文章编号: 1001-0920 (2012) 02-0211-05

一种基于信息熵的金融数据神经网络分类方法

冯 建[1,2], Starzyk Janusz[2], 邱菀华[1]

(1. 北京航空航天大学 经济管理学院, 北京 100191; 2. 俄亥俄大学 电机与计算机科学学院, 俄亥俄 45701)

摘 要: 讨论一种基于信息熵的神经网络数据分类方法, 通过所有神经元的统计权重信息对输入数据进行投票分类. 这种多层网络结构以及基于信息量的分割算法, 使得它在数据分类问题上比现有的多数神经网络具有更好的表现. 其并行的可扩展结构适合硬件实现, 能够提高实际运算速度, 适合用来处理金融方面高维度、复杂的海量数据问题.

关键词: 数据分类; 神经网络; 熵

中图分类号: C931.9 **文献标识码:** A

A classification approach of neural networks based on entropy for financial data

FENG Jian[1,2], *STARZYK Janusz*[2], *QIU Wan-hua*[1]

(1. School of Economy and Management, Beihang University, Beijing 100191, China; 2. School of Electrical Engineering and Computer Science, Ohio University, Ohio 45701, USA. Correspondent: FENG Jian, E-mail: jian.feng.bj@gmail.com)

Abstract: Through the entropy estimation, information theory-based learning is performed locally at each neuron. The input data are classified by using the weighted statistical information from all the neurons. Classification method based on multi-layer structure and information results in a better performance in data classification than many other existing methods of neural networks. This architecture can extend to a large embedded system to handle complex financial problems.

Key words: data classification; neural network; entropy

1 引 言

随着计算机技术的进步, 数据挖掘已成为数据库研究领域的一个重点. 其中数据分类作为数据挖掘的基本操作, 是数据挖掘领域中的关键技术之一[1]. 数据分类是根据数据集的特点构造一个分类器, 利用分类器对未知类别的样本赋予类别的一种技术. 构造分类器的过程一般分为训练和测试两个步骤. 在训练阶段,分析训练数据集的特点, 为各个类别产生一个对相应数据集的准确描述或模型. 在测试阶段, 利用类别的描述或模型对测试进行分类, 测试其分类准确度. 数据分类有许多不同的算法和模型, 比如 Bayesian 分类、决策树学习、统计方法和神经网络等. BP 神经网络是其中一种应用比较广泛的模型[2-3]. BP 模型虽然简捷, 但是当数据维度增大时, BP 神经网络模型在判别相似类别的差异时会遇到困难, 造成误分而影响精度[4].

本文以信息熵为基础, 提出一种基于统计学习理论的神经网络分类方法, 与传统的神经网络方法有很大不同. 它的分类算法是基于网络中所有神经元投票的统计结果, 不存在算法收敛与否的问题. 其可扩展的网络结构和松散连接方式适合可编程硬件的实现, 有利于对高维度数据特征的深入提取和分析, 对解决具有海量数据的高维度、高相关性的金融方面的问题具有重要的实践意义. 通过对几个有代表性的金融问题的分类实验, 表明了基于信息理论的统计学习算法使得这种方法对金融问题的分析准确性优于现有的多数人工神经网络.

2 基于信息熵的神经网络

近年来, 国内学者把多种改进的神经网络用于数据分类, 取得了优于传统神经网络分类的测试结果[5-6]. 本文讨论的神经网络与 BP 网络一样, 也是一种前馈结构. 在网络结构上它与传统的神经网络具

收稿日期: 2011-02-22; **修回日期:** 2011-03-29.

基金项目: 国家自然科学基金项目(70871002).

作者简介: 冯建(1970−), 男, 工程师, 博士, 从事项目管理、智能管理等研究; Starzyk Janusz(1947−), 男, 教授, 博士生导师, 从事机器学习、人工智能等研究.

有一定的相似性, 但它的内部组织以及对数据的学习机制都与传统的神经网络有较大区别. 图 1 是一个 16×14 的网络结构.

