基于单层感知器的数据挖掘分类的设计和实现

王必强,毕硕本,董学士
(南京信息工程大学 计算机与软件学院,江苏 南京 210044)

摘 要:数据挖掘是指从大型数据库或数据仓库中提取人们感兴趣的知识,这些知识是潜在有用信息。分类是数据挖掘重要研究方向之一,其目的就是分析输入数据,通过分析在训练集中的数据表现出来的特性,为每一个类找到一种准确的描述或者模型。怎样用科学合适的方式来解决分类问题,是数据挖掘研究领域的一个热点和难点。通过构造一种单层感知器神经网络的分类方法,对其进行设计分析和仿真实验,用图文并貌的界面形象直观地展示了分类效果,实验表明单层感知器神经网络可有效地进行数据挖掘分类。

关键词:单层感知器;神经网络;分类;数据挖掘
中图分类号:TP311 **文献标识码**:A **文章编号**:1673-629X(2010)09-0111-04

Design and Implementation of Classification Mining Based on Single-Layer Perceptron Artificial Neural Network

WANG Bi-qiang, BI Shuo-ben, DONG Xue-shi
(School of Computer & Software, Nanjing University of Information Science & Technology, Nanjing 210044, China)

Abstract: Data mining is to extract the interested potential knowledge from the large database and data warehouse. Classification is one of the most important research directions of data mining, which aims to find an accurate description or model for each category by analyzing the characteristics of data in the training set. How to solve the problem of classification in a scientific way is a hot spot and difficulty in the field of data mining research. In this paper, propose a data mining classification method based on the single-layer perceptron neural network. Some simulation experiments are made to verify the effectiveness and the feasibility of the proposed methods, and the classification results are graphically displayed and demonstrate that the single-layer perceptron neural network can be used to solve the problem of classification data mining effectively.

Key words: single sensor; neural network; classification; data mining

0 引 言

数据挖掘,又称为数据库中的知识发现(Knowledge Discovery in Database, KDD),顾名思义就是从数据库中大量的无序数据中进行有目的的分析,获取潜在有用的模式,"挖掘"一词便由此而来[1,2]。数据挖掘的思想主要来自于两个领域:统计学、人工智能和模式识别。它包含的内容很多,文中介绍的分类即是其中的一个重要分支。

分类技术可以应用到很多领域,包括商业[3,4]、安全[5,6]、文本分析[7]、交通[8,9]等具有明显的分类意义的行业。例如在交通管理中,可以对各路段在不同时段的繁忙程度划分为:繁忙路段,不繁忙路段。进而再对其时间上的繁忙程度进行划分,这样就可以对交通疏导提供准确有效的参考依据。分类还可以用于文献检索和网页搜索;在安全领域,其可以用于入侵检测。对于分类规则的挖掘通常有以下几种方法:决策树方法、贝叶斯方法、人工神经网络方法、粗糙集方法和遗传算法。

收稿日期:2010-01-24;**修回日期**:2010-05-07
基金项目:中国气象局公益性行业科研专项经费资助项目(GYHY 200806017)
作者简介:王必强(1983-),男,陕西澄城人,硕士研究生,研究方向为信息融合、GIS、人工智能等;毕硕本,教授,博士后,研究方向为地理信息系统(GIS)、数据挖掘、人工智能等。

1 分类数据挖掘

分类通常用于预测未知数据实例的归属类别,是数据挖掘过程的一个关键过程,它通过分析输入数据,通过在训练集中的数据表现出来的特性,为每一个类找到一种准确的描述或者模型[10]。

1.1 分类挖掘步骤

从数据库中提取出分类规则通常需要经历以下七个步骤:

1.1.1 定义目标

目标定义的任务是产生一个目标清单说明,或者是提出可能产生的假定或所期望的结果。

1.1.2 创建目标数据集

该步骤从要进行数据挖掘的数据库中提取出要进行分析的初始数据,这些数据不仅包括孤立的数据,而且包括数据之间的关系及属性,还要进行数据之间条件属性和决策属性的区分。

1.1.3 数据预处理

主要是处理噪声数据,并对缺损数据进行处理。主要包括:与分类任务有关的数据属性的选择、数据抽样和删除属性值不完全的元组。其中进行数据抽样的原因是数据挖掘中的数据常常具有多维性,搜索空间多数情况下都是与属性的个数成指数增长的,因此数据的多维性是较难解决的问题。

