人工神经网络算法在数据挖掘中的应用

赵婧宏，潘维民
北京邮电大学计算机科学与技术学院，北京 (100876)
E-mail：grace_jinghong@163.com

摘　要：数据挖掘技术是可从大量数据中，提取隐含在其中的有用的潜在的信息和知识的过程。本文主要介绍了数据挖掘和人工神经网络的相关概念，讨论了在数据挖掘中分别利用 BP 网络和竞争学习网络进行数据分类和聚类的实现过程。同时详细描述了 BP 算法和竞争学习算法，并进行了算法分析和总结了算法的优缺点。

关键词：数据挖掘；人工神经网络；BP 算法；竞争学习算法
中图分类号：TP311

1. 数据挖掘介绍

1.1 数据挖掘的概念

随着数据库技术和信息技术的飞速发展，如何从海量的数据中，提取更直观的信息进行数据分析现状和预测未来，成为推动数据挖掘产生并发展的强大动力。

在 1989 年举行的第 11 届国际联合人工智能学术会议上，首次提出了基于数据库中知识发现(knowledge Discovery in Database, KDD)技术。KDD 是一个综合的过程，包括实验记录，迭代求解与用户交互，以及许多定制要求和决策设计等。直到 1995 年，才在美国计算机年会(Association for Computing Machinery, ACM)上首次提出数据挖掘的概念[1]。

数据挖掘是指从大量的，不完全的，有噪声的，模糊的，随机的数据中，提取隐含在其中的，人们事先不知道的，又是潜在有用信息和知识的过程[1]。它是知识发现的有效手段。发现了的知识不仅可以被用于信息管理、查询优化、决策支持、过程控制等，还可以用于数据自身的维护。因此，数据挖掘是数据库研究中的一个很有应用价值的新领域，它又是一门广义的交叉学科，融合了数据库、人工智能、机器学习、统计学等多个领域的理论和技术。数据挖掘是数据库中知识发现 KDD 过程的一个基本步骤。

1.2 数据挖掘的来源

从原则上讲，数据挖掘可以在任何类型的信息存储上进行[2]。常用的有：关系数据库、事务数据库、数据仓库、高级数据库系统。其中高级数据库系统包括面向对象数据库、对象-关系数据库和面向特殊应用的数据库，如空间数据库、时间序列数据库、文本数据库、多媒体数据库、异种数据库和遗产数据库,WWW。

1.3 数据挖掘过程

从数据库中实现知识的发现，可以通过如下的数据挖掘过程实现：数据挖掘的过程大致有数据清洗、数据集成、数据转换、数据挖掘、模式评估、知识表示等过程，具体介绍如下[2]：

数据清洗(data clearing)：通过填写空缺值，平滑噪声数据，识别，删除孤立点来清除数据噪声与挖掘主题明显无关的数据；

数据集成(data integration)：通过集成多个数据库，数据立方体或文件将来自多数据源中的相关数据组合到一起；

数据转换(data transformation)：将数据转换为易于进行数据挖掘的数据存储形式；

数据挖掘(data mining)：它是知识挖掘的一个基本步骤，其作用就是利用智能方法挖掘数据模式或规律知识；

模式评估(pattern evaluation)：根据一定评估标准从挖掘结果筛选出有意义的模式知识；

知识表示(knowledge presentation)：利用可视化和知识表达技术，向用户展示所挖掘出的相关知识。

数据挖掘过程的结构图如图一所示:

