

基于 MS 神经网络算法的数据挖掘应用的探讨

刘城霞

(北京信息科技大学 计算机学院, 北京 100101)

摘要: Microsoft 神经网络算法是基于人体神经网络系统模拟而成的一种算法, 它对于数据挖掘的发展有着很大的推动性。为了进一步发展基于神经网络算法的数据挖掘系统的应用, 在 Microsoft 神经网络算法的基础上构建了一个数据挖掘商业应用实例系统, 通过研究客户的一些个人属性以及办理业务的基本情况, 预测客户的信誉情况、业务的办理趋向、银行开展新业务的趋向等。在实例系统的构建过程中, 对神经网络数据挖掘算法的挖掘过程进行了详细的分析, 促进了数据挖掘的应用实践。

关键词: 神经网络; 数据挖掘; 预测

中图分类号: TP39

文献标识码: A

文章编号: 1673-629X(2011)06-0235-04

Discussion of Application of Data Mining Based on Microsoft Neural Network Algorithm

LIU Cheng-xia

(Computer School, Beijing Information and Technology University, Beijing 100101, China)

Abstract: Microsoft neural network algorithm is the simulation of human's neural network system and enhances the development of data mining. In order to improve the application of data mining system based on neural network algorithm, a business data mining system is created based on Microsoft neural network algorithm. Using the application system analyze the customer's attributes to predict his credit and business tendency. In the creation of the instance model system the whole program of data mining is introduced in detail and this helps the development of data mining's application.

Key words: neural network; data mining; prediction

0 引言

数据挖掘(Data Mining)是当今的研究热点, 它其实就是从海量的、有噪声的、干扰的、不完全的、模糊的、随机的数据中, 提取隐含的、事先不知道的、而又潜在有用的信息和知识的一个过程^[1]。从商业的角度来说, 数据挖掘是一种进行信息处理的技术。由于各个行业中的业务都进行计算机自动处理, 在各个行业领域都产生了大量的相关数据, 需要对数据库中的海量业务数据进行数据抽取、分析转换和模型化处理后, 得到帮助判断的关键性的数据, 为商业决策提供真正有价值的信息, 这有利于商业运作、提高竞争力, 进而获得利润。如国内外有些学者对金融、证券、保险等行业信息进行数据挖掘, 研究客户的信用度^[2,3], 进而帮助商业管理者进行决策。因此, 在商业中数据挖掘的应

用可以描述为: 按企业的业务目标, 对大量的企业积累的已有数据进行综合分析, 揭示隐藏在其中的、未知的规律性, 并更进一步将其进行模型化, 方便以后分析处理的一种先进有效的方法。

1 Microsoft 神经网络算法介绍

“神经网络”(Neural Network, 简称 NN)是在对大脑组织结构和运行机制的认识理解基础之上模拟其结构和智能行为的一种工程系统。20 世纪 40 年代初期, McCulloch 和 Pitts 提出了人工神经网络的第一个数学模型, 从此开创了神经网络理论^[4,5]的研究时代。经过了六十多年的发展, 神经网络算法已经广泛地应用到模式识别、模糊控制、知识工程、专家系统、人工智能等领域。而它在数据挖掘中的应用研究^[6-8], 利用了神经网络的非线性和鲁棒性的特点, 对数据库中大量的数据进行知识发现, 提供人们有用的知识。

1.1 Microsoft 神经网络结构

Microsoft 神经网络算法是由最多三层神经元组成的网络^[9], 这些层如图 1 所示, 分别是输入层、可选隐

收稿日期: 2010-10-31; 修回日期: 2011-02-27

基金项目: 北京市人才强教计划——骨干教师资助项目(PHR2010 08428)

