基于广义径向基函数的神经网络分类预测

张 莉,姜 浩,蒲安建
(东南大学 计算机科学与工程学院,江苏 南京 211189)

摘 要:径向基函数网络是神经网络中一种广泛使用的设计方法。它把神经网络的设计看作是一个高维空间的曲线逼近问题。相对于其他的神经网络方法,径向基函数神经网络除了具有一般神经网络的优点,如多维非线性映射能力、泛化能力、并行信息处理能力等,还具有很强的聚类分析能力,学习算法简单方便等优点。针对一个实际分类问题,利用广义径向基函数网络的思想训练一个网络并实现对测试数据集的分类预测。本算法采用k-均值聚类算法训练广义径向基函数网络中心,使用奇异值分解计算输出层权值。对该网络的实现细节及待改进之处进行简要分析。实验表明广义径向基函数神经网络的思想具有很强的聚类分析能力,学习算法简单方便等优点。

关键词:神经网络;径向基函数;分类预测
中图分类号:TP183 **文献标识码**:A **文章编号**:1673-629X(2009)03-0106-04

Classification and Prediction of Neural Network Based on Generalized Radial Basis Function

ZHANG Li, JIANG Hao, PU An-jian
(School of Computer Science & Engineering, Southeast University, Nanjing 211189, China)

Abstract: Radial basis functions (RBF) is a widely used tool in the neural networks. It views the design of the neural networks as a curve approaching problem in the high-dimensional space. Besides the advantages of the general neural networks, such as multi-dimensional nonlinear mapping capabilities, generalization, parallel information processing capabilities, Gaussian radial basis functions has strong ability of cluster analysis and its learning algorithm is simple and convenient. In this paper use Gaussian radial basis functions to solve a classified problem. First train the network and then realize classifying the data in the test set. The algorithm uses k-means clustering algorithm to generalize the training centers of the RBF network and singular value decomposition algorithm to calculate the output values. Finally, the analysis of the algorithm is given. The experiment shows that Gaussian radial basis functions has strong ability of cluster analysis and its learning algorithm is simple and convenient.

Key words: neural networks; radial basis functions; classification and prediction

0 引 言

人工神经网络是一个由简单处理单元构成的规模宏大的并行分布式处理[1]。这种网络通过学习实现有用的计算。人工神经网络广泛应用在智能控制、模式分类、优化计算等方面。神经网络中用于完成学习过程的程序称为学习算法。其功能是以有序的方式改变网络的属性以获得想要的设计目标。径向基函数网络是神经网络中一种广泛使用的设计方法。它把神经网络的设计看作是一个高维空间的曲线逼近问题[1],它在模式分类和识别中有着重要的地位。文中将利用广义径向基函数网络对一个实际问题进行分类预测。

收稿日期:2008-06-15
作者简介:张 莉(1986-),女,江苏连云港人,硕士研究生,研究方向为数据库;姜 浩,副教授,研究方向为工作流应用研究。

1 实际问题

1.1 问题描述

目前,多数人使用的是第二代(2G)移动通信网络,而第三代(3G)移动通信网络已投入使用。一个成功发行3G移动通信网络的电信公司希望根据已存在的用户的一些个人信息,来预测用户使用3G移动通信网络的可能性。

1.2 问题分析

这是一类分类问题,即通过已有数据预测未来的数据趋势。已知用户的若干种属性(250种),希望通过此来把用户分类:2G用户类和3G用户类。

1.3 初步方法构想

解决分类问题必须构造分类器,即可以用于提取描述重要数据类或预测未来的数据趋势的模型。对于此问题,将用广义径向基函数的原理构造分类模型,根

据训练集(18000 个样本)训练一个广义径向基函数网络,并用训练好的网络对测试数据集(6000 个样本)进行分类。

为了提高分类过程的准确性、有效性和伸缩性,在使用训练集训练网络之前,应该对训练集数据进行预处理工作[2],比如数据清理,数据变换和规约等,然后用预处理过的数据训练网络。

