第34卷 第19期 计 算 机 工 程 2008年10月
Vol.34 No.19 Computer Engineering October 2008
·人工智能及识别技术· 文章编号：1000—3428(2008)19—0187—02 文献标识码：A 中图分类号：TP301

基于 Ensemble 的增量分类方法

刘 波，潘久辉
(暨南大学计算机科学系，广州 510632)

摘 要：针对在维护数据挖掘模型过程中须反复计算数据集、效率较低的问题，基于 Ensembles 学习思想，研究增量数据集的弱分类器生成方法，根据增量数据集分类器之间的相异度提出新的组合分类算法，分析组合分类器的出错率。实验结果表明，该分类方法是有效的。
关键词：增量；分类；Ensemble 学习；组合

Incremental Classification Method Based on Ensemble

LIU Bo, PAN Jiu-hui
(Department of Computer Science, Jinan University, Guangzhou 510632)

【Abstract】 To maintain a data mining model, it is necessary to repeat related computation for frequent changing data sets, and this will lead to the low efficiency problem. This paper uses Ensemble learning to study generation of weak classifiers on incremental data sets, and presents a new combination algorithm according to dissimilarity of classifiers of incremental data sets. It analyzes the error rate of the combined classifier. Experimental results show the effectiveness of the classification method.
【Key words】 incremental; classification; Ensemble learning; combination

1 概述

成熟的数据挖掘算法是面向静态数据集的，随着数据库规模的扩大，反复对动态数据库中的所有数据进行挖掘消耗时间巨大，因此，须对原有挖掘模型以增量方式进行修改。本文采用 Ensemble 学习思想对增量分类方法进行研究。

Ensemble技术是通过一组分类器组合，提高弱分类器的精度，改善分类结果[1]。文献[1]提出的AdaBoost算法，是Ensemble技术的常用方法之一，对每个训练集中的实例采用不同分布，产生一个弱分类器(分类准确率仅要求大于 50%)，最后按多数投票方式结合成一个强分类预测器。本文借鉴上述方法提出IncBoost算法，产生增量集分类器。

目前，在增量分类挖掘方面，已有较多研究，但Ensemble的方法较少。文献[2]虽采用该方法研究增量分类和动态集成分类器的问题，但其增量数据集(或分组数据)均只包括增加的记录，不包括已训练集中数据被删除或修改的记录，而这些记录有许多应用，例如由于电信和银行业务中固定客户转移导致的原训练集数据删除。本文研究增量分类器算法，同时考虑增加数据集和删除数据集。此外，多个分类器组合的目的是得到一个更为精确的预测分类器，被组合的分类器之间的差异性对组合后的分类器预测效果起关键作用[3]。本文基于分类器相异度，提出增量数据集分类器组合方法。

2 增量分类器及组合方法

本文将变化的数据分为 2 个集合：增加的数据集和删除的数据集，其中初始数据集作为增加数据集。当增加的数据或删除的数据累计达到一定数目时，将它们作为新的训练集调用 IncBoost 算法计算得出新的增加集或删除集分类器，再调用 ComBiner 算法生成或维护整体分类器(组合分类器)，即可通过增量集分类器而无需原数据集来实现增量分类。整个过程有 2 种组合：(1)包括在计算增量集分类器中，提高弱分类器的准确率。(2)包括在计算整体分类器中，提高整体分类或预测的准确率。

2.1 增量数据集分类方法

增量数据集分类算法 IncBoost 可分为以下 4 个部分：

(1)对每个增量数据集，假设其中实例的初始分布均匀，每一个实例的初始权重相同，通过调用弱学习算法(Weaklearn)后，得到规则集h_t。

(2)用h_t对各实例类型进行推测，并与真实值比较，计算出错率ε_t。

(3)修改实例权重，增加分类错误的实例在下次学习训练中的作用。

(4)将得到的增量数据集分类器由若干h_t分类器按投票的方式构建。

算法描述如下：

输入：新增训练集$S_n=\{(x_1, y_1), (x_2, y_2),\cdots,(x_m, y_m)\}$,$x_i \in X$, $y_i \in Y=\{1, 2, \cdots, k\}$)；分类算法WeakLearn ($S$, D), D为实例分布；组合弱分类器的次数T。

输出：新增量集的分类器及错误规范化值。

for i=1 ,2,…, m

初始化S_n中实例(x_i, y_i)的权值$w_1(i)=1/m$

for t=1, 2, …, T

$$D_t(i)=\frac{w_t(i)}{\sum_{i=1}^{m} w_i(i)};$$

调用 WeakLearn (S_n, D_t);