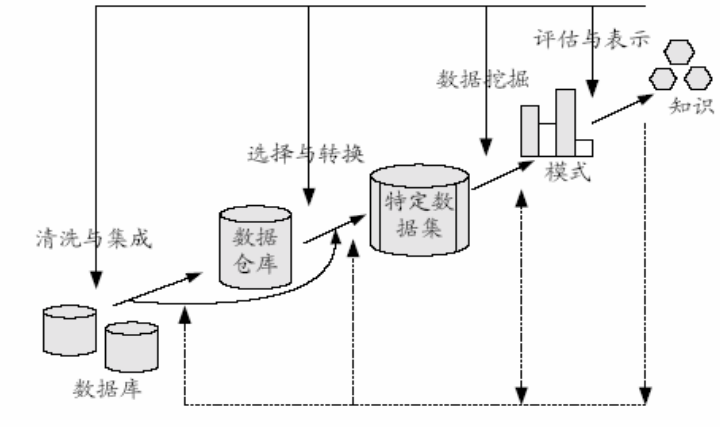

图1　数据挖掘过程

1.4 数据挖掘基本任务

在数据挖掘过程中，数据挖掘是整个挖掘过程的核心内容，其挖掘的主要功能有:分类与预测、聚类分析、孤立点分析、演变分析[2]、总结概括、关联分析等。

分类与预测：有导师的监督学习找出一组能够描述数据集合典型特征的模型，以便能够分类识别未知数据的归属或类别，即将未知事例映射到某种离散类别上。主要表示方法有：分类规则、决策树、数学公式和神经网络。

聚类分析：无导师监督学习方法。根据“各聚集内部数据对象间的相似度最大化，而各聚集对象间相似度最小化”的基本聚类分析原则，将聚类分析的数据对象划分为若干组。每个聚类分析所获得的组就可以视为一个同类别归属的数据对象集合，更进一步从这些同类别数据集，又可以通过分类学习获得相应的分类预测模型（规则）。此外,通过反复不断地对所获得的聚类组进行聚类分析，还可获得初始数据集合的一个层次结构模型。

孤立点分析：对孤立点数据的分析，其中孤立点数据通常为噪声或意外而将其排除在数据挖掘分析处理范围之内的数据。

演变分析：对随时间变化的数据对象的变化规律和趋势进行建模描述。

总结概括：一种典型的描述性任务,找出数据集(子集)的简单描述。

关联分析：从给定的数据集发现频繁出现的项集模式知识，广泛用于市场营销，事务分析等领域。

2. 人工神经网络介绍

目前数据挖掘存在几个方面的问题：

1．数据的量度和维度，面对大量复杂、非线性、时序性与噪音普遍存在的数据；

2．数据分析的目标具有多样性，使其在表述和处理上都涉及到领域知识；

3．在复杂目标下，对海量数据集的分析，目前还没有现成的且满足可计算条件的一般性理论的方法[3]。然而，神经网络在对噪声数据的高承受能力以及对未经训练的数据分类模式的能力方面有很大优势。因此设计出基于神经网络的数据挖掘方法，并将其用于真实世界问题，是可行且也是必要的。

人工神经网络可用于数据挖掘的分类、聚类、特征挖掘、预测和模式识别等方面，因此人工神经网络在数据挖掘中占有举足轻重的作用。下面将从数据挖掘的分类和聚类两个方面对神经网络的前向神经网络和自组织神经网络进行总结介绍。

2.1 人工神经网络

1943 年心理学家 McCulloch 和数学家 Pitts 合作提出了神经元模型理论(简称 MP 模型)，开创了神经科学理论[4]和人工神经网络(Artificial Neural Network 记作 ANN)研究的时代。ANN 是由大量并行分布式处理单元组成的简单处理单元，它有通过调整连接强度而从经验知识进行学习的能力，并可以将这些知识进行运用[5]。

