基于全信息矩阵的多分类器集成方法*

唐春生+, 金以慧

(清华大学 自动化系,北京 100084)

A Multiple Classifiers Integration Method Based on Full Information Matrix

TANG Chun-Sheng+, JIN Yi-Hui

(Department of Automation, Tsinghua University, Beijing 100084, China)

+Corresponding author: Phn: 86-10-62313318 ext 3140, E-mail: tangchunsheng@tsinghua.org.cn; tangcs@smartdot.com

http://www.cs.tsinghua.edu.cn

Received 2002-05-24; Accepted 2002-08-14

Tang CS, Jin YH. A multiple classifiers integration method based on full information matrix. *Journal of Software*, 2003,14(6):1103~1109.

http://www.jos.org.cn/1000-9825/14/1103.htm

Abstract: Automatic text categorization is an effective method to increase the efficiency and quality of information utilizing. The combination of a set of different classifiers can often achieve higher classification accuracy. The concept of full information matrix is first given, and then an integration method of multiple classifiers based on adaptive weight adjusting is presented in this paper. The classifiers and their weights are determined automatically and adaptively with this method. The effective integration of each classifier's result can be realized by analyzing the statistical information of the classifier on the training set. The classification performance is promoted by the improvement of the precision and the recall. The effectiveness of the method is shown by the text classification experiments on the Reuters-21578 text sets.

Key words: combination of multiple classifiers; full information matrix; text classification

摘 要: 自动文本分类是提高信息利用效率和质量的有效方法,而多分类器的有效组合能够得到更高的分类准确率.给出了样本集在多分类器下的全信息矩阵概念,并提出一种权重自适应调整的多分类器集成方法.该方法能够自适应地选择分类器组合及确定分类器权重,并利用分类统计信息指导分类结果的集成判决.通过在标准文本集 Reuters-21578 上的实验表明:该方法能从查准率和查全率两方面提高文本分类的整体性能,同时表明了该方法的有效性.

关键词: 多分类器组合;全信息矩阵;文本分类

中图法分类号: TP181 **文献标识码**: A

多分类器组合方法常用来获得更好的分类效果,它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[1~4].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[1].

* Supported by the Science and Technology Committee of Beijing of China under Grant No.2001-0075 (北京市科委科技项目基金)

第一作者简介: **唐春生**(1974－),男,广西宜州人,博士,主要研究领域为文本分类与聚类,数据挖掘,Agent 技术.

© 中国科学院软件研究所 http://www.jos.org.cn

多分类器的组合方式可以分为级联和并联两种方式.Schapire 等人提出的 Boosting 方法是级联形式的多分类器组合方法中的代表[5,6].并联形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、模糊推理法以及通过分析样本特征而动态选择分类器的方法等[1~4].多分类器组合的方法有很多,但是到目前为止仍然不太成熟,在应用到具体应用领域时还需要做很多调整和处理工作[1,3].

Internet 中存储了海量的文本信息,如何有效地发现、处理、过滤和管理这些信息资源是一个亟需解决的问题.文本分类是提高信息利用效率和质量的有效方法,而多分类器的有效组合能够得到更高的分类准确率.由于文本样本的复杂性,各个分类器在不同样本区域内的分类效果是有差别的;又由于文本类别之间以及特征之间常常存在关联,这对分类结果会产生一定的影响,需要利用一些先验知识或统计知识加以调整.针对以上问题,本文提出了一种权重自适应调整的多分类器集成方法,该方法能够根据待定样本的特征自适应地选择分类器组合并确定其权重,而且能够根据训练样本的统计知识对待定样本的分类结果进行一定程度的调整,从而修正一些易错的划分.该方法能够充分发挥各分类器的优势,使分类器组合的整体性能更佳.

1 权值自适应调整的多分类器集成方法

1.1 多分类器组合的问题描述

对于一个有 M 个数据类的分类问题,其中 $\varpi_j, \forall j \in \Lambda = \{1,...,M\}$ 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.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 $C_i, i=1,...,L$ 来完成分类任务,则可以用 $e_i(X)=(e_{i1},...,e_{ij},...,e_{iM})$ 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 ϖ_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有效地进行综合.

1.2 权值自适应调整的多分类器集成方法的基本出发点

本文提出的方法基于以下对不同文本分类方法及其结果的观察:

(1) 不同样本具有不同的特征,在样本空间中也处于不同的区域;

(2) 各分类器对不同样本的分类效果是有差别的;

(3) 同一分类器在样本空间的不同区域分类性能会有所变化;

(4) 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.

