计算机研究与发展
Journal of Computer Research and Development
ISSN 1000-1239/CN 11-1777/TP
43(7):1161~1166, 2006

小规模数据集的神经网络集成算法研究

李 凯[1] 黄厚宽[2]
[1](河北大学数学与计算机学院 保定 071002)
[2](北京交通大学计算机与信息技术学院计算智能研究所 北京 100044)
(likai@mail.hbu.edu.cn)

Study of a Neural Network Ensemble Algorithm for Small Data Sets

Li Kai[1] and Huang Houkuan[2]
[1](*School of Mathematics and Computer Science, Hebei University, Baoding* 071002)
[2](*Institute of Computational Intelligence, School of Computer and Information Technology, Beijing Jiaotong University, Beijing* 100044)

Abstract Ensemble learning has become a hot topic in machine learning. It dramatically improves the generalization performance of a classifier. In this paper, neural network ensemble for small data sets is studied and an approach to neural network ensemble (*Novel_NNE*) is presented. For increasing ensemble diversity, a diverse data set is generated as part training set in order to create diverse neural network classifiers. Moreover, different combinational methods are studied for *Novel_NNE*. Experimental results show that *Novel_NNE* for both the relative majority vote method and the Bayes combinational method achieves higher predictive accuracy.

Key words neural network ensemble; small data set; diversity; generalization

摘 要 研究了小样本数据集的神经网络分类器集成,提出了适合于小样本数据集的神经网络分类器集成方法 *Novel_NNE*,通过生成差异数据提高神经网络集成中个体的差异性,从而提高集成学习的泛化性能;最后应用不同的融合技术针对 UCI 标准数据集进行了实验研究. 结果表明,在集成算法 *Novel_NNE* 中,使用相对多数投票与贝叶斯融合方法的性能优于行为知识空间融合方法.

关键词 神经网络集成;小规模数据集;差异性;泛化

中图法分类号 TP18

1 引 言

集成学习已经成为机器学习的研究热点[1],其目的是利用多个模型的差异性提高学习系统的泛化性能. 目前研究人员已经提出了许多集成方法,大致可以归纳为如下几种:①从训练集中提取子样;②操作输入特征;③操作输出目标;④引入随机性方法;⑤具体的集成算法. 神经网络做为一种机器学习方法,已经成功应用于许多领域,但由于神经网络是一种不稳定的学习方法(即当训练数据集有较小的变化,则产生的模型有很大的不同),从而导致神经网络模型的泛化性能不同. Breiman[2] 对该问题进行了研究,提出了不同模型的产生是由于在神经网络训练时到达了误差面的不同局部最小值. 为了克服这些问题,研究人员基于交叉验证技术提高单个神经网络的泛化性能;另外一些研究人员采用选择最优模型的方法提高神经网络的泛化性能. 但是

收稿日期:2004-09-13;修回日期:2005-10-10
基金项目:国家自然科学基金项目(60443003);河北大学博士基金项目(075)

这些方法存在一些弊端:①从一组神经网络中选择出一个网络,将会丢失存储在被丢弃网络中的信息;②由于交叉验证数据集具有随机性,可能出现这组网络中的某个网络在另外的数据集上具有更好的性能. 1990 年,Hansen 与 Salamon[3] 开创性地提出了神经网络集成方法,之后很多研究人员进行了这方面的研究. 1995 年,Krogh 与 Vedelsby[4] 提出了神经网络集成泛化误差的计算公式 $E=\bar{E}-\bar{A}$,其中 $\bar{E}$ 为各个神经网络的泛化误差的加权平均,$\bar{A}$ 为神经网络集成的差异度. 由此可知,只要个体神经网络的泛化误差均值保持不变,增加差异性将会提高神经网络的泛化性能. 基于这种思想,研究人员提出了不同的神经网络集成算法,这些集成方法大致分为两种类型:显式集成方法与隐式集成方法. 显式集成方法是通过直接生成一组误差(分类错误)独立的网络. 例如 Opitz 与 Shavlik[5] 提出的 ADDEMUP 算法;Rosen[6] 提出的使用 BP 方法训练神经网络的集成方法,这种方法不仅产生希望的输出,而且使得这些网络产生的错误是不相关的. 隐式集成方法是通过生成与选择策略来确定神经网络集成的成员,也就是开始时生成许多神经网络,通常这些神经网络的独立性是很弱的;然后使用选择策略从中选取具有最大独立性的神经网络. 例如 Partridge 与 Yates[7] 基于这种设计策略提出的启发式选择集成方法;Zhou 等人[8] 使用遗传算法选择神经网络集成中的个体方法. 最近 Imamura 等人[9] 应用遗传规划方法进行分类器的集成.

