)，线性加权组合方法，模糊推理法，将分类结果作为一种新的输入特征的神经网络组合方法等[1~3]。

在进行多分类器组合时，认识到各分类器对不同样本的分类效果是不同的，同一分类器在样本空间的不同区域性能会有变化，而且分类器输出的不同候选类别与实际类别之间存在一定的相似性，对最终判决有一定支持作用。

基于以上认识，本文提出了一种权重自适应调整的多分类器集成判决方法，通过分析待定样本的特征，寻找其在训练集中的邻居，分析各分类器在这些邻域样本上的行为，自动地调整邻域范围，并根据各分类器在邻域中的统计结果，为分类器的集成判决提供依据，而最终结果为各分类器集成判决的综合，其权重根据分类器在该邻域中的分类准确率自动确定。通过在 Reuters-21578文本集上的文本分类实验，证明了该方法的有效性。

2　权重自适应调整的多分类器集成判决方法

2.1　问题描述

对于一个有 M 个数据类的分类问题，其 $\tilde{\omega}_j, \forall\, j\in\Lambda=\{1,\cdots,M\}$ 称为一个类。每一类别代表一组相似的样本，样本可以用一个特征向量 X 来表示，下文中对样本和其特征向量表示将不再区分。分类任务即是判断样本 X 的类别所属，可以用样本属于各类的可能性来表示。假设分别训练了 L 个不同的分类器 $C_i, i=1,\cdots L$，来完成分类任务，则可以用 $e_i(X)=(e_{i1},\cdots,e_{ij},\cdots,e_{iM})$ 来表示分类器 C_i 对样本 X 的分类输出，其中 e_{ij} 表示在分类器 C_i 下样本 X 属于类 $\tilde{\omega}_j$ 的可能性。多分类器组合方法即是寻找一种合适的组合准则，将各分类器的输出结果进行有效综合。

2.2　关键问题及其解决方法

不同待定样本具有不同的特征，在样本空间中也处于不同的区域，而各分类器对不同类别和不同区域的样本的分类效果是有差别的，需要针对不同样本选择一组合适的分类器，并进行恰当的加权来对分类结果进行综合。本文提出的权重自适应调整的多分类器集成判决方法正是基于以上思想来设计与实现的。在这种方法中，需要处理以下一些关键问题：

(1)如何自动地判别待定样本的有效邻域？

(2)如何进行分类结果的集成判决？

(3)如何自适应地确定各分类器的权重？

2.2.1　*多分类器行为分析及样本有效邻域的判定*　待定样本的邻域是指与之近邻的一组训练样本，通过判断分类器在邻域上的分类性能，可以初步确定合适的分类器及其权值。在此邻域中，通常会有这样一些样本，各分类器在这些样本上的分类判断一致性不够，这些样本的存在对分类器的有效组合起到了干扰作用，因此需要剔除。

利用多分类器的行为分析可以有效剔出这些无效的邻域

*)本研究得到北京市科委科技项目基金资助，基金编号为：2001－0075。唐春生　博士研究生，研究方向：数据挖掘、超文本/超媒体技术、中文信息处理技术、agent 技术。金以慧　教授，博导。

样本，从而自动确定待定样本的有效邻域。如果需要给样本指定为某一个类，通常可以将样本划归为可能性最大的那个类，即 $C_i(X)=\arg\max_j(e_{ij})$，其中 $\arg\max_j(x_j)$ 是取使 x_j 最大的 j，这说明分类器将样本划归为 $C_i(X)$ 类。从另一个角度看，$C_i(X)\in\{1,\cdots,M\}$ 可以看作分类器 C_i 在样本 X 上的行为[4]。对于 L 个分类器，多分类器在样本 X 上的行为(Multiple Classifier Behaviour，MCB)可以定义为一个向量：$MCB(X)=\{C_1(X),C_2(X),\cdots,C_L(X)\}$。对于样本 X 和 Y，可以定义两个样本的 MCB 之间的相似度为：

$$S(X,Y)=\frac{1}{L}\sum_{i=1}^{L}T_i(X,Y) \tag{1}$$

其中 $T_i(X,Y),i=1,\cdots,L$ 定义为：

$$T_i(X,Y)=\begin{cases}1 & if\ C_i(X)=C_i(Y)\\ 0 & if\ C_i(X)\neq C_i(Y)\end{cases} \tag{2}$$