从以上观察中可以引出全局准确率和局部准确率的概念.全局准确率是指分类器在整个样本集合上的准确率,局部准确率是指包含在样本空间的某个局部区域内的样本所构成的集合上的分类准确率.通过例 1 可以说明全局准确率和局部准确率的影响.

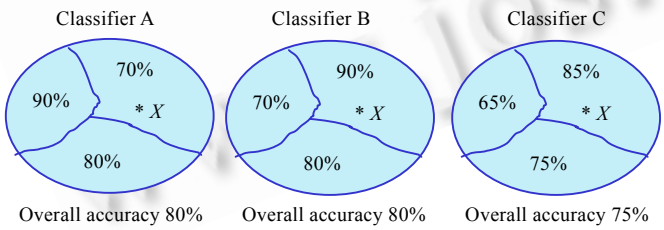

Fig.1 Example of overall accuracy and local accuracy

图 1 全局准确率和局部准确率的影响示例

例 1:假设 3 个分类器 A,B,C 在训练集合上的全局准确率分别为 80%,80%和 75%;假设样本空间可以分为 3 个区域,分类器在 3 个区域上的局部准确率如图 1 所示.X 为一个待定样本.

当利用分类器 A 和分类器 B 来进行组合时,若根据全局准确率来加权,则两个分类器的权重是相同的;如果考察局部准确率,由于分类器 B 的局部准确率为 90%,而分类器 A 的局部准确率为 70%,显然应该给分类器 B 加上更大的权重.

若在分类器 A,B 和 C 中"取分类准确率不低于 80%的分类器进行组合",则根据全局准确率,会选择分类器 A 和 B 进行组合,而根据局部准确率,则会选择分类器 B 和 C 来进行组合.

利用分类器在训练样本集上分类结果的统计知识,可以辅助对待定样本进行判断,对某些易错样本进行分类判断的调整.集成判决的思想可以通过例 2 来说明.

© 中国科学院软件研究所 http://www.jos.org.cn

例 2:假设分类器的分类结果用概率来表示,一般取概率最大的一类作为该样本所属的类别.若有一个分类器,通过统计知道,该分类器经常将属于类别 B 的样本划归为类别 A,而且概率达到 50%,则当分类器将一个待定样本划分为类别 A 时,我们会怀疑结果的可信度,猜测待定样本是否本应属于 B;当分类结果中属于 A 的概率只略微大于属于 B 的概率时,我们将会表示更大的怀疑.

如果通过某种方法利用统计的知识对结果进行调整,则有可能得到正确的类别划分.

1.3 关键问题及其解决

根据前面的分析,我们可以借鉴控制理论中的分段线性化和模型自适应的思想,针对不同待定样本所处的区域,分析在该区域上分类器的性能,从而自适应地选择分类器组合及确定分类器的权重,并利用分类器的统计信息对分类器输出进行调整,实现分类结果的集成判决.在这种方法中,需要处理以下一些关键问题:(1) 如何自动判别待定样本的有效邻域,从而在该邻域中分析分类器的性能?(2) 如何进行分类结果的集成判决?(3) 如何自适应地选择分类器组合及确定各分类器的权重?

在解决以上关键问题之前,首先引入全信息矩阵这一概念.利用全信息矩阵,可以很方便地分析分类器的行为,判断待定样本的有效邻域;计算分类器在有效邻域中的混乱矩阵,并利用这些统计信息指导分类结果的集成判决;计算各分类器在有效邻域中的分类准确率,自适应地选择分类器组合及确定分类器权重.

1.3.1 全信息矩阵

设 $\varpi_j, \forall j \in \Lambda = \{1,...,M\}$ 为 M 个不同的文本类, $C_i, i=1,...,L$ 分别为 L 个不同的分类器,则分类器 C_i 对样本 X 的分类输出可用 $e_i(X)=(e_{i1},...,e_{ij},...,e_{iM})$ 来表示,其中 $0 \le e_{ij} \le 1$ 表示在分类器 C_i 下样本 X 属于类 ϖ_j 的概率,一般选择最大的 e_{ij} 所对应的标号(设为 k,$1 \le k \le M$)作为样本 X 的类别标号,此时,样本 X 属于类 ϖ_k.