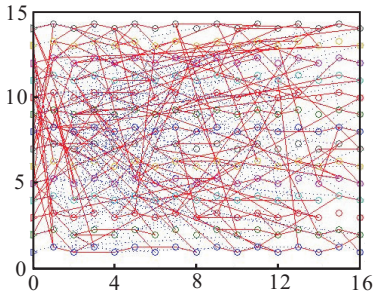

图 1 网络结构

2.1 神经元

图 1 中的圆圈代表具有相同运算能力的基本功能单元神经元. 每个神经元由它的行数和列数唯一确定. 每个神经元除了数据输入和数据输出外, 还有一个二进制的控制输入 τ_i 和两个控制输出 τ_o 和 $\bar{\tau}_o$. 每个神经元对应于它的数据输入和控制输入是一个事件驱动的处理器. 当控制信号为高时, 神经元被激活, 只对其输入空间所选中的数据作简单的算术或逻辑运算来降低输入数据的信息熵值, 从而获取数据的最大信息量. 在激活状态下, 神经元的各种统计信息和控制输出都会通过计算获得.

一个神经元通过将它的输出与一个选定的阈值 τ 进行比较, 分割它的输入空间. 分割边界定义为

$$\Psi_{k,l}(S_{k,1}) = \tau.$$

两个子空间分别为

$$S_{ot} = \{x \in S_{k,l} : \Phi_{k,l}(x) > \tau\}, \tag{1}$$

$$S_{oti} = \{x \in S_{k,l} : \Phi_{k,l}(x) \leqslant \tau\}. \tag{2}$$

这两个子空间分别对应于输入信号的两个子空间 T_k 和 T_{ki}, 即

$$S_{ot} = \Psi_{k,l}(T_k), \tag{3}$$

$$S_{oti} = \Phi_{k,l}(T_{ki}). \tag{4}$$

通过多个神经元对输入信号的处理, 也即多个基本变换的累积, 会对整个数据输入空间产生更复杂的变换, 如图 2 所示.

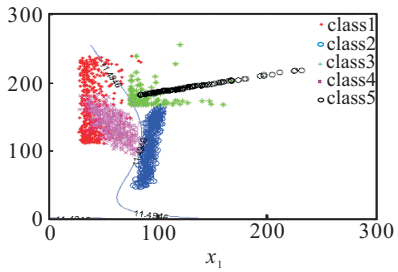

图 2 多个神经元对原始输入空间的分割

如前所述, 每个神经元仅当它的控制输入为高时才被激活. 它的控制输出由下式得到:

$$\tau_o = \tau_i \bigwedge \tau_n, \tag{5}$$

$$\bar{\tau}_o = \tau_i \bigwedge \bar{\tau}_n. \tag{6}$$

其中: τ_n 是神经元的阈值函数值 ($\Phi_{k,l}(x) > \tau$), τ_i 是控制源神经元的输出控制信号 τ_o 或 $\bar{\tau}_o$. 可见, 一个神经元的控制输入是该控制源链上的所有神经元的阈值函数值的乘积, 即

$$\tau_i = \prod_{k=1}^{m} \tau_k. \tag{7}$$

其中: m 是控制源链上的神经元数, τ_k 是链上每个神经元所用的控制输入 τ_n 或 $\bar{\tau}_n$. 根据式 (5) 和 (7), 如果一个神经元的控制输出为高, 则它所在的这条控制链上所有前级神经元的阈值函数值都为高, 即在它之前的这条控制链上的所有神经元都被激活, 最终所得到的输入子空间由这条链上的所有神经元分割得到.

2.2 神经网络结构

神经元与其附近的神经元随机相连, 构成网状结构. 每个神经元从与其相连的所有连接中选择它的输入数据来源, 并对选中的数据进行算术或逻辑变换, 将结果输出到下一级神经元. 神经元间的连接是松散式的, 并且是随机的, 间距越远, 神经元间互相连接的概率越低. 除了数据线的连接, 神经元间还有控制线连接, 这种连接也遵循局部性和松散性的原则. 每个神经元具有相同数量的数据连接线以及控制连接线. 传统的神经网络的连线数量随神经元的个数增加呈立方递增[7], 而这种结构中的连线数量随神经元个数的增加而呈线性增加, 因此它会大大节省实际应用中的软硬件资源.