产生分类规则集 h_t: X→Y;

$$\varepsilon(h_t)=\sum_{i=1}^{m} D_i(t)\left\|h_t(x_i)\neq y_i\right\|;$$

基金项目：广东省科技攻关基金资助项目(2003c101011)
作者简介：刘 波(1965－)，女，副教授，主研方向：数据挖掘，数据库，信息集成与数据仓库，智能化信息处理；潘久辉，教授
收稿日期：2007-11-12 **E-mail：**lbxldd@sohu.com

If $\varepsilon(h_t)>1/2$ then t=t-1 and continue;

Else

$\beta(t)=\varepsilon(h_t)/(1-\varepsilon(h_t))$;

$R_t=\arg\max_{y\in Y}\sum_{t:h_t(x)=y}\mathrm{lb}(1/\beta(t))$;

$\varepsilon(R_t)=\sum_{i=1}^{m}D_i(t)\left\|R_t(x_i)\neq y_i\right\|$;

If $\varepsilon(H_t)>1/2$ and k>2 then T=t-1 and abort loop

Else

$B(t)=\varepsilon(R_t)/(1-\varepsilon(R_t))$;

$w_{t+1}(i)=w_t(i)\times B(t)^{(1-\|R_t(x_i)\neq y_i\|)}$;

End if

End for

$H_n=R_T$; $B_n=B(T)$;

IncBoost算法有如下规定：(1)集合$\{(x_1, y_1), (x_2, y_2),\cdots, (x_n, y_n)\}$表示一张关系数据表，$(x_i, y_i)$代表第$i$条记录，即第$i$个实例，其中，$x_i$为条件属性数据项组合，$y_i$为实例的类型。(2)$w_t$表示在时间t，实例集的权重向量$\{w_1^t, w_2^t,\cdots, w_n^t\}$，其中$w_t(i)$表示实例$i$在时间$t$的权重。(3)$D_t$表示在时间$t$，实例的分布概率向量，其中，$D_t(i)$表示实例$i$在时间$t$的分布概率。(4)$h_t(x)$是在时间$t$，$x$的类型函数。(5)$h_t$：$X\rightarrow Y$表示在时间$t$的分类规则：if X then Y。(6)如果$h_t(x_i)\neq y_i$，$\|h_t(x_i)\neq y_i\|=1$，否则，$\|h_t(x_i)\neq y_i\|=0$。(7)$\varepsilon(h_t)$表示规则集h_t的出错率。(8)R_t表示t个弱分类器的组合。(9)H_n和B_n表示新增量集S_n的分类器及错误规范化值。

对于初始数据集和每一增加数据集可采用任意WeakLearn分类算法产生弱分类器。由于通过选择较小的群体大小和遗传次数等参数，易产生弱分类器，因此在本研究工作的实验中应用遗传分类算法GA_C[4]。

删除的数据集累计到一定数目时，同样存在一些内在分类规则，使用上述算法，也可生成删除集分类器。

2.2 多个增量分类器的组合

通过 IncBoost 算法产生的分类器，由于训练的数据子集不同，互相之间会存在一定的差异。文献[3]研究了子分类器之间的差异度与组合分类器准确度的关系：子分类器之间的差异度越大，组合分类器按多数投票方式确定未知实例的类型会越准确。这适合于增加数据集产生的分类器，它们之间的差异度越大，组合分类器的准确度越高。增加数据集产生的分类器能够直接组合并更新整体分类器，而删除数据集产生的分类器虽不能组合到整体分类器中，但也影响到组合分类器的准确度。删除数据集的分类器与某增加数据集的分类器差异度越大，则删除集对该增加数据集分类器影响越小，该增加数据集分类器在组合投票中的权重就应增加。

基于上述思想，本文提出组合分类器算法ComBiner。文献[5]指出，2 个分类模型的相异度有多种计算方法，本文按式(1)，在n个分类器中计算分类器H_m和H_p预测实例x的相异度，相异度的取值区间为[0, 1]。

$$\delta_x(H_m,H_p)=\frac{1}{|C|}\sum_{j=1}^{|C|}\frac{|PD_{mj}(x)-PD_{pj}(x)|}{R_j(x)} \tag{1}$$

其中，$|C|$表示类型总数；实例x通过每个分类器计算可得；$H_y(x)$表示类型结果，y=1, 2,⋯, n；$PD_{yj}(x)$为H_y分类器预测实例x的类型为C_j的后验概率分布，j=1, 2 ⋯ $|C|$；$R_j(x)=\max\{PD_{1j}(x),\cdots,PD_{nj}(x)\}-\min\{PD_{1j}(x),\cdots,PD_{nj}(x)\}$是$C_j$类型的后验概率分布区间。