构造一个使每个神经网络尽可能不同的集成,是集成方法所具有的重要特性. 在集成学习中,Bagging 与 Boosting 是较流行的学习方法,它们通过重取样或对训练样本加权的方法提供集成的差异性.然而当训练集规模较小时,这些方法对集成差异性具有一定的限制作用. 在本文中,我们研究了神经网络作为分类器模型的集成方法,提出了适应于小数据集的神经网络集成算法 *Novel_NNE*,并且使用不同的融合技术研究了算法 *Novel_NNE* 的有效性.

2 Bagging 与 Boosting 集成方法

2.1 Bagging 集成方法

在 Bagging 集成方法中,训练每个分类器所使用的数据集是通过有放回随机抽样技术获得的,通过这种方法得到的训练集称为 Bootstrap 训练集. 每个 Bootstrap 训练集平均含有 63.2% 的原始训练集中的数据. 在 Bootstrap 训练集中,原始训练集中的数据在 Bootstrap 训练集中可能出现多次也可能一次也不出现,在新的实例上的预测采用多数投票方法. 集成成员间的差异性是通过 Bootstrap 重取样技术获得的. Bagging 集成方法主要用于不稳定的学习算法,即当训练集中数据发生小的变化时,则会导致模型很大的变化. 因为集成中每个分类器模型不受相同实例集的影响,所以它们彼此是不同的. 通过对这些分类器的预测进行投票,以寻求减少集成学习的泛化误差. 对于稳定的学习算法,例如朴素贝叶斯方法,Bagging 集成并不能减少误差.

2.2 Boosting 集成方法

Boosting 方法有很多种变形,AdaBoost 是最流行的一种方法. 在这种算法中,假设学习算法能够处理加权实例. 若学习算法不能直接处理加权实例,则按照权分布对训练集取样以产生新的训练集. AdaBoost 对训练实例的权进行维护,并且在第 i 次迭代中,通过最小化训练集的加权误差来训练分类器 C_i,然后使用分类器 C_i 的加权误差更新训练实例上的权分布. 通过这种方法,使得误分实例的权值增加,而正确分类实例的权值减少,在训练下一个分类器时,则利用更新实例的权值分布,并重复此过程.在训练之后,使用个体分类器的一种加权 $\sum_i w_i C_i(x)$ 投票做出集成的预测,每个分类器的权是通过它使用的训练集上的加权实例上的正确率来计算的. 在 Boosting 集成方法中,集成成员间的差异性是通过实例的不同权分布获得的.

AdaBoost 方法的主要缺陷是:当数据量不足或当有大量的分类噪声(即具有不正确的类标号的训练实例)时,执行效果较差.

3 小样本数据集的神经网络集成算法 *Novel_NNE*

在集成算法 *Novel_NNE* 中,将集成中个体分类器的预测与集成的不一致性作为一种差异性度量[10,11],主要通过生成差异数据[12] 提高神经网络之间的差异性. 首先按照给定数据集的分布(实际上是一个近似分布)生成额外数据,即对于连续属性的数据,计算均值与方差;而对于非连续属性的数据,计算它们的概率分布. 通过这种方法可以获得与原数据集近似同分布的数据. 具体方法如下:

首先利用原数据集 T 训练神经网络,由此可以获得一个分类器 C_i,并将其作为集成中的一个成员,也就是说,到目前为止,集成中的分类器个数为 1;然后根据数据集 T 的数据特性生成额外数据,设获得的这些数据为 R. 使用现有的集成分类器对数据集 R 分类,这样对于数据集 R 中的每一个数据点可以获得一个隶属各类的一个概率分布,若需要确定每个数据点的具体类别,则可求每个数据点所属类别的最大概率值,它所对应的类别即为该数据点的类别. 为了使得以后新生成的神经网络分类器与集成中的分类器有较大的差异性,对生成的数据集 R 中的每个数据点隶属各个类的概率求倒数,即训练新的神经网络使用的这些数据所属类别与集成分类器对它们分类的类别成反比. 当确定了生成数据的类别后,与数据集 T 一起作为训练新的神经网络的数据集,设使用这个数据集获得的分类器为 C'. 另外,为了保证集成方法的正确率,对新生成的分类器 C' 采用了如下的方法:将分类器 C' 加入到集成中,计算集成分类器在数据集 T 上的分类错误率,若这个错误率小于未加入 C' 时的错误率,则 C' 作为集成中一个成员,否则就丢弃 C'. 最后进入下一轮,直至达到集成规模要求或达到规定的迭代次数为止. 具体算法如下:

Novel_NNE(T, *C_size*)/* T:训练数据,*C_size*:集成规模,*Max_iteration*:最大迭代次数 */

Step1. 对集成规模及迭代次数赋初值:$i=1$, $iterations=1$;

Step2. 使用数据集 T 训练神经网络,设获得的分类器为 C_i,$C_i = NN(T)$;

Step3. 将分类器 C_i 加入到集成中 $C^* = \{C_i\}$;

Step4. 计算集成分类器在数据集 T 上的错误率 $\varepsilon = \frac{\sum 1}{m}$, $x_j \in T, C^*(x_j) \neq y_j$;

Step5. 当 $i < C\_size$ 且 $iterations < Max\_iteration$ 时,重复 Step6 至 Step13;

Step6. 利用数据集 T 的数据分布生成数据集 R,数据集 R 中数据点的个数由比例因子 α 确定,$R = Data\_Generation(\alpha, T)$;

Step7. 使用获得的局部集成分类器对数据集 R 分类,其结果为每个数据点隶属各类的概率分布 P,$P = Local\_Ensemble\_Classification(C^*, R)$;

Step8. 为了产生具有差异的神经网络,利用概率分布 P,产生一个与分布 P 互为倒数的概率分布,由此分布确定生成数据集的类标号,即 $R\_label = Set\_Class\_Label(C^*, R)$;

Step9. 将新生成的数据集 R_label 与数据集 T 合并组成新的数据集 T,$T = T \cup R\_label$;

Step10. 使用新数据集 T 训练神经网络,设获得的分类器为 C',$C' = NN(T')$;

Step11. 将新生成的分类器 C' 加入到集成,$C^* = C^* \cup \{C'\}$;

Step12. 从数据集 T 去除新生成的数据集,$T = T - R\_label$;计算集成分类器在数据集 T 上的错误率,$\varepsilon' = \frac{\sum 1}{m}$, $x_j \in T, C^*(x_j) \neq y_j$;

Step13. 若加入分类器 C' 后集成分类器在数据集 T 上的错误率小于未加入分类器 C' 后集成分类器的错误率,则在集成中保留分类器 C',否则就从集成中剔除该分类器;

if $\varepsilon' \leqslant \varepsilon$ then $i = i+1$, $\varepsilon = \varepsilon'$ else $C^* = C^* - \{C'\}$;

$iterations = iterations + 1$.

Data_Generation(α, T)[①]/* 利用数据集 T 的分布生成新的数据集,其中 α 为比例因子,$|\cdot|$ 表示集合的势,$\alpha|T|$ 为实际生成的数据个数 */

{

重复下面的操作,直至达到规定的数据个数 $\alpha|T|$ 为止:

若属性为连续性属性,则计算数据集 T 的每个连续属性的均值 *Mean* 与方差 *Variance*,然后按照高斯分布生成新的数据集 *Data_con*,*Data_con* = *Gaussion*(*Mean*, *Variance*);

若属性为离散属性,则计算它们的概率分布 *P_feature*,然后按照这个分布生成数据集 *Data_nocon*,*Data_nocon* = *generation*(*P_feature*);

将各个属性数据 *Data_con* 与 *Data_nocon* 组成新的数据点 *Data*;

返回获得的数据集.