图 1 中的三角形符号表示外部的数据输入, 实线表示数据连接线, 虚线表示控制连接线. 在这种多层结构中, 每一层 (列) 都有相同个数的神经元. 第 1 列是外部数据输入, 输入数据直接连接到第 1 层上的所有神经元. 通常, 每一层上的神经元个数都取大于或等于输入数据的维度. 基于网络结构的自组织特点, 网络的层数能在对数据的学习过程中由学习算法自动确定. 对于这种结构的神经网络而言, 对数据的学习过程即是对其网络结构的优化过程. 通过对输入数据的学习, 神经元间的连接以及神经元内部的功能都会被提炼和优化.

3 基于统计理论的分类方法

3.1 学习算法

网络中神经元同步对输入数据进行学习, 通过计算信息熵来选取最优的变换函数和阈值以获得最大的信息量. 学习结束时, 与最优信息熵相对应的输入数据选择、变换函数和阈值都被保存下来. 在训练阶

段, 网络连接以及每个神经元根据输入数据的信息进行调整而趋于一个适应性的稳定结构. 当系统在测试阶段时, 选中的控制输入将决定一个神经元是否被激活并对某个测试样本的分类进行投票, 最终的分类结果将由所有被激活的神经元投票获得.

在训练阶段, 神经元选取输入数据并通过变换函数将其变换为输出数据, 通过和设定的阈值进行比较, 输入数据能被确定属于由阈值分割出的两个子空间 S 或 S_i 其中之一. 每个神经元有专门的计数器分别记录以下数据:

1) 总的数据个数 n_t;

2) 满足阈值的数据个数 n_s;

3) 属于某类的且满足阈值的数据个数 n_{sc};

4) 属于某类的不满足阈值的数据个数 n_{sic}.

学习的质量由网络从数据中获取的信息量来确定. 为了计算从数据中获得的信息量, 需要针对训练样本计算如下概率:

1) 某一类数据满足阈值的概率

$$P_{sc}=\frac{n_{sc}}{n_t}.$$

2) 某一类数据不满足阈值的概率

$$P_{sic}=\frac{n_{sic}}{n_t}.$$

3) 子空间概率 (满足阈值)

$$P_s=\frac{n_s}{n_t}.$$

4) 子空间概率 (不满足阈值)

$$P_{si}=1-P_s.$$

5) 某类数据的概率

$$P_c=\frac{n_c}{n_t}.$$

利用信息熵概念, 通过不同类的训练数据落在每个子空间的概率, 定义了从输入数据所获取的信息量, 即

$$I=1-\frac{\Delta E_s}{E_{\max}}=1-\left\{\left[\sum_{sc}P_{sc}\log(P_{sc})-P_s\log(P_s)\right]+\left[\sum_{sic}P_{sic}\log(P_{sic})-P_{si}\log(P_{si})\right]\right\}\Big/\sum_{c}P_c\log(P_c). \tag{8}$$

式 (8) 是单位化的信息量. $I=0$ 表示数据的信息熵没有减少, 数据所包含的信息量为零; $I=1$ 表示输入数据的信息熵减少为 0, 输入数据的信息完全确定. 信息量也是对神经元分割输入空间程度的度量, 代表了对输入数据所包含信息获取的程度. 对每个神经元而言, 选择不同的输入数据、变换函数以及阈值的组合都会得到不同的信息量. 图 3 和图 4 给出了一个神经元在相同输入数据情况下, 通过不同的变换产生出不同的信息量的例子. 从图中可以看出, 通过乘法变换能从数据中获取比加法变换更大的信息量. 另外, 每种变换的最大信息量都和一个阈值相对应, 合适的阈值能将数据空间进行最优化分割; 对一个神经元而言, 如果在所有的组合中这是所能获得的最大的信息量, 则乘法变换以及相应的阈值将被该神经元保存下来.