2 背景知识

2.1 径向基函数神经网络

径向基函数神经网络是一种前馈神经网络。相对于其他的神经网络方法,径向基函数神经网络除了具有一般神经网络的优点,如多维非线性映射能力、泛化能力、并行信息处理能力等,还具有很强的聚类分析能力,学习算法简单方便等优点。径向基函数网络的出发点就是把神经网络的设计看成是一个高维空间的数据拟合问题[1]。从而学习过程等价于在多维空间寻找一个能够拟合训练数据的曲面。泛化等价于利用这个多维曲面对测试数据进行插值。在神经网络的背景下,隐藏单元提供一个"函数"集,该函数集在输入描述扩展至隐藏空间时构建一个任意的"基",这个函数集中的函数就被称为径向基函数。

基本形式的径向基函数网络包括三层:输入层,隐藏层以及输出层[3],如图 1 所示。每层有着不同的作用。输入层由一些源点组成,它们将网络与外界环境连接起来。第二层是网络中的一个隐藏层,它的作用是从输入空间到隐藏空间之间进行非线性变换;大多数情况下,隐藏空间有较高的维数。输出层是线性的,它为作用于输入层的信号提供响应。其中隐藏空间的维数较高是因为一个模式分类问题如果映射到一个高维空间将会比映射到低维空间更有可能是线性可分[1]。

径向基函数是要选择一个函数 F 有以下形式:

$$F(\boldsymbol{x})=\sum_{i=1}^{n} w_i\varphi(\|\boldsymbol{x}-\boldsymbol{x}_i\|) \tag{1}$$

其中,$\|\ \|$ 表示范数,通常是欧几里德范数。w_i 是输出层权值。φ 可取 Gauss 函数 $\varphi(r)=\exp(-r^2)$。已知数据 $\boldsymbol{x}_i\in R^m$,$i=1,2,\cdots,n$ 是径向基函数的中心。其中径向基函数中心个数 n 一般为训练数据集的大小[1]。

训练一个径向基函数网络模型就是利用训练样本集 $\{(\boldsymbol{x}_i,d_i)\}_{i=1}^{n}$,构造输出层权值 w_i 和径向基函数的中心。

由文献[1]知训练公式为:$\boldsymbol{w}=(\boldsymbol{G}+\lambda I)^{-1}\boldsymbol{d}$　(2)

其中,$\boldsymbol{w}=(w_1,w_2,\cdots,w_n)$,$\boldsymbol{G}$ 为矩阵 $[\boldsymbol{G}(\boldsymbol{x}_i,\boldsymbol{x}_j)]_{i,j}$,$\lambda$ 为正则化参数,取值在 0 到 1 之间。$\boldsymbol{d}$ 是训练集中期望响应向量。

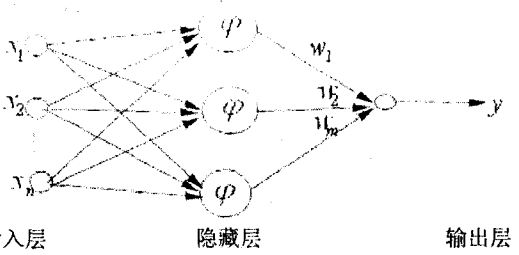

图 1　径向基函数神经网络

2.2 广义径向基函数神经网络

广义径向基函数神经网络和普通径向基函数神经网络一样具有输入层、隐藏层和输出层三层网络结构。不同点在于广义径向基函数神经网络中心的个数远小于训练集的大小 n[4]。这是因为当 n 很大时,要计算一个 $n\times n$ 阶矩阵的逆,其计算量按 n 的多项式增长(大约为 n^3)。为了克服这些计算上的困难,降低神经网络的复杂度,不妨在一个较低维的空间中求一个次优解,以此来逼近式(1)中的最优解。其径向基函数具有以下的形式:

$$F(\boldsymbol{x})=\sum_{i=1}^{m} w_i\varphi(\|\boldsymbol{x}-\boldsymbol{x}_i\|) \tag{3}$$