1.1.4 数据转换

该部分的主要工作是通过对决策表进行属性约减和值约减,实现对数据概念层次的提取,以增强决策规则的代表性。

1.1.5 选择最佳分类模式

主要从两个角度对分类模式进行评价:复杂度和稀疏度。

1.1.6 解释与评估

将数据挖掘的结果与目标进行比对,如果没有达到目标,则回滚前五步。

1.1.7 采取行动

如果得到的分类规则是有效的,则可以将之用于数据挖掘,进行实际问题的解决。

1.2 分类挖掘的准备工作

1.2.1 确定目标

确定数据挖掘的目的,制定挖掘方案和算法。

1.2.2 建立目标数据集

该步骤从要进行数据挖掘的数据库中提取出要进行分析的初始数据,这些数据不仅是孤立的数据,而且包括数据之间的关系及属性,还要进行数据之间条件属性和决策属性的区分。

2 单层感知器的学习算法

2.1 单层感知器工作原理

单层感知器可将外部输入分为两类:当感知器的输出为 +1 时,输入属于前类;当感知器的输出为 -1 时,输入属于后类,从而实现两类目标的识别。

对于只有两个输入的判别边界是直线,选择合适的学习算法可训练出满意的结果,当它用于两类模式的分类时,相当于在高维样本空间中,用一个超平面将两类样本分开。单层感知器模型如图 1 所示。

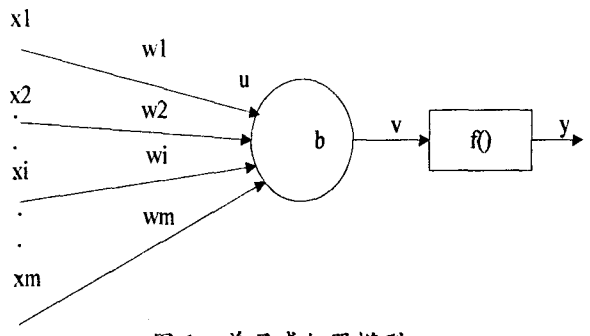

图 1 单层感知器模型

2.2 单层感知器学习算法

基于迭代的思想,通常是采用误差校正学习规则的学习算法;可以将偏差作为神经元突触权值向量的第一个分量加到权值向量中;输入向量和权值向量[11]。

为解决学习速度和分类精度的问题,采取了以下五种方法:

(1)所有节点同时训练;

(2)对各个节点的训练,加以分类正确率为 100% 和错误率为 0 时才停止该节点训练的策略;

(3)采用单样本扰动学习程序策略,即每次单个样本的类别不同,从各类的样本中循环提取样本进行训练;

(4)训练样本的输出与其期望之间允许有误差,且该误差与训练次数成反比,若误差在允许的范围内,则权值不做修正;

(5)对于总误差阈值,若设定过大分类精度将降低,过小会导致训练速度降低,故允许存在一定的误差。

单层感知器学习算法执行流程图如图 2 所示。

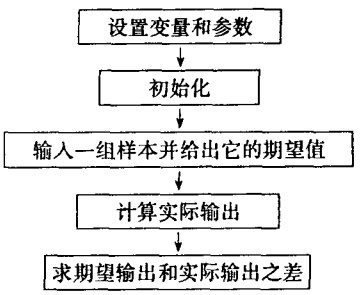

图 2 单层感知器学习算法执行流程图

3 系统程序

开始按钮部分代码:

```
private void startButton_Click(object sender, System.EventArgs e)
    {
    try
    {
    learningRate = Math.Max( 0.00001, Math.Min( 1, double.Parse( learningRateBox.Text ) ) );
    }
    catch
    {
    learningRate = 0.1;
    }
    saveStatisticsToFiles = saveFilesCheck.Checked;
    UpdateSettings( );
    EnableControls( false );
    needToStop = false;
    workerThread = new Thread( new ThreadStart( SearchSolution ) );
    workerThread.Start( );
    }
```

4　系统实现

C# 是一种安全的、稳定的、简单的、优雅的,由 C 和 C++ 衍生出来的面向对象的编程语言[12]。它在继承 C 和 C++ 强大功能的同时去掉了一些它们的复杂特性(例如没有宏和模板,不允许多重继承)。C# 综合了 VB 简单的可视化操作和 C++ 的高运行效率,以其强大的操作能力、优雅的语法风格、创新的语言特性和便捷的面向组件编程的支持成为.NET 开发的首选语言。文中按照 C# 语言面向对象设计的原理,基于 C# 语言进行系统的设计,系统运行界面如图 3 所示。

图 3 为用神经网络单层感知器进行分类运行的界面,操作步骤如下:用 Load 按钮加载数据,系统在 Data 界面中显示三种不同类型的物体。在 Data 界面内系统用不同的颜色表示不同的物体,然后单击开始系统将对三种不同颜色的物体分类,由运行界面可知系统用三条直线将物体分为三类,且使三条直线相交的内部三角形的面积最大。为使系统的分类效果最好,此次分类神经网络系统将学习率设置为 0.1,训练次数运行为 9 次,系统运行误差的动态变化在 Error's dynamics 中给出(图中九个节点分别对应了九次训练次数),由图 3 窗口的曲线动态变化可知,随着迭代次数的增加系统的误差越来越小。神经网络系统的权值和阈值在运行界面中也相应地给出,系统根据权值和阈值的值来求出 Data 界面中对应的三条分类直线。

