一种可并行的贝叶斯集合在线学习算法*)

古 平 朱庆生
(重庆大学计算机学院 重庆 400044)

摘 要 无论是 Boosting 还是 Bagging 算法,在使用连续样本集进行分类器集合学习时,均需缓存大量数据,这对大容量样本集的应用不可行。本文提出一种基于贝叶斯集合的在线学习算法 BEPOL,在保持 Boosting 算法加权采样思想的前提下,只需对样本集进行一次扫描,就可实现对贝叶斯集合的在线更新学习。算法针对串行训练时间长、成员相关性差的缺点,采用了并行学习的思想,通过将各贝叶斯分量映射到并行计算结构上,提高集合学习的效率。通过 UCI 数据集的实验表明,算法 BEPOL 具有与批量学习算法相近的分类性能和更小的时间开销,这使得算法对某些具有时间和空间限制的应用,如大型数据集或连续型数据集应用尤其有效。

关键词 贝叶斯分类器,集合,在线学习

A Parallelizable Bayesian Ensemble Online Learning Algorithm

GU Ping ZHU Qing-Sheng
(College of Computer Science and Engineering, Chongqing University, Chongqing 400044)

Abstract In situation where data is being generated continuously, storing large amounts of data is impractical for Boosting and Bagging algorithms. In this paper, we present a parallelizable Bayesian ensemble online learning algorithm (BEPOL), which follow the method of reweighting in determining the training sets, but only require one pass through the entire sets. In order to improve both the correlation and time efficiency, parallel training based on negative correlation is used in BEPOL for training each individuals in the ensemble. With experiments conducted on several UCI datasets, BEPOL are proven to perform comparably to the Adaboost in terms of classification performance and have the advantage of lower running time for large datasets. This make BEPOL particularly suitable for some applications where large or continuous datasets are available.

Keywords Bayesian classifier, Ensemble, Online-learning

1 引言

贝叶斯分类器因其处理不确定性问题的能力和良好的推理机制,在生物医学、图像处理[1]、语音识别等领域得到广泛应用,很多监督学习算法可以用于训练贝叶斯分类器,如 CL 算法,TAN 算法,BNAN 算法等。但实际应用中很多复杂的问题仅仅依赖单一的分类器无法取得令人满意的效果,而贝叶斯集合(ensemble)则显示了其远远高于单一分类器的分类能力。实验[2]证明,只要集合中各贝叶斯分量不在同一样本空间产生相同的错误,即预测结果具有多样性,则整个分类器集合的预测将优于任何一个分类器单独的预测结果。

各种分类器集合的学习算法中,最有效的是 Bagging 和 Boosting 算法,与其它学习算法一样,它们均属于批量学习算法,要求预先给定样本集,然后利用样本集的统计特征进行学习,不仅每加入一个贝叶斯分量需要至少完成一次样本集扫描,而且每个贝叶斯分量的学习也需要对样本集进行多次扫描。就时间开销而言,对大样本集或连续型样本集应用,批量学习算法显然无法适用。而在线学习算法只需利用单个样本或进行一次数据扫描,就可实现对整个分类器集合的学习,因此具有更好的适应能力和响应能力。Kivinen[3] 和 Fern[4] 将在线学习引入到 Boosting 算法中,利用单个样本顺序对所有分类器进行增量更新,取得了不错的效果。但这种串行训练方式[5]也可能导致:(1)成员间缺少相关性,可能产生正相关的分类结果;(2)时间开销过大,无法满足并行体系结构的需要。本文将从另一角度研究基于贝叶斯集合的在线学习算法,尤其针对具有多处理器的并行计算结构,我们给出一种可并行的在线 Boosting 学习算法,利用单个或很少的样本,集合中所有贝叶斯分量同时进行在线学习,这不仅可以减小串行训练所带来的时间开销,而且利用成员间的分类差异可以满足贝叶斯集合的多样性要求,从而提高系统的分类性能。

