

一种基于聚类技术的选择性神经网络集成方法

李 凯^{1,2} 黄厚宽¹

¹(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)

²(河北大学数学与计算机学院 保定 071002)

(likai_njtu@163.com)

A Selective Approach to Neural Network Ensemble Based on Clustering Technology

Li Kai^{1,2} and Huang Houkuan¹

¹(Institute of Computational Intelligence, School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044)

²(School of Mathematics and Computer Science, Hebei University, Baoding 071002)

Abstract A neural network ensemble is a very popular learning paradigm where the outputs of a set of separately trained neural network are combined to form one unified prediction. To improve the effectiveness of ensemble, neural networks in the ensemble are not only highly correct but make their errors on different parts of the input space as well. However, most existing approaches ensemble all the available neural networks for prediction. In this paper, a selective approach to neural network ensemble based on clustering technology is presented. After individual neural networks are trained, the clustering algorithm is used to select a part of the trained individual networks in order to reduce their similarity. Then many selected neural networks are combined. Experimental results show that this approach outperforms the traditional ones that ensemble all of the individual networks.

Key words neural network; ensemble; clustering

摘 要 神经网络集成是一种很流行的学习方法,通过组合每个神经网络的输出生成最后的预测.为了提高集成方法的有效性,不仅要求集成中的个体神经网络具有很高的正确率,而且要求这些网络在输入空间产生不相关的错误.然而,在现有的众多集成方法中,大都采用将训练的所有神经网络直接进行组合以形成集成,实际上生成的这些神经网络可能具有一定的相关性.为了进一步提高神经网络间的差异性,一种基于聚类技术的选择性神经网络集成方法 CLU-ENN 被提出.在获得个体神经网络后,并不直接对这些神经网络集成,而是先应用聚类算法对这些神经网络模型聚类以获得差异较大的部分神经网络,然后由部分神经网络构成集成;最后,通过实验研究了 CLU-ENN 集成方法,与传统的集成方法 Bagging 相比,该方法取得了更好的效果.

关键词 神经网络;集成;聚类

中图法分类号 TP18

1 引 言

神经网络集成是由 Hansen 与 Salamon^[1]在 1990 年提出的,旨在通过训练多个神经网络并将其进行组合来提高神经网络系统的泛化能力.由于该

方法的效果明显,因此受到了很多学者的青睐.1996 年 Sollich 与 Krogh^[2]对神经网络集成给出一个定义,即“神经网络集成是用有限个神经网络对同一个问题进行学习,且集成在某个输入实例下的输出是由构成集成的各个神经网络在该实例下的输出共同决定的”.目前,这个定义已被广泛接受.由于神经

神经网络集成所蕴涵的巨大潜力及广阔的应用前景,自 Hansen 与 Salamon 之后,很多学者进行了这方面的研究.

本文提出了一种新的构造神经网络集成的方法,在独立地训练出一批神经网络之后,通过聚类技术选择差异度较大的神经网络,用这些网络来组成神经网络集成.实验结果表明,这种方法是非常有效的.

2 神经网络的集成

神经网络集成由两部分组成,即个体生成方法与结论生成方法.个体生成方法也称为训练基模型的方法,在理想情况下,要求获得的基模型可在各自的输入空间产生错误,而网络集成后却能够对所有样本正确分类,如图 1 所示,其中直线 A、B 和 C 分别是线性分类器,黑粗体是 3 个线性分类器的集成,“+”表示线性分类器的正例,“-”表示线性分类器的负例.可以看到,当 3 个线性分类器集成后能够对所有样本正确分类,而对任一线性分类器则不能做到这一点.

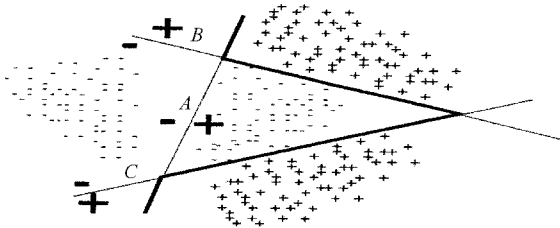


Fig. 1 Ensemble of linear classifiers.