ANN 有强大的计算能力。首先，有着庞大的并行分布式结构；其次，它有学习和由此进行归纳的能力。归纳是指 ANN 对新的输入产生合理的输出，并有其独特的性能：

1．ANN 具有信息处理的并行方式和信息存贮的分布方式；

2．ANN 的“联想记忆”方式使它具有分类和模式识别功能，并具有抗噪声干扰的能力；

3．ANN 具有自学习和自组织的特点，有很强的适应性，经过训练可以识别新的模式；

4．ANN 新的分布式存贮使其具有很强的容错能力，即使少数神经元坏了，也不会导致网络运行失误，具有很好的鲁棒性；

5．ANN 是大量神经元连接的非线性动力学系统。

2.2 人工神经网络学习方式

神经网络有两种不同的学习方法：一种是有导师的学习，也称为监督学习；一种是无导师的学习，也称为自主学习。

有导师的学习：主要是学习过程需要监督，监督的作用通过训练数据本身来完成，也即训练数据不但要包含输入数据，还要包含在特定条件下的期望输出，学习的目的是使网络的实际输出接近于网络的期望输出，这种学习系统分成三个部分：输入部、训练部和输出部。典型的网络为前向多层 ANN(BP 算法等)和 Hopfield ANN。

无导师的学习：主要是学习过程没有明确的外部监督机制，训练数据只包含输入而不包括输出，网络必需根据一定的判断标准进行权值的调整。其典型的网络为：竞争学习网络、自组织特征映射和自适应网络等。

2.3 神经网络分类方法

在 1.4 节中，介绍了数据挖掘分类方法是一种有导师的监督学习的方法，可用的方法有：判定树归纳分类、贝叶斯分类、BP 分类、自关联规则分类等方法。在本部分，主要介绍神经网络的 BP 分类方法。

2.3.1 BP 算法

2.3.1.1 算法简介

反向传播算法人工神经网络(BP人工神经网络)是80年代初发展起来的人工神经网络中最有实用价值的部分之一。早在1969年，感知器的提出者M. Misky 和S. Papert 在他们的Perceptron 专著中指出:简单的线性感知器只能解决线性可分样本的分类问题[6]。简单的线性感知器仅有一层计算单元，而要实现对复杂函数的逼近，必须采用多层前馈网络。于1988年Rumelhart 和Mcclalland 提出了多层前馈网络的反向传播算法(BP算法)，解决了感知器不能解决的多层网络学习算法的问题，其关键是引入了反向传播的误差信号来解决学习问题。

反向传播算法在多层前馈神经网络上学习，这种神经网络的一个例子如图二所示[2]

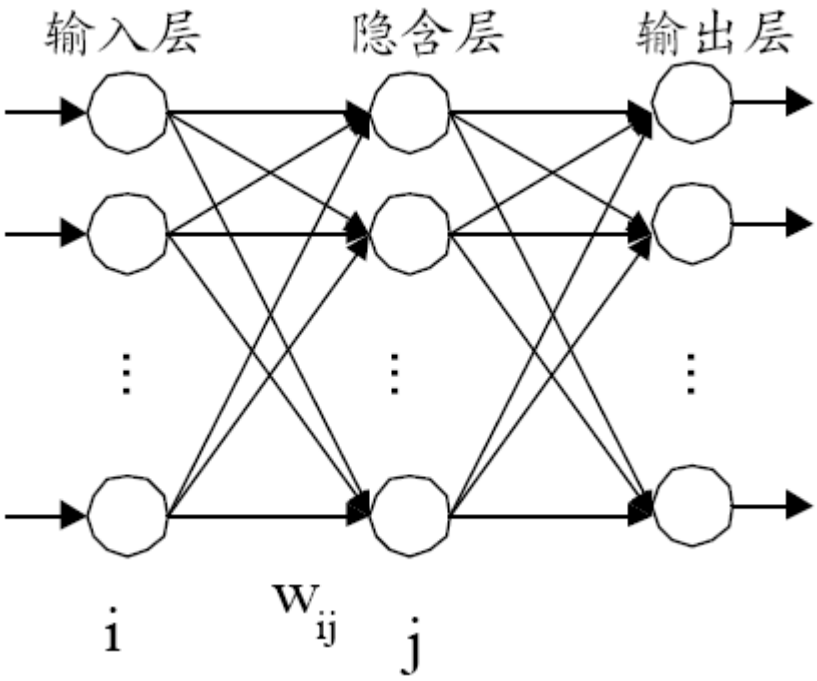

图2 多层前馈神经网络

其中输入对应于对每个训练样本的各属性取值，输入同时赋给第一层(称为输入层)单元，这些单元的输出结合相应的权重，通过赋给第二层(称为隐含层)单元，隐含层的带权输出又作为输入在赋给另一隐含层等等，最后的隐含层结点带权输出赋给输出层单元，该层单元最终给出相应样本的预测输出。