其中 m 为隐藏层节点数(中心数),$m<n$。文献[1]中给出一训练公式:

$$\boldsymbol{w}=\boldsymbol{G}^{+}\boldsymbol{d} \tag{4}$$

其中 $\boldsymbol{G}^{+}$ 是 $\boldsymbol{G}$ 的伪逆,即 $\boldsymbol{G}^{+}=(\boldsymbol{G}^{\mathrm{T}}\boldsymbol{G})^{-1}\boldsymbol{G}^{\mathrm{T}}$　(5)

2.3 Weka 机器学习平台

Weka 机器学习平台是由新西兰怀卡托大学开发的工作平台。它汇集了当今最前沿的机器学习算法及数据预处理工具,包含属性选择、回归、分类、聚类以及关联规则挖掘等标准数据挖掘的方法。Weka 的免费使用版下载网址:

http://www.cs.waikato.ac.nz/ml/weka

3 问题处理

3.1 数据预处理

3.1.1 处理缺失值

处理缺失值是使用该属性最常出现的值,或最可能的值替换缺省值。根据每个属性的不同类型来采取不同的处理方法。

对于连续性属性的数据,不妨取所有现存数据的平均值来替换缺失值。如属性 AGE,其平均值可以用以下的 SQL 语句找出:

```
SELECT AVG(AGE)
FROM training
WHERE AGE IS NOT NULL
```

对于离散型属性的数据，则可取该属性现存数据中出现频率最高的值来替换缺失值。如属性 HS_MODEL，其中出现频率最高的值也可以通过以下 SQL 语句得到：

```
SELECT HS_MODEL
FROM training
GROUP BY HS_MODEL
HAVING count( * ) > = all
(SELECT count( * )
FROM training
GROUP BY HS_MODEL)
```

其中 training 是训练数据集 training 对应的数据库表名。

3.1.2 属性的选择

使用 Weka 进行属性选择。打开 Weka 的主要图形用户界面 Explorer(探索者)，导入训练数据集 training 对应的 aref 格式的文件后选择 Select attributes 选项。搜索方法是 GreedStepwise，即不具有返回的贪心登山式的搜索，而属性选择的评估方法用 ConsistencySubsetEval，即将训练数据集映射到属性集上来检查类值的一致性。得到的结果如下：

```
Attribute Selection on all input data
Search Method:
Greedy Stepwise(forwards).
Start set: no attributes
Merit of best subset found:1
Attribute Subset Evaluator(supervised, Class(nominal):251 CUSTOMER_TYPE)
Consistency Subset Evaluator
Selected attributes: 7,11,23,26,28,30,48,111,181,219,244,248:12
COBRAND_CARD_FLAG
SUBPLAN
HS_AGE
HS_MODEL
TOT_RETENTION_CAMP
LOYALTY_POINTS
TOP1_INT_CD
AVG_CALL_T1
STD_MINS_OP
STD_VAS_GAMES
STD_CALL_FRW_RADIO
STD_MM_CALL_RADIO
```

3.1.3 属性选择结果分析

首先算出 training 中元组分类所需的期望信息为：

$$\mathrm{Info}(\mathrm{training}) = -\sum p_i \log_2 p_i = 0.5491$$

接下来对于离散型属性，计算出基于按各属性划分对 training 的元组分类所需的期望信息 $\mathrm{Info}_A(\mathrm{training})$ 和各属性的增益值 $\mathrm{Gain}(A)$[5]。其中：

$$\mathrm{Info}_A(\mathrm{training}) = -\sum \frac{D_j}{D} \times \mathrm{Info}(D_j) \quad (6)$$

$$\mathrm{Gain}(A) = \mathrm{Info}(\mathrm{training}) - \mathrm{Info}_A(\mathrm{training}) \quad (7)$$