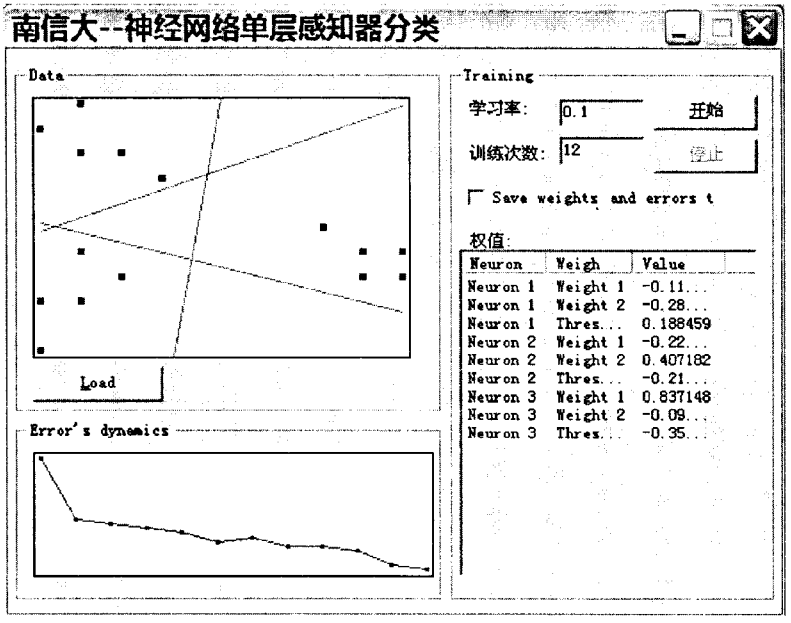

图 3　系统对三种不同的物体进行分类的界面

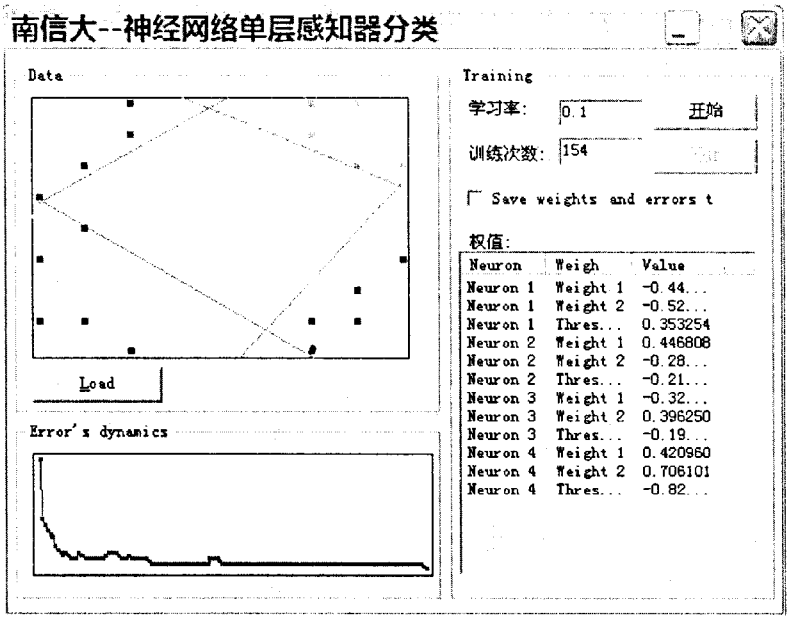

图 4　系统对四种不同的物体进行分类的界面

图 4 为用神经网络单层感知器进行分类运行的界面,操作步骤如下:用 Load 按钮加载数据,系统在 Data 界面中显示四种不同类型的物体。在 Data 界面内系统用不同的颜色表示不同的物体,然后单击开始按钮,系统将对四种不同颜色的物体分类,由运行界面可知系统用四条直线将物体分为四类,且使四条直线相交的内部四角形的面积最大。为使系统的分类效果最

好,此次分类神经网络系统将学习率设置为 0.1,训练次数运行为 74 次,系统运行误差的动态变化在 Error's dynamics 中给出,图中 74 个节点分别对应了 74 次训练次数,由图 4 窗口的曲线的动态变化可知随着迭代次数的增加系统的误差越来越小。神经网络系统的权值和阈值在运行界面中也相应地给出,系统根据权值和阈值的值来求出 Data 界面中对应的用于分类的四条直线。

5 结束语

分类作为数据挖掘的主要内容之一,主要是通过分析训练数据样本,产生关于类别的精确描述。这种类别通常由分类规则组成,可以用来对未来的对象进行分类预测,有着广泛的应用前景。文中通过构造一种单层感知器的神经网络的分类方法,并进行设计分析和实验仿真,对相关技术的研究有一定的借鉴意义。

参考文献:

[1] Pandya A S, Macy R B. Pattern Recognition with Neural Networks in C++ [M]. [s.l.]: CRC Press, 1993: 119 - 125.