两个样本的 MCB 之间的相似度 $S(X,Y)$ 的取值范围为[0,1]，当取值为1时，说明同一分类器都认为两个样本属于同一类；当取值为0时，说明每一个分类器都将两个样本判为不同的类；$S(X,Y)$ 越接近1，说明分类器在样本上的分类判断越一致。

确定待定样本有效邻域的步骤如下：

(1)利用 k 近邻法在训练集中寻找待定样本的 k 个近邻；

(2)计算这 k 个近邻和待定样本的 MCB 值；

(3)计算 k 个近邻与待定样本的 MCB 值之间的相似度；

(4)剔出相似度小于阈值的近邻，剩余的近邻构成待定样本的有效邻域。

2.2.2 *分类结果的集成判决* 其思想由肖旭红在文[1]中提出，认为由分类器输出的不同候选类别与待识别样本之间必然存在某些相似性，对该样本的识别起着一定的支持作用，并认为它们之间的关系可以用一个线性模型来近似：

$$A_i(x,j,k,l)=m_i(x,l,k)\times R_i(j,l)\times W_k \tag{3}$$

其中 $A_i(x,j,k,l)$ 表示分类器 C_i 的第 k 阶候选(类别为 $\tilde{\omega}_l$)对集成判决 $X\in\tilde{\omega}_j$ 的支持作用，$m_i(x,l,k)$ 为分类器 C_i 下样本 X 与类别 $\tilde{\omega}_l$ 之间的相似度，$R_i(j,l)$ 表示分类器 i 将类别 $\tilde{\omega}_j$ 中的样本判为属于类别 $\tilde{\omega}_l$ 的可能性，W_k 为支持因子，与候选阶次有关，阶次越低，该因子越大。公式(3)表明：分类器 C_i 将类别 $\tilde{\omega}_j$ 中的样本判为属于类别 $\tilde{\omega}_l$ 的可能性越大，对判决 $X\in\tilde{\omega}_j$ 的支持作用越大；待定样本 X 与类 $\tilde{\omega}_l$ 越相似，对判决 $X\in\tilde{\omega}_j$ 的支持作用也越大。

分类器 C_i 对样本 X 与类别 j 之间的相似度集成判决可用 $A_i(x,j,k,l)$ 的线性组合来表示，即

$$M_i(x,j)=\sum_{l=1}^{M}A_i(x,j,k,l) \tag{4}$$

而多分类器的集成判决为各分类器的集成判决的线性加权组合。其权重将通过评价分类器在邻域内的分类准确率来确定。

在分类器的集成判决中，$R_i(j,l)$ 可以通过对分类器的识别情况进行统计后得到，一般用分类器的混乱矩阵来表示：

$$CM_i=\begin{bmatrix}r_{11}^{(i)} & r_{12}^{(i)} & \cdots & r_{1M}^{(i)}\\ r_{21}^{(i)} & \ddots & & \vdots\\ \vdots & & r_{jl}^{(i)} & \vdots\\ r_{M1}^{(i)} & \cdots & \cdots & r_{MM}^{(i)}\end{bmatrix} \tag{5}$$

其中 $r_{jl}^{(i)}$ 表示分类器 C_i 将类别 $\tilde{\omega}_j$ 中的样本识别为 $\tilde{\omega}_l$ 类的数量。$R_i(j,l)$ 可用归一化后的 $r_{jl}^{(i)}$ 来近似。

2.2.3 *分类器权重的确定* 在本文的方法中，各分类器的权重是随着待定样本的特征和区域不同而变化的，因此需要让分类器权重跟随着不同样本进行自适应的调整。这里采用的方法是评价分类器在待定样本的有效邻域中的分类准确率。按照准确率来给予不同的权重，准确率越高，分类器权重越大。准确率的计算可以通过构造有效邻域的混乱矩阵后进行，计算公式如下：

$$Acc_i=\sum_{j=1}^{M}r_{jj}^{(i)}/\sum_{j=1}^{M}\sum_{l=1}^{M}r_{jl}^{(i)} \tag{6}$$