}

Local_Ensemble_Classification(C^*, R)/* 使用局部集成 C^* 对数据集 R 分类,得到每个数据点的概率分布 */

① 这里假设属性之间彼此独立,若不独立,则利用相关性也可确定.

{

将集成 C^* 中每个分类器对数据集 R 中每个数据点 x_j 分类,由此可得每个数据点 x_j 隶属各类的概率分布 $P_y^{C_i}(x_j)$;

选择一种融合方法 F 集成,这样可得集成分类器对数据集 R 中每个数据点隶属各类的概率分布 P_y,

$$P_y(x_j) = F(P_y^{C_1}(x_j), P_y^{C_2}(x_j), \cdots, P_y^{C_{|C^*|}}(x_j)),$$
$$j = 1,2,\cdots,|R|;$$

return P_y.

}

Set_Class_Label(C^*, R, P_y)/* 为了得到差异较大的神经网络,将集成分类器对集合 R 中的每个数据点的分类结果求倒数,以获得数据集 R 中每个数据点的概率分布 */

{

for $k = 1,2,\cdots,|R|$

$$P'_y(x) = \frac{1/P_y(x)}{\sum_y 1/P_y(x)};$$

return P'_y.

}

当集成分类器中的每个成员确定后,采用融合方法 F 对各分类器的分类结果融合.

Classify(C^*, x)

{

$P_y(x) = F(P_y^{C_1}(x), P_y^{C_2}(x), \cdots, P_y^{C_{c\_size}}(x))$;

$C^*(x) = \arg\max_{y \in Y} P_y(x)$.

}

4 实验结果及分析

实验中选择了 UCI 的 10 个数据集[13],以表明集成算法 *Novel_NNE* 的有效性. 训练神经网络使用的学习算法为 BP 算法,每个神经网络具有一个隐层,该隐层含有 10 个隐单元,BP 算法中其他参数(例如学习率等)采用 Matlab 中的默认值. 融合函数 F 分别采用贝叶斯平均集成方法(BAM)、相对多数投票方法(RMV)与行为知识空间方法(BKS)[14,15]. 集成规模设置为 15,最大迭代次数设置为 60,α 在 0.1~2 间取值. 实验中使用了十折交叉验证技术,实验结果见表 1 和表 2,表中的数值是每种方法的错误率,数据集名右上角的"*"号表示该数据集中的数据含有空缺值.

Table 1 Experimental Results (error rates) of *Novel_NNE* and Comparison with Bagging and Adaboost

表 1 *Novel_NNE* 算法实验结果及与 Bagging 与 AdaBoost 算法的比较

Data Set	Error Rate	
	Novel_NNE/Bagging	*Novel_NNE*/AdaBoost
breast-w	4.1/3.4	4.1/4.0
glass	32.8/33.1	32.8/31.1
iris	4.1/4.0	4.1/3.9
segment	5.6/5.4	5.6/3.3
ionosphere	9.1/9.2	9.1/8.3
soybean	6.8/6.9	6.8/6.3
house-votes-84	4.0/4.1	4.0/5.3
hepatitis	17.6/17.8	17.6/19.7
wdbc	9.8/9.95	9.8/11.1
wpbc	10.2/10.5	10.2/12.4

表 1 中的每一行数据分别是 *Novel_NNE* 算法与另一个算法(或者是 AdaBoost 或者是 Bagging)的分类错误率. 集成方法 *Novel_NNE* 在处理小数据集及具有空缺值的数据集效果更好,例如 *breast-w*, *wpbc*, *house-votes-84* 及 *hepatitis* 数据集. 另外,实验中使用双尾配对 t 检验比较了两种分类方法的结果,win,loss 与 draw 分别表示在显著性水平 0.05 下,集成分类器 *Novel_NNE* 与另外两种典型集成方法(Bagging/AdaBoost)相比,其分类正确率明显优于、明显低于以及基本相当的数据集个数分别为 7/5,3/5,0/0. 可以看出,在 10 个数据集中,*Novel_NNE* 分类器在 7 个数据集上的分类正确率高于 Bagging 分类器;而在 5 个数据集上的分类正确率高于 AdaBoost 分类器. 正如第 2.1 节与第 2.2 节所分析的那样,Bagging 与 AdaBoost 集成方法处理小数据集并不能显著提高分类器的正确率,有时还可能使分类器的正确率降低,其主要原因是集成成员之间的差异性较小;而集成算法 *Novel_NNE* 在处理小规模数据集时通过生成差异数据提高集成成员之间的差异性,同时采用选择方法保证集成成员的正确率. 表 1 中给出的是 5 次实验结果的均值,统计检验表明了这种方法的稳定性较好.