图 1 线性分类器的集成

然而,在一般情况下,却很难满足上述条件.为此研究人员提出了许多算法试图生成一组基模型,并使得它们产生的错误尽可能是不相关的,诸如 Bagging 方法^[3]、Boosting^[4]方法、交叉验证划分^[5]方法、ECOC^[6]方法、ID^[7,8]方法与 SASC^[9]方法.另外,著名学者周志华教授等^[10~12]对神经网络集成也进行了深入研究,提出了 GASEN 集成方法,并受到了神经网络集成界的广泛关注.结论生成方法也称为组合基模型的方法.最基本的组合方法是多数表决法.若分类器以概率值的形式输出,则简单平均方法与加权平均方法也是一种有效的组合方法.目前,除了这些组合方法外,也存在另外一些非线性集成方法,诸如 RBC^[13]、BBM^[14]与 OSC^[15].

Bagging 的基础是可重复取样,在该方法中,各

神经网络的训练集是通过从原始训练集中随机选取若干实例组成的,训练集的规模通常与原始训练集相当,训练实例允许重复选取.这样原始训练集中某些实例则可能在新的训练集中出现多次,而另外一些实例则可能一次也不出现.在这种方法中,通过重新选取训练集来增加神经网络集成的差异度,从而提高泛化能力. Krogh 等人^[5]通过理论研究发现,组成神经网络集成的各网络差异越大,集成的效果越好.然而使用 Bagging 技术训练的神经网络通常是各自尽可能独立地训练,网络的差异由使用的训练集、网络结构、学习算法的差异或者学习算法的随机性来产生,因此有可能产生彼此非常相似的神经网络,这些神经网络对于整个神经网络集成的泛化能力的提高可能并不起作用.

为了进一步增加神经网络集成中基模型的差异,提高整个神经网络集成的泛化能力,本文研究了采用聚类技术提高集成中基模型的差异.由于本文主要研究由基模型集成的泛化能力,对聚类所花费的时间并未考虑,因此我们选择了非常通用的硬 C 均值聚类算法,当然也可以采用其他的聚类方法.由于聚类算法与数据集的分布是有关的,对于同一个数据集,选择不同的聚类算法及不同的参数,则聚类结果及所用时间可能是不同的,因此,模型聚类算法的选择有待于我们进一步研究.

3 神经网络的聚类集成算法 CLU-ENN

神经网络集成是用有限个神经网络对同一个问题进行学习,对于某个输入来说,它在集成上的输出是由集成中的各个神经网络在该输入上的输出共同决定的.神经网络集成方法可应用于很多领域,本文仅考虑神经网络集成应用于回归分析情况.对于回归型神经网络,集成目的可视为模拟函数 $f: R^m \rightarrow R^n$. 为便于讨论,可假设 $n = 1$.

3.1 神经网络的聚类

假设应用 Bootstrapping 技术及网络训练方法生成了 N 个模型,每个模型通过参数 $w_i (i = 1, \dots, N)$ 来描述,另外定义 c 个模型,称为聚类中心,由 $m_j (j = 1, \dots, c)$ 描述.设样本集为 $S = \{(x^1, y_1), (x^2, y_2), \dots, (x^n, y_n)\}$. 针对要聚类的模型对象,首先定义模型之间的差异(神经网络差异),它可以通过模型间的距离^[16](实际上,不一定是距离)来度量,即

$$D(w_i, m_j) = \sum_k d(y(w_i, x^k), y(m_j, x^k)), (1)$$

其中 $d(r_1, r_2) = \|r_1 - r_2\|_2$ 为距离度量, $y(w_i, x^k)$ 与 $y(m_j, x^k)$ 分别是参数为 w_i 和 m_j 的模型在输入 x^k 上的输出.

$D(w_i, w_j)$ 具有如下特性:

- ① 不具有对称性, 即 $D(w_i, w_j) \neq D(w_j, w_i)$;
- ② $D(w_i, w_j) \geq 0$ 且 $D(w_i, w_i) = 0$.

如上定义的模型间的距离 $D(w_i, w_j)$, 实际上是通过模型在输入数据集上的输出来定义模型间的不同, 此距离越大, 则说明两个模型相差越大, 反之, 则相差很小. 在聚类中, 当两个模型相差很大时, 说明这两个模型有可能在不同的簇中, 相反, 则在同一个簇中. 由于在同一个簇中的模型都是相似的模型(按照上面的距离度量), 也就是说, 这些模型在相同的输入下得到的输出结果相似, 因此, 可以选择模型聚类中心(也可选择其他的模型)表示该簇中的所有模型, 通过这种方法, 可以进一步提高不同模型之间的差异.