2.3 权重自适应调整的多分类器集成判决算法

权重自适应调整的多分类器集成判决的算法可描述如下：

(1)读入待定样本；

(2)对给定的待定样本 X^*，首先在训练集或校验集中寻找它的 k 个最近邻；

(3)利用多分类器行为分析方法计算待定样本和邻域样本的 MCB 值；

(4)如果待定样本的 MCB 中的各项值都相同，说明各分类器对待定样本的类别判断一致，此时输出该类别，并转到第(11)步，否则继续；

(5)计算待定样本和邻域样本的 MCB 值之间的相似度来确定待定样本的有效邻域；

(6)构造各分类器在有效邻域上的混乱矩阵；

(7)计算各分类器在有效邻域上的准确率，并得到准确率最高的分类器；

(8)如果最大准确率与某分类器准确率之间的差值大于阈值，则该分类器权重为0，即在这次组合过程中不考虑该分类器，否则按准确率大小给剩余的分类器分配权重；

(9)利用集成判决算法得到各分类器的集成判决；

(10)利用线性加权方法综合各分类器的集成判决，得到待定样本的类别输出，其中权重由第8步中得到；

(11)如果已处理完所有待定样本，则结束，否则转到(1)。

3 权重自适应调整的多分类器集成判决方法在文本分类中的应用

3.1 测试集和分类方法

在文档分类领域，Reuters-21578文章集是常用的测试平台[5]，本文的实验也是在该文章集上进行。实验中采用了“ModApte”切分，文集中包括90类，选择文章最多的10类进行测试，其中包括了7194个训练文档和2788个测试文档。

本文的算法中组合了文本分类中常用的方法，包括 KNN 方法、SVM 方法、Naïve Bayes 方法和 TFIDF 方法。这些方法一般都将文档表示为向量形式[6,7]。

利用全局准确率来确定分类器权重，这可以看作一种权重不变的多分类器的线性加权方法；而利用局部准确率来确定分类器权重，则可以得到一种权重自适应的多分类器线性加权方法。本文进行了权重自适应调整的多分类器集成判决方法与上述方法的对比实验。在实验中主要考察了这些方法的准确率、查全率和 F_1 指标[3,6]。这些参数定义如下：

对于类 c_i，其混乱矩阵为

类 c_i		分类器判断	
		YES	NO
专家判断(实际类别)	YES	TP_i	FN_i
	NO	FP_i	TN_i

则类 c_i 的准确率和查全率可用下面的公式来估计：

$$\hat{Pr}_i=TP_i/(TP_i+FP_i) \tag{7}$$

$$\hat{Re}_i=TP_i/(TP_i+FN_i) \tag{8}$$

F_β 函数定义为：

$$F_\beta=\frac{(\beta^2+1)\cdot Pr\cdot Re}{\beta^2\cdot Pr+Re} \quad (9)$$

当β设为1，即查全率和准确率的权重相等时，得到F_1度量，

$$F_1=2Pr\cdot Re/(Pr+Re) \quad (10)$$

3.2 实验结果及比较

表1给出了这些方法在测试集上的实验结果。

表1 各种分类方法在Reuters-21578(10类)上的实验结果

分类方法	准确率	查全率	F_1指标
KNN	0.828	0.785	0.806
SVM	0.867	0.83	0.848
Naive Bayes	0.82	0.815	0.817
TFIDF	0.812	0.805	0.808
权重不变的线性加权	0.847	0.812	0.829
权重不变的集成判决	0.841	0.83	0.836
权重自适应的线性加权	0.858	0.829	0.843
权重自适应的集成判决	0.874	0.846	0.858

比较实验结果，可以得出以下几点结论：

(1)比较多分类器组合与单个分类器的分类结果可以看出，多分类器的有效组合可以取得较好的分类效果。多分类组合的分类查准率、查全率和F_1指标普遍高于单个分类器(SVM除外)。本文的多分类组合方法主要在查全率方面有较大提高，从而从整体上提高了分类性能。

(2)比较权重不变(全局准确率)的方法和权重自适应(局部准确率)的方法，可以看出针对不同待定样本的特征和分布区域来自适应地选择分类器组合及其权重，可以从一定程度上提高分类性能。

