基于最近邻规则的神经网络训练样本选择方法

郝红卫 [1] 蒋蓉蓉 [1]

摘 要 训练集中通常含有大量相似的样本, 会增加网络的训练时间并影响学习效果. 针对这一问题, 本文将最近邻法 (Nearest neighbor, NN) 简单快捷和神经网络高精度的特点相结合, 提出了一种基于最近邻规则的神经网络训练样本选择方法. 该方法考虑到训练样本对于神经网络性能的重要影响, 利用改进的最近邻规则选择最具有代表性的样本作为神经网络的训练集. 实验结果表明, 所提出的方法能够有效去除训练集中的冗余信息, 以少量的样本获得更高的识别率, 减少网络的训练时间, 增强网络的泛化能力.

关键词 神经网络, 样本选择, 最近邻规则, 手写字符识别
中图分类号 TP391.41

Training Sample Selection Method for Neural Networks Based on Nearest Neighbor Rule

HAO Hong-Wei[1] JIANG Rong-Rong[1]

Abstract Training sets usually contain large amount of similar samples, resulting in a longer training time and poor performance. To deal with this problem, a training sample selection method for neural networks based on nearest neighbor (NN) rule was proposed. Considering the significance of train sets for the performance of neural networks, the proposed method combined simplicity of nearest neighbor (NN) with high accuracy of neural networks and utilized the modified NN rule to select the most representative samples as a new training set. Experimental results show that the presented method can eliminate the redundancy, achieve higher recognition accuracy and better generalization ability with fewer samples and less training time.

Key words Neural network, sample selection, NN rule, handwritten character recognition

1 引言

人工神经网络的研究与计算机的研究几乎是同步发展的. 现在神经网络的应用已经渗透到许多领域中, 在模式识别、智能控制、信号处理、系统辨识等领域得到了广泛的应用[1, 2]. 目前已有多种模型和学习算法, 仅以前馈网络为例, 应用较多的有反向传播 (Backpropagation, BP) 网络、径向基网络 (Radial basis function, RBF) 等. 其中, 基于误差反向传播算法的多层感知机 (Multi-layer perceptron, MLP) 网络是研究最为深入、应用最为广泛的模型. BP 网络因其结构简单、容易实现而得到普及. 标准 BP 算法是一种监督学习方式, 它使用梯度下降法, 以期最小化网络的实际输出与期望输出的均方差. 后来在标准 MLP 网络和 BP 算法的基础上针对各种具体问题提出了许多改进措施, 极大地改善了网络的性能. 归纳起来, 各种改进措施主要集中在以下三个方面: 1) 对网络结构的优化, 如权值修剪 (Weightprune)、权值衰减 (Weight decay) 和利用遗传算法 (Generic algorithm, GA) 确定网络结构[3, 4] 等; 2) 对标准 BP 算法的改进, 如增加动量项, 改变学习速率, 利用验证集确定适当的训练深度等; 3) 对训练样本的处理, 如增加 "人工制造" 的数据、加噪声的训练方法和样本的重采样技术 (Bagging, boosting, adaboost) 等. 其中, 在对训练样本的处理中, 由于过去样本数量有限, 所以各种处理方法往往以增加样本为主. 对于如何通过减少样本数量来改善网络性能, 则很少涉及.

样本在神经网络的学习中占有非常重要的地位, 其中蕴含的信息直接影响着网络的性能. 样本集是否具有代表性, 决定了网络的学习效果. 样本的收集是一项极其繁琐的工作, 大规模样本库的建立是十分困难的. 这就导致了在早期的研究中, 样本数量往往不足. 因此对训练样本的处理, 一般是通过人工制造数据和产生噪声等方法添加样本, 使网络具有足够大的训练集. 而对于样本数量较多的情况则关注不多, 研究很少. 随着一系列大规模样本库的建立, 训练集中样本的冗余问题逐渐显现出来. 例如国际标准手写数字样本库 MNIST 训练集包含 60 000 个样本, 由近 250 人书写而成, 每个人都写了多组样