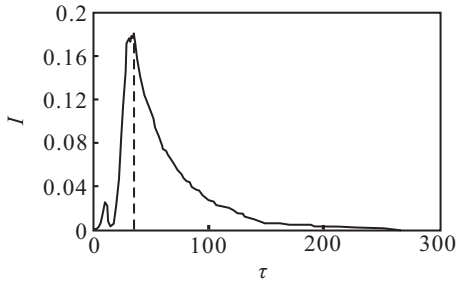

图 3　通过加法变换计算神经元获取的信息量

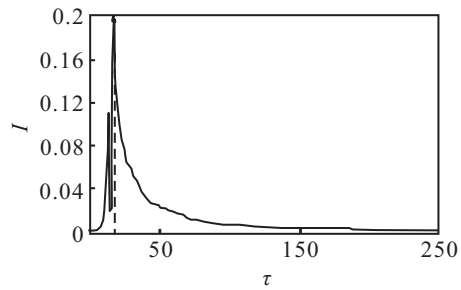

图 4　通过乘法变换计算神经元获取的信息量

在训练阶段, 每个神经元对输入数据进行选择, 积极地调整变换函数和阈值来获得最大化的信息量. 当网络完成对所有样本数据的学习后, 每个神经元的输入连接、变换函数、阈值等都已根据训练数据所包含的信息而确定, 即网络根据从训练数据中获取的信息进行了自组织优化.

3.2　数据分类投票

完成训练阶段的学习过程后, 可以将测试样本数据输入到神经网络中, 网络会根据训练结果对测试样本进行分类. 每个神经元在训练阶段都记录了每类数据在其输出空间的识别概率 P_{sc} 和 P_{sic}, c 代表 C 类数据. 如果一个输入数据落在激活神经元的输入子空间内, 则这个神经元便会用它所记录的概率作为其对 C 类数据正确识别的概率 P_{cc} 来对这个数据属于 C 类的概率投票. 单个神经元的识别概率不能反映出真实的分类情况, 在这里, 采用一种 MRC(maximum ratio combination) 权重函数的投票机制来对输入数据进行分类[8].

$$B_c=1-1\Big/\left\{1+\sqrt{\sum_{i=1}^{n}\left(1\Big/\left\{\frac{1}{P_{cci}}-1+\varepsilon\right\}\right)^2}\right\}. \tag{9}$$

其中: P_{cci} 为每个参与投票的神经元识别其为C类数据的概率; n 为参与投票的神经元个数; ε 为防止分母为 0 的极小数. 这个权重函数提供了一种对所有神经元投票的统计结果. 当 B_c 在所有类中为最大时, 网络会将输入样本归为C类.

另外, 由于初始网络连接是随机的, 每个网络具有不同的初始连接, 这可能导致每个网络会有表现上的差异. 为了均衡这种表现上的差异, 可以同时产生多个神经网络, 对它们同步进行训练, 最终的分类结果由所有网络集体投票来确定. 实验证明, 由多个神经网络集体投票的机制能提高分类结果的精确程度.

4 金融数据分类仿真

这种神经网络适于对高维的海量数据进行处理, 又具有适合硬件实现的结构特点, 这样能加快数据的处理速度, 因此用来分析数据量大、数据复杂的金融问题尤为适宜. 为了验证模型的精确性, 通过Matlab软件仿真, 对3个金融领域比较典型的问题进行了分析:

1) 对公司破产预测分析. 在过去的几十年中, 很多经济学者热衷于对公司破产预测的研究. 其中较具代表性的是Atiya对现存的预测方法作了综合的介绍, 并用传统的神经网络对该问题进行了研究, 得出了比较准确的结果[9]. 他收集了716家美国的有偿债能力的公司和195家(1∼36个月内)财务拖欠破产公司的数据. 考虑到一些破产前的不同案例, 他把数据集扩展成1 160个数据, 并把它们分为样本集和非样本集, 分别作为训练和测试数据集. 在他的实验中, 采用了两种指标系统: 一种只基于财政比率; 另一种同时基于财政比率和净资产. 通过采用这两种指标系统, 对公司破产预测的准确率相比传统的预测方法得到了很大提高. 这其中很大原因要归功于他选择了合适的指标系统. 在本文的实验中, 首先也采用基于财政比率和资产净值的指标系统, 利用本文提出的神经网络对非样本集中的公司进行预测, 两种方法的结果如表1所示. 可以看出, 基于信息熵的神经网络表现略好.

如前所述, Atiya通过专家知识选择了适当的分类指标系统, 从而大大提高了预测的准确性. 由于基于信息熵的神经网络适合处理高维度、复杂的海量数据, 本文放弃上述选择的指标系统, 而是将所有指标(共63个)用来作为预测的依据. 结果显示, 最好的神经网络预测的正确率达到了90.04%, 预测精度得到了明显提高, 如表1第3列所示.