[2] 钱卫宁,魏 黎,王 焱,等.一个面向大规模数据库的数据挖掘系统[J].软件学报,2002,13:1540 - 1545.

[3] Berson A, Smith S, Thearling K. Building Data Mining Application for CRM [M]. [s. l.]: McGraw - Hill, 2001: 180 - 230.

[4] 王红军,陈庆新,陈 新,等.基于效用分析的客户聚类方法研究[J].计算机集成制造系统,2003,9(3):237 - 241.

[5] Lee W, Stolfo S J. Data mining approaches for intrusion detection[C]//Proc. of the 7th USENIX Security Symposium. Berkeley, USA: USENIX Assoc, 1998: 79 - 90.

[6] 宋世杰,胡华平,胡笑蕾,等.数据挖掘技术在网络型异常入侵检测系统中的应用[J].计算机应用,2003,23(12):20 - 23.

[7] Wang J C, Huang Y, Wu G S, et al. Web Mining: Knowledge Discovery on the Web Systems, Man, and Cybernetics [C]// IEEE SMC '99 Conference Proceedings. Tokyo: IEEE Computer Society, 1999: 116 - 121.

[8] Schroedl S, Wagstaff K, Rogers S, et al. Mining GPS traces for map refinement[J]. Data Mining and Knowledge Discovery, 2004(9): 59 - 87.

[9] 魏宏业,吕永波,刘志硕.基于数据挖掘的智能交通系统的决策方法研究[J].交通运输系统工程与信息,2003,3(1): 23 - 27.

[10] Pawlak Z. Rough sets[J]. International Journal of Information and Computer Science, 1982, 11: 256 - 341.

[11] 袁曾任.人工神经元网络及其应用[M].北京:清华大学出版社,1999: 66 - 117.

[12] Watson K, Nagel C. C# 入门经典[M].北京:清华大学出版社,2006: 331 - 379.

(上接第 110 页)

了一种子结构发现情形下的近似图匹配算法,该算法具有多项式时间复杂度。实验表明混合变异和个体协同提高了算法的性能,可以获得更优的解;图的近似匹配则在不降低解质量的同时提高了算法的运行效率。不断改进算法的性能以及将文中算法应用于图分类、图聚类等图数据挖掘任务是下一步研究的方向。

参考文献:

[1] Inokuchi A, Washio T, Motoda H. An Apriori - based Algorithm for Mining Frequent Substructures from Graph Data [C]//Proc. of the 4th European Conf. on Principles and Practice of Knowledge Discovery in Databases (PKDD). Lyon, France: Springer, 2000: 13 - 23.

[2] Kuramochi M, Karypis G. Finding Frequent Patterns in A Large Sparse Graph[C]//Proc. of the 2004 SIAM Data Mining Conf.. Lake Buena Vista, Florida, USA: Morgan Kaufmann, 2004.

[3] Yan Xi - feng, Han Jia wei. gSpan:. Graph - based Substructure Pattern Mining[C]//Int. Conf. on Data Mining. Maebashi City, Japan: IEEE Computer Society, 2002: 721 - 724.

[4] Huan J, Wang W, Prins J. Efficient Mining of Frequent Subgraph in the Presence of Isomorphism[C]//Third IEEE International Conference on Data Mining (ICDM 2003). [s.l.]: [s.n.], 2003: 549 - 552.

[5] Grunwald P. A Tutorial Introduction to the Minimum Description Length Principle[EB/OL]. 2009 - 12 - 12. http://www.csee.wvu.edu/natalias/ee568/grunwald04.pdf.

[6] Cook D J, Holder L B. Graph - based data mining[J]. IEEE Intelligent Systems, 2000, 15(2): 32 - 41.

[7] Yoshida K, Motoda H, Indurkhya N. Graph - based induction as a unified learning framework[J]. Journal of Applied Intelligence, 1994(4): 297 - 328.

[8] 常新功,李敏强,寇纪淞. 基于进化算法的图形数据模式发现[J].模式识别与人工智能,2008,21(1): 116 - 121.

[9] 常新功,寇纪淞,李敏强.一种基于混杂 EA 的子结构发现算法[J].系统仿真学报,2008,20(6): 1626 - 1629.

[10] Bandyopadhyay S, Maulik U, Cook D J, et al. Enhancing structure discovery for data mining in graphical databases using evolutionary programming [C]//Proceedings of the Florida Artificial Intelligence Research Symposium. Pensacola Beach, Florida, USA: AAAI Press, 2002: 232 - 236.