对于连续性属性，由于该属性对 training 集的分裂将导致大量的划分，使得用增益衡量属性选择结果不够准确。为了克服这种偏移，采用增益率的办法对其离散化。计算连续性属性的分裂信息 SplitInfoA(training) 和增益率 GainRatio(A)[5]。其中：

$$\mathrm{SplitInfo}_A(\mathrm{training}) = -\sum \frac{|D_j|}{|D|} \times \log_2 \frac{|D_j|}{|D|} \quad (8)$$

$$\mathrm{GainRatio}(A) = \frac{\mathrm{Gain}(A)}{\mathrm{SplitInfo}(A)} \quad (9)$$

选择增益率大的属性作为分裂属性，对其进行二分分裂，仍选取所需取值的平均值作为分裂点，这样该属性就已离散化。然后对离散化后的该属性重新计算增益。

最后结果如下表：

attribute	$\mathrm{Info}_A(\mathrm{training})$	Gain(A)
COBRAND_CARD_FLAG	0.5393	0.0098
SUBPLAN	0.494	0.0551
HS_AGE	0.4482	0.1009
HS_MODEL	0.4098	0.1393
TOT_RETENTION_CAMP	0.5244	0.0247
LOYALTY_POINTS	0.5374	0.0117
TOP1_INT_CD	0.5278	0.0213
AVG_CALL_T1	0.5292	0.0188
STD_MINS_OP	0.5347	0.0144
STD_VAS_GAMES	0.5156	0.0335
STD_CALL_FRW_RATIO	0.5254	0.0237
STD_M2M_CALL_RATIO	0.5255	0.0236

最后就算出上表所选出属性的信息增益的总和为 0.4768，占 training 集总信息量的 86.8%。由此可见，用这些属性对 training 集划分，纯度基本能满足后面分类的要求。

用 access 建立一名为 data 数据库，把 training 集和 test 集中选择出的属性和 SERIAL_NUMBER(用户 ID 作为关键码)以及 CUSTOMER_TYPE(预测属性)分别以名为 training 和 test 的表的形式写入数据库中。

3.1.4 数据规范化

在训练网络之后，应当对训练集合测试集中的数据进行规范化处理[5]。为了节省工作量，把本步骤放在属性选择之后。这样只需对那些被选择出的属性中的数据规范化处理。

对于连续类型的数据，直接做最大最小规范化，把该属性的取值范围变换到[0,1]之间。对于离散型的

数据,先对其不同的取值依次记为0,1,2…,然后再做最大最小规范化处理。对于待预测的属性CUSTOMER_TYPE,当其值为2G记为0,值为3G记为1。

3.2 分　类

3.2.1 训练网络中心

首先,根据训练集样本来训练网络中心。

网络中心集的选取方法有多种。对于径向基函数网络,一般用所有的训练集元组作为网络的中心集。但如上文所讲的那样,当训练样本数很大时,训练该网络包括测试的计算量将非常大。所以采用广义径向基函数网络的思想,选取较小的中心集。然而,从训练集样本中选取部分元组直接作为网络的中心集随意性较强,不利于网络的泛化性能。所以采用传统的k-均值聚类算法[5]。用该算法对训练集样本(即网络的输入数据集)进行聚类,用所得到的各聚类中心作为广义径向基函数网络的中心集。在执行算法前,不妨设中心数为 m,而 m 取何值依靠后面实验来决定。

3.2.2 中心个数 m 的确定

把训练集分成两部分,一部分17000个样本,用作训练数据训练网络;另一部分1000个样本用来测试网络的泛化性能。每次训练时选择不同的 m(保证 m 输入向量的维数),选取使网络泛化性能最佳的 m。经验证,$m = 90$ 时,对于此问题网络的泛化性能较为理想。

3.3 得出结果

最后,使用测试集test中的6000元组作为训练好的网络的输入数据。注意,该网络是用18000个样本训练的,而不只是3.2.2中的17000个样本。执行算法,得到各测试对象的输出值,并按其值从大到小的顺序(即从最有可能使用3G通信网络的用户到最没有可能使用3G通信网络的用户)把测试集对象的ID写入output.txt文件中。