2) 对股票投资决策分析. 在这个例子中, 将基于信息熵的神经网络用于对股价变化的预测, 并与支持向量机(SVM)分类器的结果进行比较. 数据采用Research Insight从标准普尔超过10 000家在NYSE, AMEX, NASDAQ, OTC公开上市交易的美国公司提取的一个数据库[10]. 这些公司公布了其最近20年的财务数据, 包括收入、支出、现金流等财务指标. 以每3年为一个时间段, 训练样本和测试样本基于从数据库中提取的192个特征值来构建. 实验中, 根据上市公司的股价在3年期紧随的后一年的变化情况, 定义了两种分类类别: 一年中股价增值低于市场平均增值水平的公司为第1类; 股价增值超过市场平均增值水平的公司为第2类. 每3年一个的训练样本集作为输入数据对分类器进行训练. 对数据库的准备工作包括: 将有数据缺失的公司和其他公司分离; 对缺省数据的恢复; 由于这些财务数据极其庞大而且维数很高, 可通过非线性PCA分析法[11]来降低数据特征值间的高相关性. 通过这种方法, 数据的特征值个数可以从192个下降到22∼25个.

表1 公司破产预测正确率

破产期限	预测正确百分率/%		
	传统神经网络[9]	基于信息熵的神经网络	
	采用选择的指标系统	采用选择的指标系统	采用全部指标
少于6个月	86.15	85.11	87.23
6∼12个月	81.48	84.09	86.36
12∼18个月	74.60	76.19	90.24
18∼24个月	78.13	55.17	72.24
24个月以上	66.67	64.29	75.00
1∼36个月	78.13	75.13	83.96
有偿还能力公司	90.07	92.74	93.42
总共	85.50	85.80	90.04

通过这种实验方法, 得到了对不同数据集的交叉验证结果. 表2给出了使用基于信息熵的神经网络方法以及采用C-SVM方法[12]的分析结果. 表中的每一行代表了由不同年的数据训练的分类器, 如C2000, C2001, C2002, C2003. 每一列显示了用不同年的数据测试所得到的分类正确率. 比如分类器C2000是由1998∼2000年期间的数据训练得到的, 并用来对2001年, 2002年和2003年的股价变化进行预测. 如果预测的准确率超过50%, 则表明可以获得超过市场平均表现水平的投资决策. 从表2中可以看出, 基于信息熵的神经网络在对股价变化的预测上, 和SVM方法一样都能取得超过50%的正确率.

3) 信用卡申请的批准. 这是一个研究比较多的例子, 很多分类算法都可以应用于这个问题. 数据来源于UCI机器学习数据库, 每条数据包含14种属性. 在将样本数据应用到模型前需对其预处理, 即将定性属性进行量化, 并对于离散型属性的缺失值, 采用与该样例类别相同的样例集中最频繁的取值填充; 同时为了排除变量之间单位的差异, 需对数据进行标准化处理.

鉴于这种模型与神经网络和SVM算法具有一定

表 2　股票投资决策预测准确率

分类器		测试年代					
		2000	2001	2002	2003	平均准确率/%	标准偏差
基于信息熵的神经网络	C2000	——	0.620 17	0.607 18	0.539 6	0.589 0	0.043 3
	C2001	0.546 73	——	0.544 54	0.515 68	0.535 7	0.017 3
	C2002	0.567 3	0.580 83	——	0.509 28	0.552 5	0.038 0
	C2003	0.527 55	0.490 51	0.499 08	——	0.505 7	0.019 4
SVM	C2000	——	0.584 05	0.566 25	0.573 05	0.574 5	0.009 0
	C2001	0.561 97	——	0.608 75	0.461 93	0.544 2	0.075 0
	C2002	0.549 15	0.572 65	——	0.478 4	0.533 4	0.049 1
	C2003	0.606 84	0.558 4	0.62	——	0.595 1	0.032 4

的相似性, 实验中主要关注这几种算法的比较. 采用与文献 [13] 中一样的交叉验证方法来消除统计误差, 实验结果和其他几种算法[13-14]的测试结果在表 3 中进行了比较. 可以看出, 对同样的数据样本的识别, 这种方法的错误率是很低的. 尽管比决策树算法 CAL5 略微高一些, 却是所有神经网络和 SVM 算法中表现最好的. 值得指出的是, 融合了专家知识的决策数算法本身适宜于解决这种类型的分类问题[11], 而实验中采用的基于信息熵的神经网络却不是专门为这种问题而设计的, 它具有更广泛的适用性. 另外, 网络中单个神经元的行为和 SVM 算法有一定程度的相似性, 都是通过阈值来分割数据输入空间. 表 3 中的两种 SVM 算法, SVM-light 和 BSVM 在对这一问题的表现上都不如这种神经网络的精确度高.