2 贝叶斯集合与 AdaBoost

贝叶斯网可以简单地表示为 $B=\{B_S, B_P\}$,其中 B_S 为贝叶斯网结构,是由 n 个随机变量构成的有向无环图,B_P 为贝叶斯网的条件概率表,表示每个变量 v_i 对其父结点 $u(v_i)$ 的条件概率。如果网络结构已知,贝叶斯学习就是自动选择与样本数据最佳拟合的贝叶斯条件概率参数,它可以归结为一个连续空间的寻优问题,梯度下降法、极大似然估计法、Gibbs 采样法等均可用于上述问题的解决。贝叶斯网和决策树、神经网络一样,可以用作分类器集合中的分量模型,利用各分类器在不同样本空间的分类能力提高分类器总体的分类性能。

所有分类器集合的学习算法中最常见的是 AdaBoost[6] 算法,它采用一种样本加权采样的思想,为每个样本赋予一个权值 $w_k(i)$,表示其被选入分类器 C_k 训练集的概率,如果样

*)重庆市自然科学基金项目(编号 2005BB2224)资助。古 平 博士生,讲师,研究方向为贝叶斯网,数据挖掘,模式识别。朱庆生 教授,博导,研究方向为多媒体技术,数据挖掘,模式识别等。

本被正确分类，则其权值降低；否则，该样本在构造下一个训练集时被选中的概率就会增加，这使得随后的分类器 C_{k+1} 更关注 C_k 的分类错误，提高分类器的多样性。最终的分类结果可以通过所有分量预测的加权平均取得：$h(x)=\sum_{k=1}^{k_{max}}\alpha_k h_k(x)$，$\alpha_k$ 表示各分类器的权重。

3 可并行的贝叶斯集合在线学习算法

3.1 BEPOL 分析与设计

AdaBoost 算法利用样本加权采样的思想可以提高分类器的总体性能，但同时由于每个分类器均需要对原始样本集多次采样，这也限制了算法在在线学习中的应用。为解决该问题，我们提出一种折衷思路：在保持 AdaBoost 算法对样本加权采样的前提下，减少对同一样本的采样次数，通过概率方式产生与实际采样相似的训练样本集。

通过对 AdaBoost 算法的分析，我们注意到：假定样本集中所有样本的权值相同，则每个样本出现在各分类器对应的训练子集中的次数 c 满足二项分布：

$$P(X=c)=\binom{N}{c}\left(\frac{1}{N}\right)^{c}\left(1-\frac{1}{N}\right)^{N-c} \tag{1}$$

当样本集很大，即 $N\to\infty$，式(1)将趋于泊松分布：$Possion(\lambda=1)=\frac{e^{-1}}{c!}$。进一步考虑样本权值的变化，如果样本被错误分类，我们可以增大参数 λ 的取值，这样它出现在训练子集中的次数也会增加，反之，随着 λ 的减小采样次数就会减少，因此它仍然满足泊松分布。由于每个样本在各训练子集中的分布已经确定，我们可以模拟 AdaBoost 算法中的采样过程：对每个训练样本，首先根据分布 Possion(λ)确定每个样本的副本数 c，贝叶斯学习时，分类器使用 c 个同样的样本学习即可。可以验证，随着样本容量的增加或趋于无穷，算法 BEPOL 将产生与 AdaBoost 相同的训练样本分布。根据上述思路，我们设计在线学习算法 BEPOL 如下：假定初始贝叶斯集合 E 已经给定，由 M 个贝叶斯分量组成，所有参数初始化为一任意小随机数。

算法输入：E，样本$(x_i,y_i\in\{-1,1\}))$，贝叶斯参数在线学习算法 $L_{onlineLearn}$；
算法输出：基于(x_i,y_i)更新后的 E；
　初始化 $\lambda_k^i=1$；
　for 每个贝叶斯分量 $C_k\in E$
　　并行计算 $h_k(x_i),k=1..M$；
　　$E_k=\frac{1}{k-1}\sum_{j=1}^{k-1}y_i h_j(x_i)$
　　$\alpha_k=\frac{1}{2}\ln(\frac{1+E_k}{1-E_k})$
　　$\lambda_k^i=\lambda_k^i\exp(-\alpha_k y_k h_k(x_i))$
　　$m_k=Poisson(\lambda_k^i)$
　　$C_k=L_{onlineLearn}(C_k,(x_i,y_i),m_k)$
end for