如图二所示的多层神经网络，包含两层输出单元，称为两层神经网络；同样包含两个隐含的神经网络称为三层神经网络，如此等等。该网络是前馈的，即每一个反馈只能发送到前面的输出层或隐含层。它是全连接的，即每一个层中单元均与前面一层的各单元相连接。

给定足够多的隐含单元，线性阀值函数的多层前馈神经网络可以逼近任何函数。

反向传播算法通过不断处理一个训练样本集，将网络处理结果与每个样本实际类别相比较所获误差，来帮助完成学习任务。对于每个训练样本，反向修改其权重，即从输出层开始，通过之后的隐含层，直到最前面的隐含层，通过迭代修改，当权重收敛时学习过程终止，因此它具有误差小、收敛性好、动态性好、结果客观等优势。

2.3.1.2 算法伪代码

神经网络利用反向传播算法学习分类权重，通过逐个更新的方法,迭代进行[2]。

输入：训练样本，samples；学习速率l，一个多层前馈网络network。

输出：一个训练的，对样本分类的神经网络。

方法：

初始化network的权值Wij和偏差$\theta_{j;}$

while不满足训练终止条件{

for samples中的各训练样本X{//正向传播输人

for隐藏层或输出层的每个单元j{

$I_j=\sum_i Wij\ Oi+\theta_j$；//相对于前一层i，计算单元j的净输人

$Oj=1/(1+e^{-Ij})$;//使用对数型的单极性Sigmoid函数将各神经元J的输出映射到区间[0,1]

}

//反向传播误差

for输出层的每个单元j

$Err_j=O_j*(1—O_j)*(T_j—O_j)$; //根据训练样本的已知类标号真实输出Tj，计算神经元j的误差Err_j

for从最后1个到第1个隐含层,对于隐含层的每个单元j

$Err_j=O_j*(1-O_j)*\sum_k(Err_k*W_{jk})$; //根据下一较高层中连接到j的所有神经元的误加权值来计算隐含层神经的误差Err_j

for network中的各权值Wij{/更新权值

$\triangle W_{ij}=(l)\ Err_jO_j$// 权增值

$W_{ij}=W_{ij}+\triangle W_i$ }//权更新

for network中每个偏差θ_j {//更新偏差

$\triangle\theta_j=(l)\ Err_{j;}$//偏差增值

$\theta_j=\theta_j+\triangle\theta_j$//偏差更新

}}}

2.3.1.3 算法描述

1．预处理。在训练之前，将样本的属性值规范到[0，1]范围内，作为输入，如果为离散值，则将其用二进制进行转化。

2．设计网络拓扑结构，理论上讲，隐藏层的神经元数越多，逼近越精确。但实践中，隐藏层神经元数不宜过多，否则会极大加长训练时间，并造成网络容错能力下降。

3．初始化网络的权值和偏差，将其初始化为(0,1)内的随机小数。

4．向前传播输入，根据输入层的输入，权重和偏差，计算出隐含层和输出层的每个单元的净输入和输出。对于隐含层和输出层的输入，为上一层(i)每个连接该单元的输入 Oi 与其对应的权重 W_{ij} 的乘积，以及对应单元偏差之和，然后求和,如图三所示[2]。给定隐含层和输出层的单元 j，到单元 j 的净输入为 I_j