基于上面的思想, 为了对模型聚类, 将每个模型在样本点 $x^k(k = 1, 2, \dots, n)$ 上的输出组成一个向量, 即 $z_i = (y(w_i, x^1), y(w_i, x^2), \dots, y(w_i, x^n))$ ($i = 1, 2, \dots, N$), 这样可得到 N 个维数为 n 的样本, 这些样本是输出空间上的样本. 现在对它们应用硬 C 均值聚类算法, 假设得到的聚类中心为 $v_j(j = 1, 2, \dots, c)$, 其中 $v_j = (v_j^1, v_j^2, \dots, v_j^n)$. 可以看到, 经过硬 C 均值聚类算法得到的是模型输出空间上的中心, 这种聚类中心与集成中所使用的模型是不同的, 由于最终目的是确定聚类后的模型, 因此, 需要将输出空间上的每个聚类中心经过重新训练转换成实际模型. 首先, 对每个聚类中心与输入样本进行组合, 获得 c 组样本集, 即

$$\{(x^i, v_j^i)\}_{i=1}^n, j = 1, 2, \dots, c.$$

然后将这 c 组样本集分为训练集与验证集, 应用与本节所使用的相同网络训练方法来训练网络模型. 在训练过程中, 当在训练集上的均方误差达到最小且当在验证集上的误差开始增加时, 则训练过程结束. 通过这种方法可以获得 c 个模型.

3.2 神经网络的集成

按照第 3.1 节介绍的聚类算法, 假设模型聚类中心为 $\{m_j\}_{j=1}^c$, 模型输出为 $y(m_j, x^k)$ ($j = 1, \dots, c; k = 1, \dots, n$).

应用简单平均方法, 则根据网络模型聚类的结果得到的集成模型为(相对于输入 x 的输出):

$$y_{ensemble} = \frac{1}{c} \sum_{j=1}^c y(m_j, x).$$

4 实验结果及分析

为了验证集成方法 CLU-ENN 的有效性, 我们进行了两组实验. 一组是针对 Freidman # 1 与 Mexican Hat^[10]数据集; 另一组是 Boston housing^[17]数据集. 这些数据集都为回归分析型的. 对于 Freidman # 1 与 Mexican Hat 数据集, 首先通过 Bootstrapping 技术独立地训练出 20 个(这里有意选择了 20 个网络, 主要与文献[10]做比较)BP 网络(1 个隐层, 有 6 个神经元), 根据聚类方法选择部分神经网络集成. 注意到, 由于聚类方法涉及到许多因素, 例如数据的分布、聚类算法中参数的选择与聚类有效性等问题, 实验中使用聚类技术进行集成的结果会有一些偏差. 为了进行比较, 我们也列出了使用 Bagging 方法(即使用全部 20 个神经网络组成的集成)及使用 GASEN^[10, 12]方法得到的结果. 实验中, 对于每个数据集使用了 3 种方法共进行了 6 次实验. 另外, 针对 Boston housing 数据集, 我们也研究了簇数对聚类集成的影响. 首先通过 Bootstrapping 技术独立地训练出 300 个 BP 网络(1 个隐层, 有 14 个神经元), 然后按照聚类方法选择部分神经网络集成.

数据集 Freidman # 1 中的每个输入向量含有 5 个分量, 数据集根据如下式子产生:

$$t = 10\sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5 + \epsilon,$$

ϵ 服从正态分布 $N(0, 1)$, $x_i(i = 1, 2, \dots, 5)$ 服从 $[0, 1]$ 上的均匀分布. 数据集共包含 1400 个实例, 训练集大小为 400, 其中 200 个实例作为验证集, 测试集大小为 1000. 运行结果如图 2 所示. 其中 x 轴为实验序号, y 轴为误差.

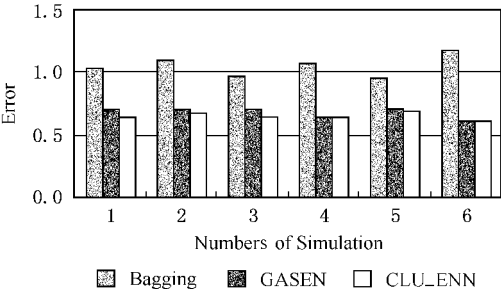


Fig. 2 Errors with different ensemble methods for Freidman # 1 data set.