收稿日期 2006-9-1 收修改稿日期 2007-5-9
Received September 1, 2006; in revised form May 9, 2007
国家自然科学基金 (60675006) 资助
Supported by National Natural Science Foundation of China (60675006)
1. 北京科技大学信息工程学院 北京 100083
1. School of Information Engineering, University of Science and Technology Beijing, Beijing 100083
DOI: 10.1360/aas-007-1247

本; 中国科学院自动化研究所收集的自由手写体数字样本库含有 1 100 000 个样本, 是由 22 000 人书写而成, 每人书写了 5 组数字. 因为每个人的书写习惯相对稳定, 使得这些样本库中包含着许多相近甚至相同的样本, 出现了大量的冗余信息. 以全部的样本进行训练不仅需要花费更多的时间, 使网络训练速度减慢, 而且这些冗余信息可能会造成样本空间不平衡, 产生对某些样本的 "过学习" 现象, 降低网络的泛化能力.

为了消除样本中的冗余信息, 我们可以从原始样本库中挑选部分典型样本, 形成新的训练集. 目前, 在这方面的研究还很少, 已有的一些样本选择方法, 如随机选择、动态选择等均有各自的不足. 随机选择简单易行, 可以减少训练时间, 但是由于样本的选择是随机的, 得到的结果代表性差, 识别率随着所选样本数量的增加而逐渐增长, 基本起不到样本选择的作用, 在改进网络性能上也几乎没有效果[5]. 动态选择是按照某种规则在训练过程中根据网络的训练情况, 决定将哪些样本添加到训练集中[6], 运算量极大, 很难用于大型样本库的处理[7]. 为了快速高效地选择具有代表性的训练样本, 我们提出了一种新的训练样本选择方法, 该方法将最近邻法简单快捷和神经网络高精度的特点相结合, 利用改进的最近邻规则从初始训练集中选择最具有代表性的样本作为神经网络的训练集. 实验结果表明, 本文提出的方法能够以少量的样本获得更高的识别率, 增强网络的泛化能力. 论文安排如下: 第 2 节介绍本文提出的样本选择方法; 第 3 节给出了在两个国际通用手写体数字样本库 MNIST 和 USPS 上的实验结果; 最后是结论.

2 基于 NN 规则的样本选择方法

我们将以 BP 网络为例来展开研究, 结果可以适用于任意的前馈网络.

样本选择的关键在于如何判断哪些样本是冗余的. 为此可以建立如下判定规则: 设有训练集 T, 其子集为 S, S 对于 T 的补集为 C, 由 S 训练得到的分类器为 X, 若 X 能正确识别 C 中的样本, 则 C 为冗余.

样本选择的任务就是由 T 得到 S, $S \subseteq T$. 最直观的方法是使用 BP 算法训练分类器 X, 由于 BP 算法和样本选择都需要反复迭代, 这样必然导致选择算法的时间开销过大. 考虑到最近邻规则与 BP 网络决策的相似性, 分类器 X 可以采用最近邻模型. 事实上, BP 网络也是一种最近邻分类器, 只不过它的模板以权值的形式存储在网络结构中, 是一种优化的近邻方法. 相比之下, 近邻法直接以代表性样本作为模板, 没有反复迭代调整的过程. 显然, 近邻法简单快捷但精度较低, BP 网络训练复杂但精度较高. 以近邻法构造分类器 X 来选择样本, 能够发挥其简单快速的优点, 而用由此生成的子集 S 来训练神经网络得到的分类器, 则依然保持了神经网络高精度的特点.