算法首先初始化所有样本的权值为 1，然后所有贝叶斯分类器并行地对样本(x_i,y_i)进行分类。如果样本(x_i,y_i)被之前的多数分类器和当前分类器 C_k 错误分类，则 $\alpha_k y_k h_k(x_i)<0$，λ_k^i 取值增大，相应的分类器 C_k 使用该样本学习时的采样次数也会增加；反之，如果样本(x_i,y_i)被当前分类器和其它多数分类器正确分类，则意味着集合已经能正确识别该样本，相应的学习次数就会减少。根据样本(x_i,y_i)以及泊松分布所确定的不同样本副本数，各贝叶斯分量再并行地进行在线更新，并返回更新后的贝叶斯集合，最后的分类结果可以通过各分量预测的加权平均取得：$h(x)=\sum_{k=1}^{M}\alpha_k h_k(x)$。与其它在线集合学习算法不同，算法 BEPOL 具有两个重要特征：(1)引入一错误相关参数，使得后续分类器的学习集中于被其它分类器错误分类的样本，增强了分类器之间的相关性。(2)采用了并行学习的思想，如果有合适的多处理器系统或 VLSI，算法可以方便地映射到该并行结构上去，以充分地发挥并行机的计算能力，缩短运行时间。

3.2 贝叶斯参数在线学习

为提高总体分类性能，BEPOL 算法中除对样本采用加权采样外，改进贝叶斯在线学习性能也是一重要因素。通常贝叶斯参数学习中常用的是极大似然法，给定样本集 D 和贝叶斯网结构 S，使对数似然(LL)函数 $\ln P(D,w)=\sum_{l=1}^{m}\ln P_w(D_l)$ 最大化的参数即为对未知参数 w 的极大似然估计。实际上，它与分类器优化的目标：条件对数似然(CLL)函数 $\sum_{l=1}^{m}\ln P_w(C_l|E_l)$ 并不完全一致，它们之间的关系[7]为：

$$LL_S(w|D)=CLL_S(w|D)\sum_{l=1}^{m}\ln P_w(E_l) \tag{2}$$

如果给定的贝叶斯分类器结构正确，则两者目标是一致的，反之，可能产生次优的贝叶斯分类器。因此在 BEPOL 算法中，我们以 $\ln P_w(C|E)$ 为优化目标，并结合梯度下降法对贝叶斯参数进行在线学习，应用贝叶斯公式可得：

$$\ln P_w(C|E)=\sum_{l=1}^{m}\ln P_w(C_l,E_l)-\ln P_w(E_l)) \tag{3}$$

其中，m 是训练样本的个数，E_l 为样本属性，C_l 为样本标记。

$$\frac{\partial\ln P_w(C|E)}{\partial w_{ijk}}=\sum_{l=1}^{m}\left(\frac{\partial\ln P_w(C_l)}{\partial w_{ijk}}-\frac{\partial\ln P_w(E_l)}{\partial w_{ijk}}\right) \tag{4}$$

E_l 和 C_l 均为样本 D_l 的组成部分，通过求解 $\sum_{l=1}^{m}\frac{\partial\ln P_w(D_l)}{\partial w_{ijk}}$ 代入(4)式即可求出条件对数似然函数的梯度。假定 x_{ij} 是对学习样本中变量 X_i 的第 j 个取值，u_{ik} 是变量 X_i 双亲的第 k 个取值，w_{ijk} 是变量 X_i 给定其双亲取第 k 个值时，X_i 自身取第 j 个取值的概率。

$$\sum_{l=1}^{m}\frac{\partial\ln P_w(D_l)}{\partial w_{ijk}}=\sum_{l=1}^{m}\frac{\partial P_w(D_l)/\partial w_{ijk}}{P_w(D_l)}=\sum_{l=1}^{m}\frac{P_w(D_l|x_{ij},u_{ik})P_w(u_{ik})}{P_w(D_l)}=\frac{P_w(x_{ij},u_{ik}|D_l)P_w(D_l)P_w(u_{ik})}{P_w(x_{ij},u_{ik})P_w(D_l)}=\frac{P_w(x_{ij},u_{ik}|D_l)}{w_{ijk}} \tag{5}$$