表 2 是使用 BAM,RMV 与 BKS 融合技术获得的结果. 使用 RMV 方法与 BAM 方法获得的结果接近;而使用行为知识空间方法(BKS)时,由于需要确定阈值,所以这种融合方法在很大程度上依赖于该阈值的选择,同时这种方法需要较多的训练数据,因此,使用 BKS 获得的结果略差于 BAM 与 RMV 融合方法.

Table 2 Comparison with *Novel_NNE* in Different Combinational Methods

表 2 集成算法 *Novel_NNE* 使用不同融合方法的实验比较

Data Set	Combinational Methods		
	BAM	RMV	BKS
breast-w *	4.1	4.2	4.3
glass	32.8	32.9	33.1
iris	4.1	4.2	4.3
segment	5.6	5.5	5.6
ionosphere	9.1	9.3	9.2
soybean	6.8	7.0	7.2
house-votes-84 *	4.0	4.1	4.3
hepatitis *	17.6	17.4	17.6
wdbc	9.8	9.9	10.0
wpbc *	10.2	10.1	10.3

为了研究集成算法 *Novel_NNE* 的性能与集成规模间的关系,对规模为 3,5,10,15,20,25,30,35,40,45,50,60 的集成进行了实验研究,图 1 给出了集成规模与集成正确率间的关系.

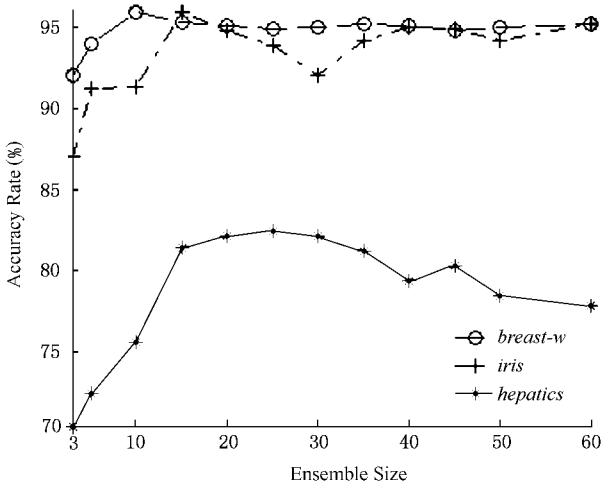

Fig. 1 Relation between ensemble size and accuracy rate.

图 1 集成规模与正确率间的关系

由图 1 可以知道,开始时集成的正确率是随着集成规模的增大而增加的,一般情况下,当集成规模处于 10 至 25 之间时,集成的正确率最高,当在这个区间以外,集成的正确率开始下降. 然而,不同的集成方法会影响这个集成规模范围.

5 结 论

我们分析了两种重要的集成技术 AdaBoost 与 Bagging,特别是对于小数据集的集成问题,指出了这两种方法的不足;针对这些问题,我们研究了针对小数据集提高神经网络差异性的集成算法,同时研究了集成算法使用不同融合方法的性能. 另外,我们针对 10 个数据集进行了实验研究,结果表明,集成算法 *Novel_NNE* 对于小规模数据集要优于 AdaBoost 及 Bagging 方法;同时也研究了集成规模与集成性能之间的关系.