定义 1. 矩阵 $\boldsymbol{FIM}(\boldsymbol{X})=[e_1(X)', e_2(X)', ..., e_L(X)']'=\begin{bmatrix} e_{11} & \cdots & e_{1M} \\ \vdots & e_{ij} & \vdots \\ e_{L1} & \cdots & e_{LM} \end{bmatrix}$ 称为样本 X 在多分类器 $\{C_i, i=1,...,L\}$ 下的全信息矩阵.

定义 2. 对样本集合 $D=\{\boldsymbol{X}_l\}, l=1,...,N$,称 $\boldsymbol{FIM}(\boldsymbol{D})=[\boldsymbol{FIM}(\boldsymbol{X}_1),..., \boldsymbol{FIM}(\boldsymbol{X}_l),..., \boldsymbol{FIM}(\boldsymbol{X}_N)]$ 为样本集 D 在多分类器 $C_i, i=1,...,L$ 下的全信息矩阵.

全信息矩阵包含了各个分类器在样本或样本集上的全部分类结果信息,从中可以统计和分析出分类准确率、类别相关情况等重要信息,从而奠定本文方法的基础,并支持关键问题的解决.

1.3.2 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所构成的区域,通过分析分类器在该区域上的分类性能,可以判断分类器在该类上的可信度,从而选择合适的分类器组合和权重.然而,在待定样本的邻域中,通常会有这样一些样本,分类器认为待定样本与这些样本不同类,而如果多个分类器在这些样本上都出现了“不同类”的判断,那么这些样本与待定样本不属于同一类的可能性就比较大,对分类器的性能分析造成干扰,并影响到分类器的选择与加权.对这些干扰样本进行剔除后可以得到待定样本的有效邻域.

利用多分类器的行为分析可以有效地剔除无效的邻域样本,自动确定待定样本的有效邻域.

定义 3[7]. 给定 M 个数据类和 L 个分类器,令 $C_i(\boldsymbol{X}) \in \{1,...,M\}$ 表示分类器 C_i 对样本 X 的分类标签,则称 $C_i(X) \in \{1,...,M\}$ 为分类器 C_i 在样本 X 上的行为.而 $MCB(X)=\{C_1(X), C_2(X),..., C_L(X)\}$ 则称为 L 个分类器在样本 X 上的行为(multiple classifier behavior,简称 MCB).

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 $T_i(X,Y), i=1,...,L$,定义为

$$T_i(X,Y)=\begin{cases}1, & \text{if } C_i(X)=C_i(Y)\\ 0, & \text{if } C_i(X)\ne C_i(Y)\end{cases}.$$

© 中国科学院软件研究所 http://www.jos.org.cn

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,说明每一个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 0,说明分类器在样本上的分类判断分歧越大,也就是说,X 和 Y 不属于同一类的可能性越大.利用这种方法可以剔除与待定样本 MCB 的相似度小于阈值的邻域样本.

样本 X 在多分类器 $C_i, i=1,...,L$ 下的全信息矩阵为

$$\boldsymbol{FIM}(\boldsymbol{X})=[e_1(X)',e_2(X)',...,e_L(X)']'=\begin{bmatrix} e_{11} & \cdots & e_{1M} \\ \vdots & e_{ij} & \vdots \\ e_{L1} & \cdots & e_{LM} \end{bmatrix},$$

则对于分类器 C_i,它在样本 X 上的行为 $C_i(X)\in\{1,...,M\}$ 可以通过式(2)求得.

$$C_i(X)=\arg\max_j(e_{ij}),\quad j=1,...,M. \tag{2}$$

其中 $\arg\max_j(e_{ij})$ 是取使 e_{ij} 最大的 j,即取分类结果中概率最大的值所对应的分类标号.

1.3.3 分类结果的集成判决

集成判决的思想由肖旭红在文献[3]中首先提出,认为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 ϖ_l)对集成判决 $X\in\varpi_j$ 的支持作用, $m_i(x,l,k)$ 为分类器 C_i 下样本 X 与类别 ϖ_l 之间的相似度, $R_i(j,l)$ 表示分类器 C_i 将类别 ϖ_j 中的样本判为属于类别 ϖ_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. $m_i(x,l,k)$ 可以由分类器的输出中直接得到,W_k 可以通过多种方式来计算,如 $W_k=e^{-\alpha(k-\delta)}$ 或 $W_k=1.0-k*\beta$ 等,其中 k 为候选阶次,α,β,δ 均为非负常数[3].