4 算法分析

4.1 关于中心数 m 的选取

中心数 m 的选取对于整个网络的选取是较为关键的。本算法采用3.2.2中所述的方法,试探性地取若干个值代入算法中。根据网络泛化的性能来决定 m 的值。这种方法具有一定的任意性,是有缺陷的,有待于进一步的改进。

文献[6]中使用EM算法确定网络中心的个数,而文献[7]中给出计算中心个数的经验公式,也是一个不错的方法。但是这两个方法到底能否提高网络泛化性能,有待于进一步验证。

4.2 关于伪逆矩阵的算法

由于求伪逆矩阵涉及到矩阵求逆,当样本集较大的时候,若用Gauss消元法等常见方法,计算非常复杂。一个比较好的处理方法是用奇异值分解。

因为 $\boldsymbol{G}$ 是 $n \times m$ 实对称矩阵,故存在 n 阶正交矩阵 $\boldsymbol{U}$,m 阶正交矩阵 $\boldsymbol{V}$,使得:

$$\boldsymbol{U}^{\mathrm{T}}\boldsymbol{G}\boldsymbol{V} = \mathrm{diag}(\sigma_1, \sigma_2, \cdots, \sigma_k) \tag{10}$$

其中 $k = \min(n, m)$,$\sigma_1 \geqslant \sigma_2 \geqslant \cdots \geqslant \sigma_k > 0$[8]。所以矩阵 $\boldsymbol{G}$ 的 $m \times n$ 阶伪逆矩阵为:

$$\boldsymbol{G}^{+} = \boldsymbol{V}\sum{}^{+}\boldsymbol{U}^{\mathrm{T}} \tag{11}$$

其中 $\sum^{+}$ 是一个 $\boldsymbol{G}$ 的奇异值决定的 $n \times n$ 阶矩阵:

$$\sum{}^{+} = \mathrm{diag}\left(\frac{1}{\sigma_1}, \frac{1}{\sigma_2}, \cdots, \frac{1}{\sigma_k}, 0, \cdots, 0\right) \tag{12}$$

这样,伪逆矩阵的计算,将大大简化,接着使用式(4)计算网络的输出层权值。

5 结束语

利用广义径向基函数的思想构造一个神经网络对移动通信网问题进行分类。本算法采用k-均值聚类算法训练广义径向基函数网络中心,使用奇异值分解计算输出层权值。对该网络的实现细节及待改进之处进行了简要分析。

参考文献:

[1] Haykin S. Neural Networks: A Comprehensive Foundation [M]. 2nd edition. 北京:清华大学出版社,2001.

[2] Witten I H, Frank E. Data Mining: Practical machine Learning Tools and Techniques[M]. 2nd edition. [s. l.]: Elsevier Inc, 2005.

[3] Sing J K, Basu D K, Nasipuri M, et al. Self-Adaptive RBF Neural Network-Based Segmentation of Medical Images of the Brain[J]. Proceedings of ICISIP, 2005, 18(7): 447-452.

[4] Guo J J, Luh P B. Selecting Input Factors for Clusters of Gaussian Radial Basis Function Networks to Improve Market Clearing Price Prediction[J]. IEEE Transactions on Power Systems, 2003, 18(2): 665-672.

[5] Han J, Kamber M. Data Mining: Concepts and Techniques [M]. 2nd edition. [s. l.]: Morgan Kaufmann, 2006.

[6] Mclachlan G, Krishnan T. The EM Algorithm and Extension [M]. New York: John Wiley & sons, 1997.

[7] Chen S, Cowan C F N, Grant P M. Orthogonal Least Squares Learning algorithm for Radial Basis Function Networks[J]. IEEE Transactions on Neural Networks, 1991, 2(2): 302-309.

[8] Maciejewski A A, Klein C A. The Singular Value Decomposition and Application to Robotics[J]. The International Journal of Robotics Research, 1989, 8(6): 63-79.