Hart 提出了压缩近邻 (Condensed nearest neighbor, CNN) 算法[8], 其算法如下所示.

```
Algorithm CNN
Input: training set T ;
Output: reduced subset D ;
begin
  pick the first sample x_1 from T ;
  D = x_1 ;
  T = T − x_1 ;
  repeat
      append = FALSE ;
      for all patterns in T pick x from T,
          Find s in D such that Distance (x, s) =
          min_{s_j ∈ D} Distance (x, s_j);
          if Class (x) ≠ Class (s) then
              D = D ∪ x ;
              T = T − x ;
              append = TRUE ;
          end if
      end for
  until append = FALSE ;
  return D ;
end
```

CNN 算法利用原有样本集 T, 逐渐生成一个新的样本子集 D, 使 D 在减少了样本数量的条件下, 仍能对 T 中的每个样本用最近邻法正确分类. 当 T 中的某个样本不能被 D 正确分类时, 就被加入到 D 中, 直到某次循环完成时 D 不发生任何变化为止.

CNN 算法十分直观, 按照最近邻规则, 靠近各类边缘的样本对分类有较大的作用, 而聚类中心附近的样本则对分类作用很小, CNN 试图删除聚类中心附近的样本, 达到降低存储、缩短运算时间的目的. 但事实上, 按照上述算法并不能完全达到这个目的. 原因在于, 其样本的选择是依次进行的, 顺序靠前的样本比靠后的样本被保留下来的可能性要大得多, 这样势必会造成部分边缘样本被删除而某些中心样本却被保留下来, 使得选择结果不仅代表性较差而且仍会有冗余. 我们的实验结果也充分表明了这一点. 另外, 通过 CNN 算法选择的样本数量是固定的, 缺乏灵活性, 难以实现对样本数量的控制.

针对上述问题, 我们通过以下两种措施对其进行改进. 对于某些边缘样本未被选中的问题, 可以将 CNN 过程重复数次, 以确保边缘样本均被选中; 对于中心样本仍被保留的问题, 通过对样本加权评估[9]、再次选择的方式来解决. 同时, 在此选择的过

程中可以根据需要灵活控制样本数量. 我们称之为加权 CNN 方法 (Weighted CNN, WCNN), 其算法如下所示.

```
Algorithm WCNN
Input: training set T ;
Output:
reduced subset S with predetermined number m ;
begin
  initialize A = ∅, n = iterations, k = 0, V, j = 0, S = ∅ ;
  do k = k + 1
      D = CNN(T) ;
      A = A ∪ D ;
      T = T − D ;
  until k = n ;
  for each pattern x in T do
      find x_i in A such that Distance (x, x_i) =
          min_{x_i ∈ A} Distance (x, x_i) ;
      if Class (x) =Class (x_i) then
          v_i + + ;
      end if
  end for
  sort V in descending order ;
  do j = j + 1
      select x_i from A according to V ;
      S = S ∪ x_i ;
  until j = m ;
end
```

上述算法通过 CNN 过程的循环确保了子集 A 中包含了足够多的边界样本, 但是其中仍存在大量冗余, 我们采用投票原则对 A 中样本的代表性进行评估并再次选择. 具体方法是对于 T 中的每个样本 x 找出 A 中与它距离最近的样本 x_i, 如果 x 和 x_i 类别相同, 那么 x_i 就获得一票. 一个样本的得票数越多, 说明它的代表性越高. 根据得票情况和所需样本数量得到最终子集 S. 由此得到的子集 S 比由 CNN 选择的子集 D 包含更多的代表性样本和更少的冗余, 同时还可以灵活控制子集中所包含的样本数量.

3　实验结果

手写数字识别是一个经典的模式识别问题. 由于类别数小, 使得一些复杂的或运算量较大的算法实现起来比较容易. 因此, 数字识别始终是模式识别中各种新方法研究的实验对象, 在算法研究中占据着重要的地位. 此外, 10 个阿拉伯数字是全世界唯一一套通用的字符集, 是各国学者共同关注的研究热点[10, 11]. 为了方便人们的研究与交流, 国际上一些组织建立了通用的手写体数字样本库, 其中典型的有 MNIST 库和 USPS 库, 很多识别方法都以这两个样本库作为评测的标准[12]. 本文同样使用了这两个数据库进行实验.