如果将函数 $\ln_w P(C|E)$ 视为以参数 w 和函数值为坐标的多维空间曲面，采用梯度下降法即可实现对贝叶斯参数的在线更新学习。

4 实验比较

为验证算法 BEPOL 的正确性和有效性，我们以朴素贝叶斯分类器为分量模型，采用 UCI 机器学习库[8]中 5 个典型的数据集进行测试。表 1 给出了对这 5 个数据集的具体描述：

表 1 实验用数据集描述

数据集	属性数	分类数	样本数
Breast Cancer	9	2	699
Mushroom	22	2	8124
Chess	36	2	3196
Waveform-40	40	3	5000
Credit-g	20	2	1000

为比较和验证算法的收敛性，我们首先从样本集 Chess

中选择了500,1000,1400,1800,2200,2600个样本组成训练集,600个样本组成测试集,比较单个朴素贝叶斯分类器,BEPOL算法,以及AdaBoost算法的学习曲线,除朴素贝叶斯分类器外,每个算法运行10次,并以这10次运行结果的平均值作为最终的结果,如图1所示。

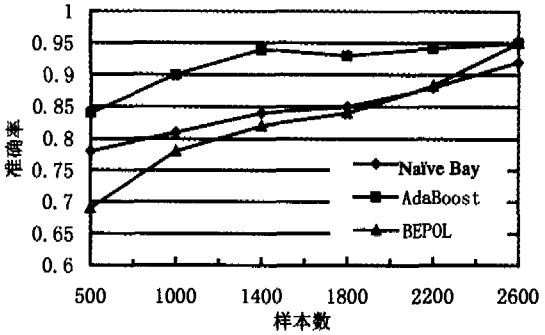

图1 基于数据集Chess的学习曲线

从图1可以看出,在线学习算法BEPOL具有与AdaBoost算法相近的分类性能,尽管算法BEPOL的初始性能不如AdaBoost,甚至比单一贝叶斯分类器更差,但随着训练样本集的增加,算法BEPOL的性能稳定增长,并最终取得与批量集合学习算法相同的分类能力。

表2是算法BEPOL与朴素贝叶斯分类器、AdaBoost算法基于样本集Breast Cancer、Mushroom、Waveform-40、Credit-g的分类性能比较结果。对朴素贝叶斯分类器和AdaBoost算法,我们采用5倍交叉验证方法,将交叉验证的平均值作为算法的有效结果。并行算法因为受样本顺序性的影响,我们采用5轮样本随机序列作为在线样本集,并以其平均值作为有效结果。

表2 BEPOL、朴素贝叶斯分类器、AdaBoost算法在UCI样本集上的分类正确率

数据集	Naïve Bayesian	AdaBoost	BEPOL
Breast Cancer	0.752	0.938	0.896
Credit-g	0.795	0.821	0.814
Waveform-40	0.842	0.826	0.837
Mushroom	0.927	0.969	0.975

通过比较我们看到,在Breast Cancer中,AdaBoost算法明显好于BEPOL,但在Waveform-40和Mushroom上,BEPOL表现则优于AdaBoost算法,在数据集Credit-g上,则性能几乎相差不大,而且我们注意到,随着样本集容量的增长,BEPOL算法的优势体现得更加明显。