式(3)表明:分类器 C_i 将类别 ϖ_j 中的样本判为属于类别 ϖ_l 的可能性越大,对判决 $X\in\varpi_j$ 的支持作用越大;待定样本 X 与类 ϖ_l 越相似,对判决 $X\in\varpi_j$ 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

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

在分类结果的集成判决中, $R_i(j,l)$ 中包含了分类器 C_i 的分类性能信息,说明了在分类器 C_i 下各个类别样本的划分情况以及类别之间的相关关系.这些信息可以通过对分类器在训练集上的识别情况进行统计后得到,一般用分类器的混乱矩阵(confusion matrix)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 ϖ_j 中的样本识别为 ϖ_l 类的概率.

$R_i(j,l)$ 可用归一化后的 $r_{jl}^{(i)}$ 来近似,而混乱矩阵可以通过全信息矩阵来计算.

样本集 D 在多分类器 $C_i, i=1,...,L$ 下的全信息矩阵为

$$\boldsymbol{FIM}(\boldsymbol{D})=[\boldsymbol{FIM}(\boldsymbol{X}_1),..,\boldsymbol{FIM}(\boldsymbol{X}_l),..,\boldsymbol{FIM}(\boldsymbol{X}_N)].$$

$\boldsymbol{FIM}(\boldsymbol{D})$的第 i 行包含了分类器 C_i 对样本集 D 中所有样本的分类输出,对这些分类情况进行统计即可得到样本集 D 在分类器 C_i 下的混乱矩阵.

对于训练样本集 D 中的样本,其类别标号是已知的,设训练样本 X 的类别标号为 j;从 $\boldsymbol{FIM}(\boldsymbol{X})$中提取出分类器 C_i 下 X 的分类输出,设为 $[e_{i1},...,e_{ij},...,e_{iM}]$,令最大的 e_{ij} 所对应的标号为 l,则混乱矩阵的计算如下:

方法 1. 初始化 $M\times M$ 矩阵 R=0,对样本 X,其所属类别为 j,分类器给出的类别标签为 l ,则矩阵 R 的相应位置上加 1,即 $R(j,l)=R(j,l)+1$;对训练样本集中的所有样本进行相同处理,得到一个元素为整数值的样本划分情况矩

© 中国科学院软件研究所 http://www.jos.org.cn

阵 R,对矩阵 R 的每一行作归一化处理,即 $r_{jl}=R(j,l)\Big/\sum_{k=1}^{M}R(j,k)$,可以得到样本集在分类器 C_i 下的混乱矩阵.

方法 2. 初始化 $M\times M$ 矩阵 R=0,对样本 X,其所属类别为 j,则可将分类器 C_i 对 X 的分类输出 $[e_{i1},...,e_{ij},...,e_{iM}]$ 直接加到矩阵 R 的第 j 行,依此类推,处理样本集中的所有样本,最后对矩阵 R 的每一行作归一化处理即可得到混乱矩阵.

方法 1 是布尔形式的求取混乱矩阵的方法,方法 2 是概率形式的方法,从理论上说,方法 2 能够更真实地反映类别划分的情况及类别之间的关系.

混乱矩阵的估计精度主要和训练样本集中的样本数目有关,训练样本数越多,估计越准确.

1.3.4 分类器的自适应组合及加权

各分类器的分类效果是随着待定样本的特征和区域不同而变化的,因此需要针对不同样本进行分类器组合及权重的自适应调整.本文方法考察分类器在待定样本的有效邻域中的分类准确率,按照准确率给予不同的权重,准确率越高,分类器权重越大.如果某些分类器在有效邻域中的分类准确率太低,则在这次分类器的组合中不考虑该分类器.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:

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

1.4 算法描述

权重自适应调整的多分类器集成算法的流程如图 2 所示.

算法中利用多分类器行为分析解决了有效邻域判定问题,通过分析有效邻域的混乱矩阵,得到了分类结果的集成判决,并自适应地选择了分类器组合及分类器权重.

在分类器行为分析中,如果待定样本的 MCB 中的各项值都相同,则说明各分类器对待定样本的类别判断一致,此时可以直接输出该类别作为样本的类别,而无须进行后续的分析,从而减少计算时间.

当利用混乱矩阵计算各分类器的分类准确率时,可以得到一个最大准确率,如果某分类器的准确率与最大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分类器分配权重.这种方法得到的分类器组合是根据样本特征而自适应变化的,有可能直接选出最好的分类器来输出分类结果,也可能是用几个较好的分类器来进行组合,然后输出综合后的结果.