1) MNIST 样本库实验结果

MNIST 由训练集和测试集两部分组成, 分别含有 60 000 个训练样本和 10 000 个测试样本, 每个样本用 28×28 的向量表示[13], 图 1 显示了训练集中的部分样本图像.

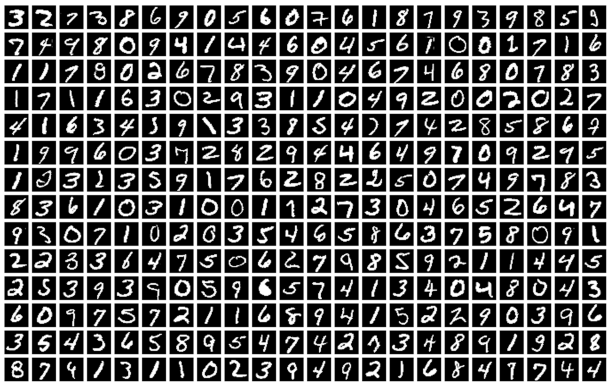


图 1　MNIST 库的部分样本图像

Fig. 1　Images of samples in MNIST database

实验中采用三层结构的 BP 网络, 为了便于分析和比较, 网络的基本参数保持不变, 具体步骤如下:

a) 对样本进行预处理后, 提取方向特征[14], 每个样本得到 60 维的特征;

b) 用 WCNN 算法从 60 000 个训练样本中选择固定数量的样本子集 S;

c) 以样本子集 S 作为输入, 训练 BP 网络, 得到权值;

d) 对 10 000 个测试样本进行识别.

逐步增加样本子集中的样本数量, 重复 b) ∼ d), 实验结果如表 1 所示, 其中第 1 列数据为采用 CNN 算法得到的结果, 其余各列为采用 WCNN 得到的结果. 显然, CNN 算法得到的样本数目固定, 且样本压缩量大并且含有冗余, 因而识别率低, 表明其代表性较差. 采用 WCNN 算法可以控制样本的数量, 表中给出了随着样本数目的增加, 对测试样本识别率的变化情况. 当样本数目增加到 27 000 即全部样本量的 45% 时, 识别率达到最高点; 样本数量在 24 000 ∼ 36 000 时, 识别率均高于用全部样本训练的测试结果. 图 2 显示了上述变化曲线.

2) USPS 样本库实验结果

USPS 含有 7 291 个训练样本和 2 007 个测试样本, 每个样本用 $16 \times 16 = 256$ 维的向量表示, 向量取值范围在 0 ∼ 2 之间[15], 图 3 显示了该字库的部分样本图像. 在 USPS 上的实验过程与 MNIST 相似, 实验结果如表 2 所示.

表 1 MNIST 库实验结果

Table 1 The test results on MNIST database

样本选择方法	CNN	WCNN							
样本数目	12 286	18 000	24 000	27 000	30 000	36 000	42 000	48 000	54 000	60 000
样本压缩比 (%)	20.48	30	40	45	50	60	70	80	90	100
识别正确率 (%)	94.67	95.42	95.55	**95.71**	95.63	95.60	95.11	95.25	95.2	95.43

表 2 USPS 库实验结果

Table 2 The test results on USPS database

样本选择方法	CNN	WCNN								
样本数目	1 598	2 400	3 000	3 600	4 200	4 800	5 400	5 600	6 000	6 600	7 291
识别正确率 (%)	88.34	90.18	90.63	90.88	91.18	91.33	91.43	**91.62**	91.33	91.43	91.28