作者简介: 刘城霞(1978-), 女, 讲师, 硕士, 研究方向为数据挖掘、数据融合、信息安全及数据结构与算法。

藏层和输出层。

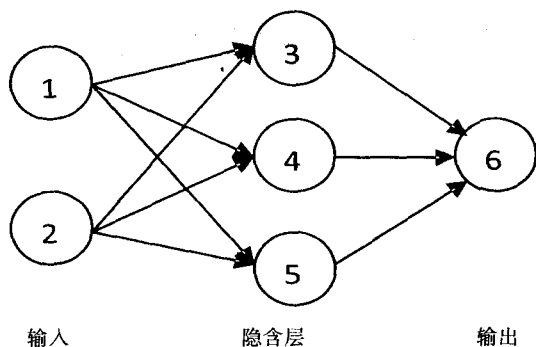


图 1 神经网络模型

输入层:输入神经元定义数据挖掘模型的所有输入属性值及其概率。

隐藏层:隐藏神经元接收来自输入神经元的输入,并向输出神经元提供输出。隐藏层是向各种输入概率分配权重的位置。权重说明某一特定输入对于隐藏神经元的相关性或重要性。输入所分配的权重越大,则输入的值越重要。权重可为负值,表示输入抑制而不是促进某一特定结果。

输出层:输出神经元代表数据挖掘模型的可预测属性值。

1.2 Microsoft 神经网络算法过程

数据从输入经过中间隐含层到输出,整个过程是一个从前向后的传播数据和信息的过程,后面一层结点上的数据值从与它相连的前面的结点传来,之后把数据加权后经过一定的函数运算得到新的值,继续传播到下一层结点。这个过程称为前向传播。

当结点的输出发生错误时,也就是与预期不同,神经网络就要自动“学习”。后一层结点对前一层结点有一个“信任”程度(结点间连接的权重),采用惩罚的方法来学习:如果结点输出出错,那就要查看这个错误是受哪些输入结点的影响,降低导致出错的结点连接的权重,惩罚这些结点,同时提高那些做出正确建议结点的连接的权重。对那些受到惩罚的结点来说,也用同样的方法来惩罚它前面的结点,直到输入结点为止。这称为回馈。

对训练集中的所有数据重复这个过程:即通过前向传播得到输出值,用回馈法进行学习。当把训练集中的所有数据都运行过一遍之后,则称完成了一个训练周期。要完成整个神经网络的训练经常需要几百个训练周期。训练后得到的神经网络模型,包含了训练集中响应值受预测值影响变化的规律。

1.3 Microsoft 神经网络算法函数

隐含层的传输函数通常使用 Sigmoid 型或者 Tanh 型。这些均是非线性函数,并且类似于生物学中神经网络的基本传输特征,即:输入值发生的细微变化有时

会产生较大的输出变化。

Sigmoid 和 Tanh 的定义是:

$$\text{Sigmoid: } O = 1 / (1 + e^{-a}) \quad (1)$$

$$\text{Tanh: } O = (e^a - e^{-a}) / (e^a + e^{-a}) \quad (2)$$

其中 a 是输入值,而 O 是输出值。

处理反向传播,计算误差,更新权值时输出层所用到的误差函数为交叉熵:

$$\text{Err}_i = O_i * (1 - O_i) * (T_i - O_i) \quad (3)$$

上述公式中 O_i 是输出神经元 i 的输出,而 T_i 是基于训练样例的该输出神经元实际值。

隐含神经元的误差是基于下一层中的神经元的误差和相关权值来计算的。计算公式:

$$\text{Err}_i = O_i(1 - O_i) \sum_j (\text{Err}_j * w_{ij}) \quad (4)$$

其中 O_i 是输出神经元 i 的输出,该单元有 j 个到下一层的输出。 Err_j 是神经元 j 的误差, w_{ij} 是这两个神经元之间的权值。

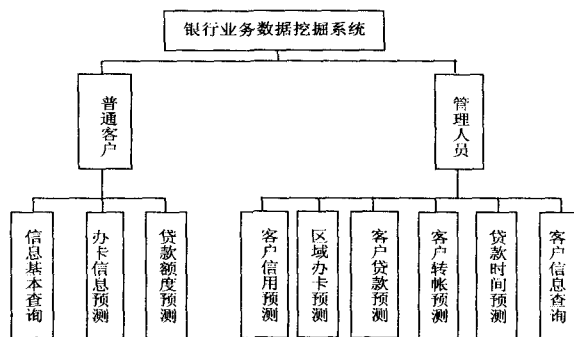
一旦计算出每个神经元的误差,则下一步是使用以下方法来调整网络中的权值。

$$w_{ij} = w_{ij} + l * \text{Err}_i * O_i \quad (5)$$

其中 l 为 0~1 范围内的数,称之为学习函数。

2 数据挖掘系统的构建

数据挖掘的目标是从数据库中发现隐含的、有意义的知识,而数据挖掘应用系统的目标就是将发现的知识用于信息的管理、查询的优化以及决策支持和过程控制等。基于神经网络的数据挖掘应用系统的研究也正在如火如荼地进行中^[10-12]。文中以银行应用为例,以关系数据库中的数据为基础,以神经网络数据挖掘算法为主体,建立为银行客户和管理者共同使用的数据挖掘系统(见图2),通过预测未来趋势及行为,为做出前摄的、基于知识的决策提供帮助。



2.1 系统需求分析

该系统的使用者为普通用户和管理员。其中,银行业务管理者为管理员;其他用户为普通用户。普通用户的权限是查询个人的贷款信息以及转账信息,预测个人贷款最大额度等。而管理员不仅可以查询所有

用户的相关信息,还可以对客户的信誉、业务方向等进行预测。

在系统设计时除了注重客户的基本需求功能外,还要尽可能地将预测的数据清晰明了地展现给用户以方便用户的分析和制定未来方案,要求:

(1)正确性:预测结果是根据以往数据训练后的预测模型得到,结果符合用户需要。

(2)完整性:系统结构完整,提供用户完整的信息及操作功能。

(3)清晰性:将预测结果以表格的方式显示外还提供了图形化显示。

(4)易用性:提供方便的可视化查询方式,允许用户从多个角度查看预测结果,方便用户简单快捷地分析每条预测信息。

(5)扩充性:系统既能高效完成现有的业务处理需求,将来又能根据需求增加功能。

2.2 数据挖掘模型的建立

2.2.1 数据库中各个表间的关系

数据库中包含一个 account 表,一个 client 表,一个 disp 表,一个 order 表,一个 trans 表,一个 loan 表,一个 card 表,一个 district 表,一共八张表。数据库表间关系图见图3。