图 2 不同集成方法在 Freidman # 1 数据集上的误差

数据集 Mexican Hat 只有一个输入分量 ,数据集包含 480 个实例 ,训练集大小为 80 个 ,以其中的 40 个实例作为验证集 ,测试集大小为 400 个 . 运行结果如图 3 所示 :

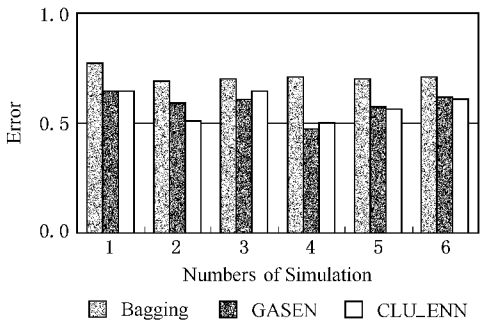


Fig. 3 Errors with different ensemble methods for Mexican Hat data set.

图 3 不同集成方法在 Mexican Hat 数据集上的误差

Boston hosing 数据集具有 13 个特征 ,共有 506 个实例 . 实验中选择训练集大小为 200 ,以其中的 100 个实例作为验证集 ,剩余的实例作为测试集 (206 个) ,在实验中 ,分别选择簇数为 6 ,15 ,25 ,30 与 35 进行了模型集成 ,得到的平方和误差(均值)如表 1 所示 :

Table 1 Relation Between Errors and Number of Clusters for Boston Housing Data Set

表 1 使用 Boston housing 数据集误差与簇数间的关系

Number of Clusters	Error
6	0.167
15	0.168
25	0.164
30	0.171
35	0.169

若使用单个模型测试 ,则由所有单个模型产生的平方和误差均值为 0.245 ;使用全部模型集成得到的平方和误差均值为 0.165 .

实验结果表明 ,使用基于聚类技术的选择性集成方法与基于遗传算法的选择性集成方法一样 ,都能够取得比传统方法更好的结果 . 另外 ,由表 1 知道 ,聚类时选择不同的簇数由集成得到的平方和误差的均值是不同的 ,其中 ,在我们的实验研究中 ,选择簇数 25 较为合适 . 当然 ,由于聚类有效性是一个公开的问题 ,很多学者都对其产生了很大兴趣 ,并提出了很多聚类有效性准则^[18~20] ,但是他们的研究也是各抒己见 ,也没有一个统一的标准确定聚类数

据的簇数 ,因此如何确定模型聚类的簇数是需要进一步探索的 .

5 结 论

本文提出了一种基于聚类技术的选择性集成方法 ,在独立地训练出一批神经网络之后 ,用聚类算法来选择部分神经网络 ,对它们组成神经网络集成 . 实验结果表明 ,这种构造神经网络集成的方法能够有效地降低神经网络集成的泛化误差 . 但也应注意到 ,由于本文主要考虑如何提高集成的泛化能力 ,未考虑聚类算法所用时间 ,所以我们选择了通用的硬 C 均值聚类算法对模型聚类 ,并取得了较为满意的效果 . 应用聚类技术对集成中的模型进行选择 ,是我们的一次尝试 ,在这个问题中还有很多工作要做 ,因为对数据的聚类涉及到聚类有效性等问题 ,因此 ,如何确定聚类数 ,选择什么样的聚类算法对模型聚类更省时间并得到更好的结果 ,是我们需要进一步研究的问题 .

参 考 文 献

1 L. K. Hansen , P. Salamon . Neural network ensembles . IEEE Trans. Pattern Analysis and Machine Intelligence , 1990 , 12 (10) : 993 ~ 1001

2 P. Sollich , A. Krogh . Learning with ensembles : How over-fitting can be useful . In : D. Touretzky , M. Mozer , M. Hasselmo , eds . Advances in Neural Information Processing Systems , Vol 8 . Cambridge , MA : MIT Press , 1996 . 190 ~ 196

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

4 Y. Freund , R. Schapire . Experiments with a new boosting algorithm . In : Proc . the 13th Int ' l Conf . Machine Learning . Bari , Italy : Morgan Kaufmann , 1996

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

6 T. Dietterich , G. Bakin . Solving multiclass learning problems via error-correcting output codes . Journal of AI Research , 1995 , 2 , 263 ~ 286

7 N. C. Oza , K. Tumer . Dimensionality reduction through classifier ensembles . NASA Ames Research Center , Tech. Rep . : NASA-ARC-IC-1999-126 , 1999

8 N. C. Oza , K. Tumer . Input decimation ensembles : Decorrelation through dimensionality reduction . In : J. Kittler , F. Roli , eds . Multiple Classifier Systems . Second International