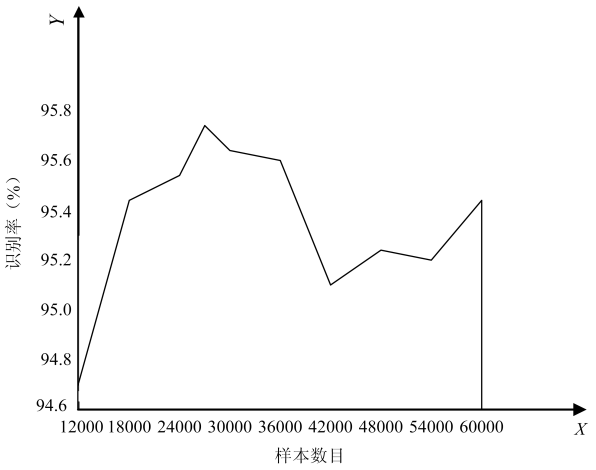


图 2 MNIST 库实验结果

Fig. 2 The test results of MNIST database

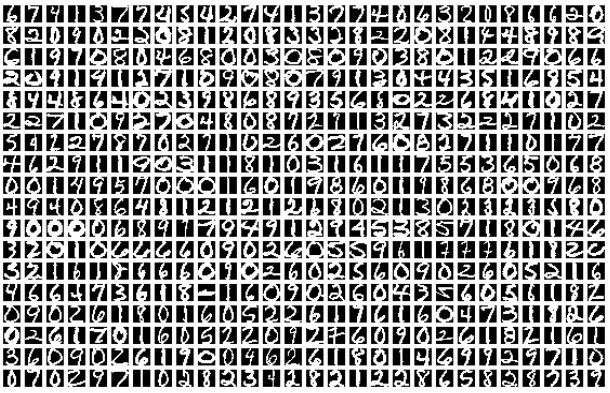


图 3 USPS 库的部分样本图像

Fig. 3 Images of samples in USPS database

同样, CNN 算法选择的样本数量少, 效果差. 随着样本数量的增加, 识别率也相应地变化, 样本量达到 4 800 后, 识别率超过了用全部样本训练的结果; 当样本数量为 5 600 即约为全部样本的 77% 时, 识别率达到最高.

在 MNIST 和 USPS 上的测试结果表明, 利用改进的近邻规则对网络的训练集进行选择, 实现了以少量样本得到更高识别率的目的. 以近邻法构造分类器来选择样本, 简单快速, 并且保留下来的样本具有典型性, 能够近似表达全部样本蕴含的信息; 而用新的样本子集来训练神经网络得到的分类器, 依然保持较高的精度. 由于训练集中的冗余样本被去除, 降低了对某些样本 "过学习" 的可能性, 从而增加了网络的泛化能力. 同时, 训练集中的样本数量减少, 缩短了网络的训练时间. 当样本库规模巨大时, 可以在此基础上采用快速最近邻方法来进一步提高算法的效率.

对比两个实验结果中最高识别率所对应的样本数量, 我们发现样本的压缩比与训练集的大小有密切的联系, 训练集越大, 则被去除的样本比例越高. 可见, 随着样本库规模的增大, 其中含有的冗余信息增多, 采用 WCNN 算法就能发挥更大的作用, 不仅可以减少训练样本、降低存储空间、极大地提高训练速度, 而且还可以大大提高识别率. 分析在 MNIST 上的样本压缩比与识别率之间的关系, 我们发现当压缩比处于两个重要的 Fibonacci 数 38.2% 和 61.8% 之间时, 对应的识别率均超过使用全部训练样本所得的识别率. USPS 样本量很少, 压缩比过高会导致训练样本数量不足, 因而不满足这一规律. 研究最佳压缩比的确定方法, 是我们下一步的工作.

4 结论

本文考虑到训练样本对于神经网络性能的重要影响, 结合最近邻法简单快捷和神经网络识别精度高的特点, 提出了基于 NN 原则的样本选择方法. 在国际通用的两个手写体数字样本库 MNIST 和 USPS 上的实验结果表明该方法可以有效去除训练集中的冗余, 使保留下的样本更好地表达全部样本蕴含的信息. 用新的样本子集训练神经网络, 能达到

以较少样本得到更高的识别率的目的. 特别是在大型训练样本库的处理中采用这种方法, 可以显著加快网络的训练速度, 节省存储空间, 并且能够提高识别率, 增加网络的泛化能力. 神经网络以待识别样本与训练样本的相似性作为决策依据, 而相似度则是通过对训练样本的学习获得的, 因此, 我们提出的样本选择方法对于提高网络性能具有重要意义. 该方法不仅适用于 BP 网络, 也可应用于其他类型的神经网络模型.