2 文本分类实验结果及分析

2.1 实验设计

在文档分类领域,Reuters-21578 文章集是常用的测试平台[8].本文的实验也是在该文章集上进行的.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 7 194 个训练文档和 2 788 个测试文档,训练文档还可以分成训练样本集和校验样本集,其中校验样本集中的文档占总训练文档的 1/5.

本文方法组合了文本分类中常用的方法,包括 KNN 方法、SVM 方法、Naïve Bayes 方法和线性分类方法[9].文档表示采用向量空间模型,利用 tfidf 方法进行词项加权.通过在训练集上进行训练并在校验集上进行校验,得到各方法的最佳分类参数.后续的分类器组合都基于最好的分类器参数来进行.

利用全局准确率来确定分类器权重,可以看做是一种权重不变的多分类器线性加权方法;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方法与上述方法的对比实验.实验中主要考察了这些方法的查准率、查全率和 F_1 指标[9].

本文方法的实验参数如下:

在确定待定样本的有效邻域时,邻域中的初始样本数为 50,MCB 相似度阈值为 0.5,即相似度小于 0.5 的邻域样本可看做是干扰样本,需要剔除,这样,有效邻域中的样本数是根据待定样本特征而动态变化的.

© 中国科学院软件研究所 http://www.jos.org.cn

在进行集成判决时,阶次支持因子[4]的计算公式为 $w_k = e^{-k}$, $W_k = w_k \Big/ \sum_{k=1}^{M} w_k$,其中 $\alpha = 1, \delta = 0, k = 1,2,...,M$,进行归一化处理后得到的 W_k 前 10 个分量依次为(0.6321,0.2326,0.0856,0.0315,0.0116,0.0043,0.0016,0.0006,0.0002,0.0001).混乱矩阵的计算采用了概率的形式.

在进行分类器的自适应组合和加权时,需要考察分类器准确率与最大准确率之间的差值,当差值大于阈值时,则在当前组合中不考虑该分类器.实验中的准确率阈值为 0.05.

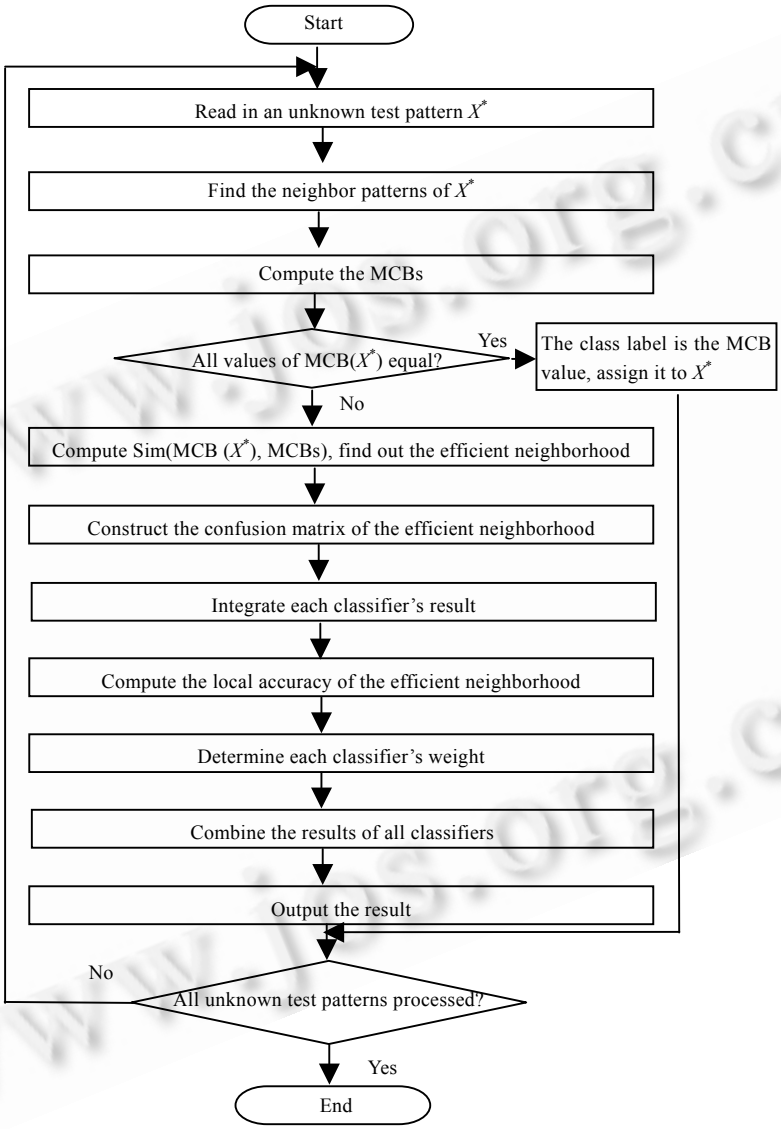

Fig.2 Flow chart of multiple classifiers integration algorithm based on adaptive weight adjusting
图 2 权重自适应调整的多分类器集成算法流程图