- Workshop (MCS 2001), LNCS 2096. Berlin: Springer, 2001. 238~247
- 9 Z. Zheng, G. Webb. Integrating boosting and stochastic attribute selection committees for future improving the performance of decision tree learning. The 10th IEEE ICTAI, Los Alamitos, 1998
 - 10 Wu Jianxin, Zhou Zhihua, Shen Xuehua, *et al.* A selective constructing approach to neural network ensemble. Journal of Computer Research and Development, 2000, 37(9): 1039~1044 (in Chinese)
(吴建鑫, 周志华, 沈学华, 等. 一种选择性神经网络集成构造方法. 计算机研究与发展, 2000, 37(9): 1039~1044)
 - 11 Zhou Zhihua, Chen Shifu. Neural network ensemble. Chinese Journal of Computers, 2002, 25(1): 1~8 (in Chinese)
(周志华, 陈世福. 神经网络集成. 计算机学报, 2002, 25(1): 1~8)
 - 12 Z. H. Zhou, J. X. Wu, W. Tang. Ensembling neural networks: Many could be better than all. Artificial Intelligence, 2002, 137(1-2): 239~263
 - 13 K. Al-Ghoneim, B. V. K. Vijaya Kumar. Learning ranks with neural networks. Applications and Science of Artificial Neural Networks, SPIE 1995, Chicago, USA, 1995
 - 14 G. Rogova. Combining the results of several neural network classifiers. Neural Networks, 1994, 7(5): 777~781
 - 15 A. C. Singer, M. Feder. Universal linear prediction by model order weighting. IEEE Trans. Signal Processing, 1999, 47(10): 2685~2699
 - 16 B. Bakker, T. Heskes. Model clustering by deterministic annealing. In: Proc. ESANN. Bruges: D-Facto Public, 1999. 87~92
 - 17 C. L. Blake, C. J. Merz. UCI repository of machine learning databases. <http://www.ics.uci.edu/~mllearn/MLRepository.html>, 1998
 - 18 X. L. Xie, G. Beni. A validity measure for fuzzy clustering. IEEE Trans. Pattern Analysis and Machine Intelligence, 1991, 13(8): 841~847
 - 19 N. R. Pal, J. C. Bezdek. On cluster validity for the fuzzy c-means model. IEEE Trans. Fuzzy Systems, 1995, 3(3): 370~378
 - 20 J. Yu, Q. S. Cheng. The upper bound of the optimal number of clusters in fuzzy clustering. Science in China (series F), 2001, 44(2): 119~125



Li Kai, born in 1963. Ph. D., candidate of the School of Computer Science & Information Technology. His research fields include neural networks, clustering, and data mining, etc. He has published several papers on neural network, clustering, ensemble learning, and so on.

李凯, 1963年生, 博士研究生, 主要研究方向为神经网络、聚类及数据挖掘等, 发表了一些有关神经网络、聚类及集成学习的论文。



Huang Houkuan, born in 1940. Professor and Ph. D. supervisor. His research fields include artificial intelligence, machine learning, data warehousing, data mining, decision support system, and multi-agent system, etc. with more than 100 journal and

conference publications.

黄厚宽, 1940年生, 教授, 博士生导师, 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。在期刊及会议上共发表了100多篇论文。

Research Background

This work is supported by the National Natural Science Foundation of P. R. China (Grant No. 60443003). Ensemble learning has become one of research fields of machine learning, it dramatically improves generalization performance of learning machines. Recently, neural network ensemble techniques have gained widespread use among neural network practitioners. There are many different varieties, but the most popular include some elaboration of bagging, boosting or stacking. In a regression problem, Krogh and Vedelsby showed that as long as the average of the individual network generalization errors in an ensemble remains constant, increasing diversity will improve generalization performance. It also confirms intuition—an ensemble that consists of a thousand identical networks will not perform any better than an individual network. Although they have established the importance of generating diversity in neural network ensembles, the question remains: how is diversity created so that ensemble generalization performance is improved? In this case, we attempt to study an approach to improving diversity of neural networks, that is to use clustering technology. After individual neural networks are trained, the clustering algorithm is used to select a part of the trained individual networks in order to increase their diversities. Then many selected neural networks are combined. Experimental results show that this approach outperforms the traditional ones that ensemble all of the individual networks. We believe that this method is very promising and is suitable to engineering design of neural network ensembles.