算法描述如下：

输入：已经生成的$n(n\geq 1)$个增量数据集分类器，包括r个增加数据集分类器$H_1, H_2,\cdots, H_r$，s个删除数据集分类器$H_{r+1}, H_{r+2},\cdots, H_n$；已计算出的$n(n\geq 1)$个增量数据集分类器的错误规范值$B_1, B_2,\cdots, B_n$；$C$类型集。

输出：组合分类器H_{final}(预测x实例的类型)。

If (s=0) then

$$H_{final}=\arg\max_{c\in C}\sum_{t=1}^{r}\mathrm{lb}(1/B_t)\times\|H_t(x)=c\| \tag{2}$$

If (s≥1) then

$$H_{final}=\arg\max_{c\in C}\sum_{t=1}^{r}(\mathrm{lb}(1/B_t)\times\|H_t(x)=c\|\times\sum_{u=r+1}^{n}\delta_x(H_t,H_u)) \tag{3}$$

ComBiner算法规定，如果$H_t(x)=c$，式(2)和式(3)中的$\|H_t(x)=c\|=1$；否则$\|H_t(x)=c\|=0$。算法处理分以下 2 种情况：

(1)不存在删除数据集分类器，则式(2)直接采用 AdaBoost 算法中提出的组合方法，根据各分类器出错率的大小确定投票权值，取加权和最大值对应的类型作为 x 的预测结果。

(2)存在删除数据集分类器，则式(3)考虑删除数据集分类器与增加数据集分类器之间的相异度，由于删除数据集分类器与增加数据集分类器的相异度越大，其间的影响越小，因此对加权和的值贡献越大，反之，则越小。

2.3 组合分类器的出错率分析

文献[2]证明：按式(2)组合的分类器出错率为

$$E\leq 2^T\prod_{t=1}^{T}\sqrt{E_t(1-E_t)}$$

其中，T是组合分类器的数；E_t是第t个组合分类器的出错率，并且，$E_t\leq 2^t\prod_{s=1}^{t}\sqrt{\varepsilon_s(1-\varepsilon_s)}$，$\varepsilon_s$是每一子分类器$H_s$的出错率。

根据IncBoost算法，由包含m个实例的训练集S产生的分类器出错率按下式计算。

$$\varepsilon(H_t)=\sum_{i=1}^{m}D_i(t)\left\|R_t(x_i)\neq y_i\right\|,\ x_i\in S \tag{4}$$

按原数据集计算出错率，当 S 中少量数据被删除，对实例数据分布 D 和原分类器的出错率影响不大；当 S 中大量数据被删除，原分类器的出错率就会增加。但如按式(3)的组合分类器对 S 中的数据进行分类，得到的数据类型则会通过 $\delta_x(H_t, H_u)$ 进行调整，从而减少分类出错率。

3 实验结果

从UCI公共数据库[6]中选取一组数据集Car Evaluation，在Windows XP 操作系统、P4 2.4 G环境下，用VC6.0 软件开发工具进行实验。Car Evaluation数据集包括 1 728 个实例，分为 4 类(多为 1, 2 类型)，每一个实例有 6 个条件属性。将该数据集分为 6 个训练集$S_1, S_2,\cdots, S_6$和一个测试集*Test*。每一个训练集各包括 200 个实例，作为增加数据集；*Test*包括 528 个实例，用于测试不同分类器的准确率；S_1和S_2仅包括 1, 2 类型的实例，S_3和S_4包括 1, 2, 3 类型的实例，S_5, S_6和*Test*包括所有 4 种类型的实例；S_7是从$S_1, S_2,\cdots, S_6$和Test中分别抽取 30 个实例组成的删除数据集，抽取数据后$S_1, S_2,\cdots, S_6$和*Test*变为S_1', S_2',⋯, S_6'和*Test*'。以$S_1, S_2,\cdots, S_6, S_7$为增量集调用IncBoost算法($T$=5)分别产生的增加集分类器是$H_1, H_2,\cdots, H_6$和删除集分类器$H_7$。在弱分类器采用的遗传算法GA_C中，设群体大小为 500，遗传代数为 5。

GA算法分别以$S_1, S_2,\cdots, S_6$和*Test*为训练集，以*Test*为测试集，经过多次运行得到的平均分类准确率(符合弱分类器的要求)，如表 1 所示。多个组合分类器对*Test*数据集多次测试得到的平均分类准确率，其中(H_1, H_2)表示H_1和H_2的组合器，(H_1, H_2, H_3)表示H_1, H_2, H_3的组合器，以此类推，如表 2 所示。

（下转第 191 页）