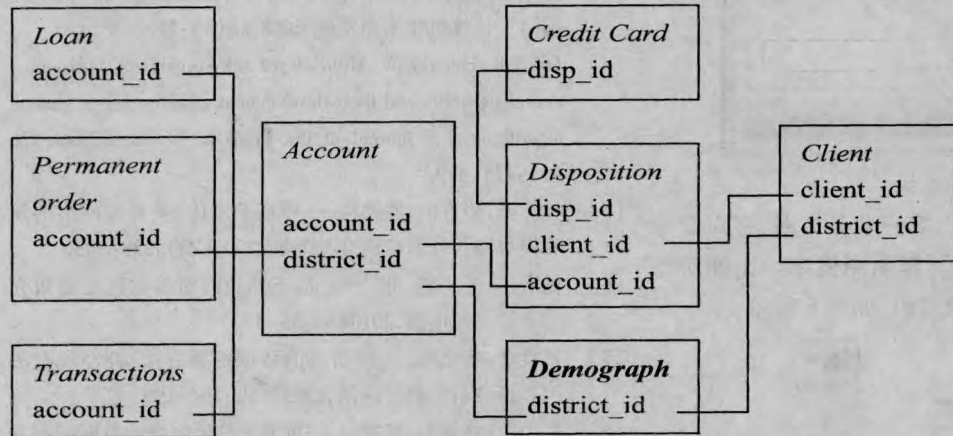


图3 数据库表间关系图

2.2.2 建立数据挖掘模型

第一:建立微软分析服务的项目。

第二:新建数据源及视图。

第三:分析可预测项。本系统针对客户信誉、客户业务方向、客户贷款、客户业务频率、贷款时间、转账信誉和转账趋势进行预测。

第四:建立神经网络挖掘结构。

第五:训练数据并预测。

对于前两步比较简单,文中主要介绍如何建立预测挖掘模型及训练预测过程。

2.2.3 建立挖掘预测模型

根据用户的需求,银行业务管理员可以对客户信

誉、客户贷款时间、客户贷款趋向、客户还款时间、客户转账趋向、区域业务等进行预测。而客户本身可以对贷款额度等进行预测。

下面以客户信誉预测为例,编写出预测模型的代码:

```

create mining model CreditPrediction
(
  account_id text key,
  amount text discrete,
  duration text discrete,
  payments text discrete,
  [status] text discrete predict_only
)
using Microsoft_Neural_Network
insert into CreditPrediction ( account_id, amount, duration, pay-
ments, [status] )
openquery([ Warehouse ], 'select account_id, amount, duration,
payments, [status] from loan')

```

```

/* model prediction */
select T. account_id , T. amount, T. duration, T. payments, pre-
dict histogram([ status])
from CreditPrediction prediction join
openquery([ Warehouse ], 'select account_id, amount, duration,
payments, [status] from loan')
as T
on CreditPrediction. ac-
count_id = T. account_id
and CreditPrediction. a-
mount = T. amount
and CreditPrediction.
[Status] = T. [status]
and CreditPrediction. Du-
ration = T. duration
and CreditPrediction. Pay-
ments = T. payments

```

通过模型,可以得到用户信誉等级的预测,为银行是否可以给客户进行贷款等做参考。另外其他方面的预测也可以帮助银行进行客户的评判和业务开展好坏的判断。比如区域业务预测可以帮助银行决定是否在某个地区开展某项业务。用户可以预测与自己有关的信息,比如贷款能否获批,贷款额度限额等。

3 结果分析

系统主界面以表格的形式展示出客户经挖掘模型预测后的数据,也可以选择对银行的客户群的整体情况直观方便地进行图形展示,图形可以改变其形状和展示方式。系统主界面如图4所示。