References

1 Duda R O, Hart P E, Stork D G. *Pattern Classiffcation (Second Edition)*. Beijing: China Machine Press, 2004

2 Hao H W, Xiao X H, Dai R W. Handwritten Chinese character recognition by metasynthetic approach. *Pattern Recognition*, 1997, **30**(8): 1321∼1328

3 Hao H W, Liu C L, Sako H. Comparison of genetic algorithm and sequential search methods for classifler subset selection. In: Proceedings of the 7th International Conference on Document Analysis and Recognition. Edinburgh, Scotland: IEEE, 2003. 765∼769

4 Liang Hua-Lou, Dai Gui-Liang. Combination of genetic algorithm and artificial neural networks: review and prospect. *Acta Electronica Sinica*, 1995, **23**(10): 194∼200
(梁化楼, 戴贵亮. 人工神经网络与遗传算法的结合进展及展望. 电子学报, 1995, **23**(10): 194∼200)

5 Baum E, Haussler D. What size nets give valid generalization? *Neural Computation*, 1989, **1**(1): 151∼160

6 Plutowski M, White H. Selecting concise training sets from clean data. *IEEE Transactions on Neural Network*, 1993, **4**(2): 305∼318

7 Robel A. The Dynamic Pattern Selection Algorithm: Effective Training and Controlled Generalization of Backpropagation Neural Networks. Technical Report 93-23, Techniche University Berlin, Germany: 1993

8 Hart P E. The condensed nearest neighbour rule. *IEEE Transactions on Information Theory*, 1968, **14**(3): 515∼516

9 Dasarathy B V. Minimal consistent set (MCS) indentification for optimal nearest neighbor decision systems design. *IEEE Transactions on Systems, Man, and Cybernetics*, 1994, **24**(3): 511∼517

10 Yang L C, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998, **86**(11): 2278∼2324

11 Hao H W, Liu C L, Sako H. Confidence evaluation for combining diverse classifiers. In: Proceedings of the 7th International Conference on Document Analysis and Recognition. Edinburgh, Scotland: 2003. 760∼764

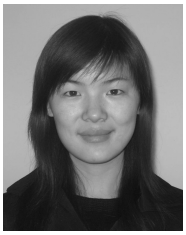
12 Decoste D, Scholkopf B. Training invariant support vector machines. *Machine Learning Journal*, 2002, **46**(13): 161∼190

13 Yang L C, Cortes C. The mnist database of handwritten digits [Online], available: http://yann.lecun.com/exdb/ mnist, October 10, 2004

14 Hao H W, Dai R W. An integration approach to handwritten Chinese character recognition system. *Science in China (Series E)*, 1998, **41**(1): 101∼105

15 Inmage Processing Research Laboratory in Hefei University of Technology [Online], available: http://wwwi1.hfut.edu.cn/organ/images/imagelab/download/usps.htm, November 19, 2007

郝红卫 北京科技大学教授, 主要研究方向为图像处理和模式识别. 本文通信作者. E-mail: hhw@ies.ustb.edu.cn
(**HAO Hong-Wei** Professor at School of Information Engineering, University of Science and Technology Beijing. His research interest covers image processing and pattern recognition. Corresponding author of this paper.)



蒋蓉蓉 硕士研究生, 主要研究方向为图像处理和模式识别.
E-mail: rongrongj@126.com
(**JIANG Rong-Rong** Master student at University of Science and Technology Beijing. Her research interest covers image processing and pattern recognition.)