结论 Boosting算法可以用于分类器集合的学习,并提高总体的分类性能。但对于连续型样本集,它需要缓存大量的数据,这会显著增加算法的时间和空间开销。对此我们提出了一种Boosting方式的在线集合学习算法,它利用与AdaBoost算法相同的样本加权采样思想,通过减少对同一样本的采样次数,满足在线学习中对样本的要求。进一步的实验证明,在数据量充足的情况下算法BEPOL具有与AdaBoost算法相近的分类能力,并可同时用于在线学习和批量学习过程。

参考文献

1 Shi X, Manduchi R. A study on bayes feature fusion for image classification. In: Workshop on Statistical Analysis in Computer Vision, (Madison, WI), 2003

2 Bauer E, Kohavi R. An empirical comparison of voting classification algorithms: bagging, boosting, and variants. Machine Learning, 1999, 36 (1-2): 105~139

3 Kivinen J, Warmuth M K. Additive versus exponentiated gradient updates for linear prediction. 2000

4 Fern A, Givan R. Online ensemble learning: An empirical study. In: Proc. of the Seventeenth International Conference on Machine Learning, 2000. 279~286

5 Zhang T. On Sequential greedy approximation for certain convex optimization problems: [Technical report]. IBM T. J. Waston Research Center, 2002

6 Freund Y, Schapire R E. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. Computer and System Sciences, 1997, 55(1):119~139

7 Chelba C, Acero A. Conditional maximum likelihood estimation of naive Bayes probability models using rational function growth transform: [Technical Report MSR-TR-2004-33]. Microsoft, 2004

8 Blake C, Keogh C J M. UCI repository of machine learning databases. 1999

(上接第140页)

11 Demers A, Greene D, Hauser C, et al. Epidemic algorithms for replicated database maintenance. In: 6th ACM Symposium on Principles of distributed computing. New York: ACM Press, 1987. 1~12

12 Saito Y. Consistency management in optimistic replication algorithms. http://www.hpl.hp.com/personal/Yasushi_Saito/replica.pdf, June 2001

13 Adly N. Management of replicated data in large scale systems: [PhD dissertation]. UK: Corpus Christi College, University of Cambridge, August 1995

14 Golding R A. Modeling replica divergence in a weak-consistency protocol for global-scale distributed data bases: [Technical Report UCSC-CRL-93-09]. University of California at Santa Cruz, February 1993

15 Saito Y, Levy H, Bershad B H. Manageability, availability and performance in Porcupine: a highly scalable, cluster-based mail service. ACM Transactions on Computer Systems, 2000, 18(3): 298~332

16 Saito Y, Levy H M. Optimistic replication for Internet data services. In: 14th International Conference on Distributed Computing. London: Springer-Verlag, October 2000. 297~314

17 Thomas R H. A majority consensus approach to concurrency control for multiple copy databases. ACM Transactions on Database Systems, 1979, 4(2): 180~209

18 Terry D B, Theimer M M, Petersen K, et al. Managing update conflicts in Bayou, a weakly connected replicated storage system. In: 15th ACM Symposium on Operating Systems Principles. New York: ACM Press, December 1995. 172~183

19 Keleher P J. Decentralized replicated-object protocols. In: 18th ACM symposium on Principles of distributed computing. New York: ACM Press, May 1999. 143~151

20 Birman K P, Joseph T A. Reliable communication in the presence of failures. ACM Transactions on Computer Systems, 1987, 5 (1): 47~76

21 Qin Xiao, Jiang Hong. Data Grid: Supporting Data-Intensive applications in Wide-Area Networks: [Technical Report TR-03-05-01]. University of Nebraska-Lincoln, May 2003

22 Lamehamedi H, Szymanski B, shentu Z, et al. Data replication strategies in Grid environments. In: Proceedings of the Fifth International Conference on Algorithms and Architectures for Parallel Processing (ICA3PP'02). Oakland: IEEE Computer Press, 2002. 378~383

23 Cai Min, Chervenak A, Frank M. A peer-to-peer replica location service based on a distributed hash table. Proceedings of the ACM/IEEE SC2004 Conference. Pittsburgh, Pennsylvania, 2004