2.2 实验结果及分析

表 1 给出了这些方法在测试集上的实验结果.比较实验结果,可以得出以下结论:

(1) 通过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能够取得较好的分类效果.多分类器组合的分类查准率、查全率和 F_1 指标普遍高于单个分类器(SVM 除外).本文的多分类器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.

© 中国科学院软件研究所 http://www.jos.org.cn

(2) 通过比较权重不变(全局准确率)的方法和权重自适应(局部准确率)的方法,发现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并确定其权重,可以更好地发挥和综合各分类器的优点,使得查准率和查全率都有一定程度的提高,从而提高了分类的整体性能.

(3) 通过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在一定程度上提高查全率,从而从整体上(F_1指标)提高了分类性能.

(4) 本文提出的权重自适应调整的多分类器集成方法分类效果最好.通过分析样本特征,可以更有针对性地选择分类器组合及其权重;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,并对分类结果进行调整,有助于纠正一些易错的划分,部分地解决一个样本属于多个分类的问题,从而使组合分类器的整体性能更佳.

Table 1 Results comparison among the eight classifiers

表 1 8 种分类方法的实验结果比较

Classifiers	Presision	Recall	F_1
KNN classifier	0.828	0.785	0.806
SVM classifier	0.867	0.83	0.848
Naïve Bayes classifier	0.82	0.815	0.817
Linear classifier	0.812	0.805	0.808
MCS based on overall accuracy	0.847	0.812	0.829
MCS based on results integration	0.841	0.83	0.836
MCS based on local accuracy	0.858	0.829	0.843
MCS based on local accuracy and results integration	0.874	0.846	0.858

3 小 结

随着问题复杂度的增加,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,将各分类器的结果有效地综合起来.

针对多分类器组合用于文本分类时存在的问题,提出了样本集在多分类器下的全信息矩阵概念,并提出了一种权重自适应调整的多分类器集成方法.该方法能够针对不同待定样本自动选择不同的分类器组及其权重,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,指导分类结果的集成判决,提高了分类的查准率和查全率,最终从整体上提高了分类性能.这在 Reuters-21578 文本集上的文本分类实验中得到了验证.

本文提出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,需要针对具体情况进行相应的处理,并需要做进一步的优化和调整.另外,可以考虑进一步引入控制理论中的思想,实现更好的自适应调节.

References:

[1] Sebastiani F. A tutorial on automated text categorization. In: Analia A, Ricardo Z, eds. Proceedings of the 1st Argentinian Symposium Artificial Intelligence (ASAI-99). Buenos Aires, 1999. 7~35.

[2] Li YH, Jain AK. Classification of text documents. The Computer Journal, 1998,41(8):537~546.

[3] Xiao XH, Dai RW. A metasynthetic approach for handwritten Chinese character recognition. Acta Automatica Sinica, 1997, 23(5):621~627 (in Chinese with English abstract).

[4] Jing XY, Yang JY. Combining classifiers based on analysis of correlation and effective supplement. Acta Automatica Sinica, 2000, 26(6):741~747 (in Chinese with English abstract).

[5] Schapire RE, Singer Y, Singhal A. Boosting and Rocchio applied to text filtering. In: Croft WB, Moat A, van Rijsbergen CJ, eds. Proceedings of the 21st SIGIR-98 ACM International Conference on Research and Development in Information Retrieval. New York: ACM Press, 1998. 215~223.

[6] Freund Y, Schapire RE. A short introduction to boosting. Journal of Japanese Society for Artificial Intelligence, 1999,14(5): 771~780.

[7] Giacinto G, Roli F. Dynamic classifier selection based on multiple classifier behavior. Pattern Recognition, 2001,34(9):1879~1881.

[8] http://www.research.att.com/~lewis/reuters21578.html. 1998.

[9] Yang YM, Liu X. A re-examination of text categorization methods. In: Hearst MA, Gey F, Tong R, eds. Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99). New York: ACM Press, 1999. 42~49.

附中文参考文献:

[3] 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.

[4] 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.

© 中国科学院软件研究所 http://www.jos.org.cn