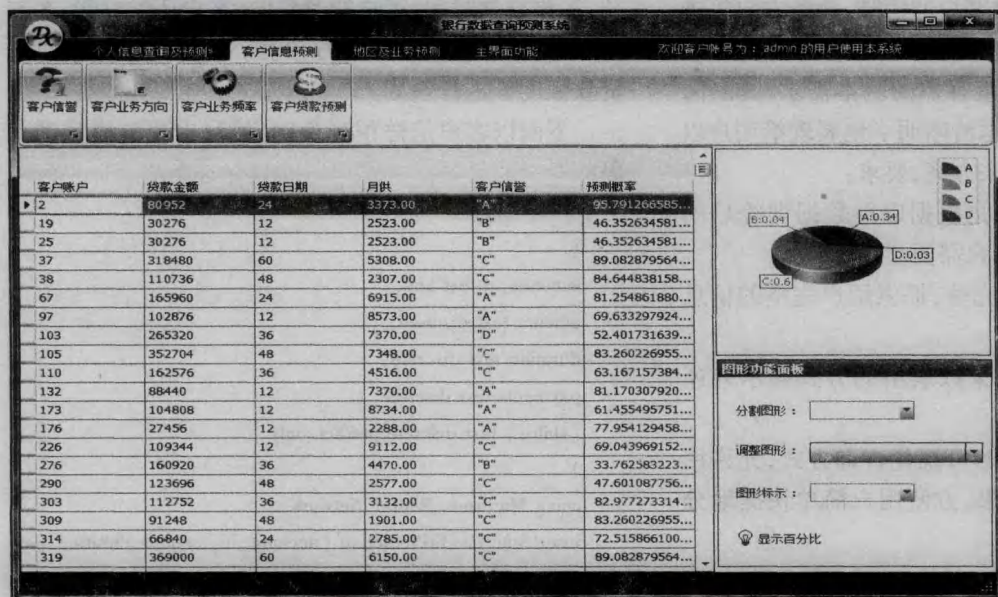


图 4 系统主界面

还可以对于查询结果进行再度查询,比如对客户信誉在某个范围内的信息进行查询,如图 5 所示。

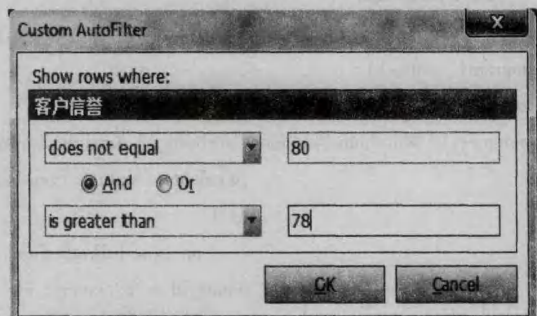


图 5 范围查找

系统还分别提供客户信誉、客户业务方向、客户业务频率和客户贷款的预测,并计算预测概率。比如对所有客户信誉分布情况预测的图形,如图 6 所示。

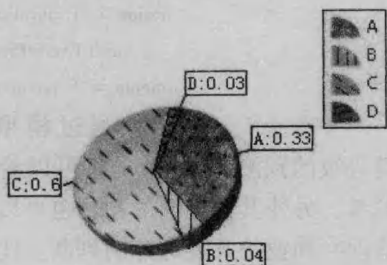


图 6 信誉分布预测

表示所有客户中,A等级的客户占33%,B等级的客户占4%,C等级的客户占60%,D等级的客户占3%。系统还有其他的预测功能的图形,篇幅关系不再赘述。

4 结束语

人们对数据的应用从简单的查询数据,到从数据中挖掘知识并提供决策支持,这就是数据挖掘的工作。

基于神经网络算法的数据挖掘以其在处理不确定性问题中独特的优势越来越被关注和使用。文中针对基于神经网络算法的数据挖掘系统进行了分析和设计,并实现了一个完整的银行业务数据挖掘系统,对客户数据及区域情况等进行信息的挖掘预测,促进了基于神经网络的数据挖掘系统的应用研究。

参考文献:

- [1] 安淑芝. 数据仓库与数据挖掘[M]. 北京:清华大学出版社, 2005.
- [2] 赵裕啸,倪志伟,王园园,等. SQL Server 2005 数据挖掘技术在证券客户忠诚度的应用[J]. 计算机技术与发展, 2010,20(2):229-232.
- [3] 陈 艳,张燕平. 数据挖掘技术在保险客户信用评估的应用[J]. 计算机技术与发展, 2008,18(5):179-181.
- [4] Oh S K, Pedrycz W. Multi-layer self-organizing polynomial neural networks and their development with the use of genetic algorithms[J]. Journal of the Franklin Institute, 2006, 176(5):475-489.
- [5] 林 香,姜青山,熊腾科. 一种基于遗传 BP 神经网络的预测模型[J]. 计算机研究与发展, 2006(Z3):338-343.
- [6] 杨启仁,王 娟,张 科. 基于神经网络的数据挖掘研究[J]. 信息与电脑, 2010(8):36-37.
- [7] 蒋良孝,蔡之华. 一种新兴的数据挖掘方法:神经规则法[J]. 计算机工程与应用, 2003(15):194-199.
- [8] 郑志军,林霞光,郑守淇. 一种基于神经网络的数据挖掘方法[J]. 西安建筑科技大学学报, 2000(3):28-30.
- [9] 微软公司技术文档. Microsoft 神经网络算法技术参考[EB/OL]. 2008. <http://msdn.microsoft.com/zh-cn/library/cc645901.aspx>.
- [10] Zhang Defu, Jiang Qinshan, Li Xin. Application of neural networks in financial data mining[J/OL]. The 2004 Int'l Journal of Computational Intelligence. 2004. <http://www.enformatica.org/ijci/>.
- [11] 齐莉丽,孙可娜. 神经网络在商业银行金融风险评价系统建模中的应用[J]. 天津职业技术师范学院学报, 2004(4):46-49.
- [12] Zhang Ling, Zhang Bo. A Geometrical Representation of McCulloch-Pitts Neural Model and Its Applications[J]. IEEE Trans. on Neural Networks, 1999, 10(4):925-929.