表 3　信用卡申请问题多种算法的比较

算法	误判率/%	算法	误判率/%
CAL5	0.131	ITule	0.173
基于熵的神经网络	0.135	Naivebay	0.151
SVM-light	0.138	CASTLE	0.148
BSVM	0.138	ALLOC80	0.201
DIPOL92	0.141	CART	0.145
Logdisc	0.141	NewID	0.181
SMART	0.158	CN2	0.204
C4.5	0.155	LVQ	0.197
IndCART	0.152	Kohenen	-
BP NN	0.154	Quadisc	0.207
Discrim	0.141	Default	0.440
RBF	0.145	AC2	0.181
Baytree	0.171	k-NN	0.181

5　结　论

本文讨论了一种基于信息熵的神经网络的数据分类方法, 它结合以信息理论为基础的统计学习算法, 能够处理非线性问题. 它和传统的神经网络在结构上具有一定的相似性, 但其单个神经元的行为又与 SVM 算法相似. 这种结构和算法特点使其具有比传统的神经网络更强大的对海量、复杂数据的处理能力, 尤其适合对数据量大且复杂的金融问题的研究. 它的松散连接方式以及可扩展的规则网络结构适合硬件的实现, 能帮助提高实际应用的运算速度, 增大了在实践中应用的可能. 尽管本文应用这种方法主要对金融方面的问题作出了分析, 但它是一种通用的统计学习模型, 而不是针对某一类型的问题设计的, 可以根据不同的应用领域重复配置使用, 提高了实际应用的灵活性. 软件仿真表明, 基于信息熵的神经网络的数据分类方法是可靠的, 在对高维、复杂的金融数据的分类问题上, 比现存的多种神经网络以及 SVM 算法具有更高的精确度.

参考文献(References)

[1] Jiawei Han, Micheline Kamber. Data Mining: Concept and Technology. Beijing: China Machine Press, 2000.

[2] 崔丽群, 刘万军, 包明宇. 基于神经网络数据分类方法. 辽宁工程技术大学学报, 2004, 23(4): 507-509.
(Cui L Q, Liu W J, Bao M Y. Research of data classification based on neural networks[J]. J of Liaoning Technical University, 2004, 23(4):507-509.)

[3] 王野乔. 遥感及多源地理数据分类中的人工神经网络模型[J]. 地理科学, 1997, 17(2): 105-111.
(Wang Y Q. Artificial neural network models in remote sensing and multisource geographical data classification[J]. Scientia Gelgeraphica Sinica, 1997, 17(2): 105-111.)

[4] 黄旭钊, 梁月明, 李桐林. 基于 BP 神经网络的航空物探数据分类方法[J]. 物探与化探, 2010, 34(4): 485-488.
(Huang X Z, Liang Y M, Li T L. The classification of cerogeophysical data based on BP neural network[J]. Geophysical and Geochemical Exploration, 2010, 34(4): 485-488.)

[5] 商琳, 王金根, 姚望舒, 等. 一种基于多进化神经网络的分类方法[J]. 软件学报, 2005, 16(09): 1577-1583.
(Shang L, Wang J G, Yao W S, et al. A classification approach based on evolutionary neural networks[J]. J of Software, 2005, 16(9): 1577-1583.)

[6] 黄国宏, 熊志化, 邵惠鹤. 一种新的基于构造型神经网络分类算法[J]. 计算机学报, 2005, 28(9): 1519-1523.
(Huang G H, Xiong Z H, Shao H H. A new classification algorithm based on constructive neural networks[J]. Chinese J of Computers, 2005, 28(9): 1519-1523.)

(下转第226页)