参 考 文 献

1 Thomas G. Dietterich. Machine learning research: Four current directions [J]. AI Magazine, 1997, 18(4): 97~136

2 L. Breiman. Bagging predictors [J]. Machine Learning, 1996, 24(2): 123~140

3 Lars Kai Hansen, Peter Salamon. Neural network ensembles [J]. IEEE Trans. Pattern Analysis and Machine Intelligence, 1990, 12(10): 993~1001

4 Anders Krogh, Jesper Vedelsby. Neural network ensembles, cross validation, and active learning [G]. In: G. Tesauro, D. S. Touretzky, T. K. Leen, eds. Advances in Neural Information Processing Systems 7. Cambridge MA: MIT Press, 1995. 231~238

5 David W. Opitz, Jude W. Shavlik. Actively searching for an effective neural-network ensemble [J]. Connection Science, 1996, 8(3): 337~353

6 B. Rosen. Ensemble learning using decorated neural networks [J]. Connection Science, 1996, 8(3): 373~384

7 Derek Partridge, W. B. Yates. Engineering multiversion neural-net systems [J]. Neural Computation, 1996, 8(4): 869~893

8 Zhi-Hua Zhou, Jianxin Wu, Wei Tang. Ensembling neural networks: Many could be better than all [J]. Artificial Intelligence, 2002, 137(1/2): 239~263

9 Kosuke Imamura, Terence Soule, Robert B. Heckendorn, *et al*. Behavioral diversity and a probabilistically optimal GP ensemble [J]. Genetic Programming and Evolvable Machines, 2003, 4(3): 235~253

10 L. Kuncheva, C. Whitaker. Measures of diversity in classifier ensembles and their relationship with ensemble accuracy [J]. Machine Learning, 2003, 51(2): 181~207

11 P. Cunningham, J. Carney. Diversity versus quality in classification ensembles based on feature selection [C]. The 11th European Conf. Machine Learning, Barcelona, Catalonia, Spain, 2000

12 P. Melville, R. J. Mooney. Diversity ensembles for active learning [C]. ICML 2004, Banff, Canda, 2004

13 C. L. Blake, C. J. Merz. UCI repository of machine learning database [OL]. http://www.ics.uci.edu/~mlearn/MLRepository.html, 1998

14 Y. S. Huang, C. Y. Suen. Combination of multiple experts for the recognition of unconstrained handwritten numerals [J]. IEEE Trans. Pattern Analysis and Machine Intelligence, 1995, 7(1): 90~94

15 L. Xu, C. Krzyzak, C. Suen. Methods of combining multiple classifiers and their applications to handwritting recognition [J]. IEEE Tran. Systems, Man and Cybernetics, 1992, 22(3): 418~435

Li Kai, born in 1963. Received his Ph. D. degree in Computer Science and Information Technology from Beijing Jiaotong University in 2005. His main research fields include neural networks, clustering, and data mining, etc.

李凯,1963 年生,博士,主要研究方向为神经网络、聚类及数据挖掘等.

Huang Houkuan, born in 1940. Professor and Ph. D. supervisor. He has published more than 100 journal and conference publications, senior member of CCF. His main research fields include artificial intelligence, machine learning, data warehousing, data mining, decision support system, and multi-agent system, etc.

黄厚宽,1940 年生,教授,博士生导师,在期刊及会议上共发表了 100 多篇论文. 中国计算机学会高级会员,主要研究方向为人工智能、机器学习、数据仓库、数据挖掘、决策支持系统和多 Agent 系统等.

Research Background

This research work is supported by the National Natural Science Foundation of P. R. China (Grant No. 60443003) and Doctoral Research Fund for Hebei University.

Ensemble learning has become a hot topic in machine learning. It dramatically improves the generalization performance of classifier. Nowadays, there exist many popular ensemble methods such as Bagging, Adaboost, etc. However, there are some problems with small data sets for these ensemble methods. In this case, we attempt to study neural network ensemble for small data sets and present an approach to neural network ensemble (*Novel_NNE*). For increasing ensemble diversity, a diverse data set is generated as part training set in order to create diverse neural network classifier. Moreover, to ensure high accuracy of ensemble, the performance of ensemble is tested when a neural network classifier is added to ensemble. Finally, we experiment on UCI data sets with different combinational methods. Experimental results show that *Novel_NNE* achieves higher predictive accuracy in both relative majority vote method and Bayes method than in the behavioral knowledge space method.

The new method has the merits what Bagging and Adaboost has, such as high predictive accuracy rate in dealing with big data sets. In addition, *Novel_NNE* has the merits in dealing with small data sets. Experimental results are given to verify the effectiveness of *Novel_NNE*. Further research on the method will go on in the future.