(3)比较集成判决和线性加权方法的分类结果，发现集成判决方法可以从整体上(F_1指标)提高分类性能，其查全率也有一定提高。

(4)本文提出的权重自适应调整的多分类器集成判决方法的分类效果最好，该方法采用了基于待定样本的特征分析和局部准确率的权重自适应调整技术，使得分类器组的选择和加权更有针对性；利用分类器在样本集上的统计信息来获取分类之间的关系，这有助于解决一个样本属于多个分类的问题，从而使得组合分类器的整体性能更佳。

小结 随着问题的复杂度增加，模式识别中的多分类器组合方法得到了更多的关注并成为研究的热点。多分类器组合的关键是寻找一种合适的组合准则，能将各分类器的结果有效地综合起来。

本文提出了一种权重自适应调整的多分类器集成判决方法，针对不同待定样本自动选择不同的分类器组及其权重，从而发挥各分类器在不同样本和不同区域上的分类优势；利用样本集上的统计信息来描述类别之间的关系，从而指导分类结果的集成判决，最终从整体上提高了分类的准确率和查全率，提高了分类性能。这在Reuters-21578文本集上的对比实验中得到了验证。

文中的方法只是在标准文本集合上进行了实验，在解决实际问题时，需要做有针对性的处理，并需要做进一步的优化和调整；另外，可以考虑引入控制理论中的思想，实现更好的自适应调节。

参考文献

1 肖旭红，戴汝为. 一种识别手写汉字的多分类器集成方法. 自动化学报，1997，23(5)：621～627

2 荆晓远，杨静宇. 基于相关性和有效互补性分析的多分类器组合方法. 自动化学报，2000，26(6)：741～747

3 Sebastiani F. A tutorial on automated text categorization. In: Proc. of Argentinian Symposium Artificial Intelligence(ASAI-99, 1st) Buenos Aires, 1999. 7～35

4 Giacinto G, Roli F. Dynamic classifier selection based on multiple classifier behavior. Pattern Recognition, 2001, 34: 1879～1881

5 http://www.research.att.com/～lewis/reuters21578.html

6 Yang Y. An evaluation of statistical approaches to text categorization. Journal of Information Retrieval, 1999, 1(1/2): 67～88

7 Salton G, et al. A vector space model for automatic indexing. Communications of the ACM, 1975, 18: 613～620

(上接第110页)

4)此外，LMDS系统所处Ka波段的电波还易受天气的影响，雨、雪、雾等都会引起电波的衰减，较强的降雨甚至可能导致信号的完全中断。对此，LMDS系统通常采用动态自适应发信功率控制技术，在信号衰减较大的情况下，自动增大信号的发射功率，以便为系统提供足够的增益储备。

5)技术仍不完善。目前提供LMDS网络设备的厂家相当有限，设备性能也存在欠缺。

LMDS的未来 据预测，LMDS的频段分配原则不久将有定论，频谱拍卖的可能性较大，届时在中国已经宣传铺垫了近两年时间的LMDS市场显然将全面启动，而且其市场容量将会迅速膨胀。从目前了解的情况看，各大运营商甚至包括像广电、公安这样的专网用户都对LMDS表示了足够的关注，估计在频谱政策公布的两年之内，LMDS都将成为城域接入网建设的重点之一。

另一方面2000年年中划出3.5 GHz频段的30 MHz作为固定无线接入系统使用可以说是为这种产品在中国市场的推广提供了一个千载难逢的机会，在LMDS频段正式公布之前甚至公布之后的一段时间内，该类产品会成为各运营商关注的重点，因为它是目前运营商唯一可以申请使用的宽带无线接入频段(2.4 GHz ISM频段除外)。而无线LAN常常被看做有线网络和用户端计算机之间的"最后一公里"链路，从IEEE 1997年批准第一个无线LAN标准--802.11规范开始，这项技术正在不断地取得重大的进步。

当前，产品的价格、统一的标准、共享外设、宽带Internet连接的家庭应用等多种因素推动着无线LAN市场的发展。总之，各种宽带固定无线接入正面临广阔的市场前景和发展机遇，只有抓住当前的机遇，方能在未来的接入网市场中占据一席之地。

参考文献

1 唐雄燕. 城域宽带网的发展策略. 通信世界，2000(3)

2 熊四郎. 宽带接入技术发展新动态. 电信技术，2000(7)

3 袁栾峰. LMDS技术分析. 电信技术，2001(10)