$$I_j=\sum_i Wij\ Oi+\theta_j \quad (1)$$

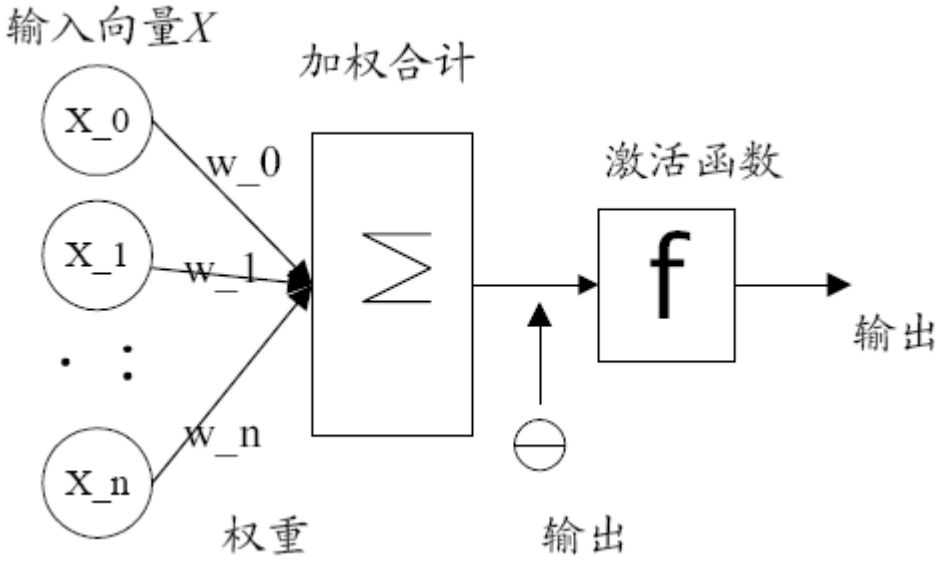

图3　多层前馈神经网络示意图

隐含层和输出层的每个单元取其净输入，然后将一个激活函数作用于它，该函数用符号表现单元代表的神经元活性，使用 logistic 或者 simoid 函数。给定单元 j 的净输入 I_j，则单元的输出 O_j 用下式计算：

$$Oj= 1/ (l +e^{-Ij}) \quad (2)$$

此公式可把较大的输入值,映射到较小的[0,1]。logistic 为非线性的和可微的，使得后向传播算法可以对线性不可分的分类问题建模。

5． 后向传播误差:：通过更新权值和偏差以反应网络预测的误差，向后传播误差.对于输出层单元 j.误差 Err_j

$$Err_j=O_j* (1 — O_j)*(T_j — O_j) \quad (3)$$

其中, O_j 为实际输出，T_j 为已知类标号的正确输出，$O_j* (1 — O_j)$为 logistic 函数的导数.

对于隐含层单元 j 的误差，考虑其下一层中连接 j 的单元的误差加权和。公式如下

$$Err_j = O_j* (1- O_j)* \sum_k(Err_k* W_{jk}) \quad (4)$$

其中，W_{jk} 为下一较高层中单元 k 到单元 j 的连接权重。 Err_k 为单元 k 的误差.

更新权重和偏差，以反映传播的误差，其中$\triangle W_{ij}$为权 W_{ij} 的改变

$$\triangle W_{ij}=(l) Err_jO_j \quad (5)$$

$$W_{ij} = W_{ij}+\triangle W_i \}/ \quad (6)$$

其中，l 为学习率,通常为 0 到 1 之间的一个常数

偏差的更新如下，其中 $\triangle\theta_j$ 为偏差 θ_j 的改变

$$\triangle\theta_j=(l) Err_{j;} \quad (7)$$

$$\theta_j = \theta_j +\triangle\theta_j \quad (8)$$

此算法，每处理一个样本就会更新权重和偏差，此方法为实例更新，权重和偏差的更新也可以积累到变量中，可以在处理完训练集中的所有样本以后再更新权重和偏差，此方法为周期更新。扫描训练集的一次迭代为一个周期。

6．结束迭代，在以下三种情况下，迭代将结束:所有的$\triangle W_{ij}$都小于某个阀值；前一周期不正确的样本输出百分比达到某个阀值；超过预先指定的周期数。

7．算法改进。此算法可通过修改网络拓扑结构，学习率或其他参数的动态调整来变形和替代后向传播。

2.3.1.4 算法优缺点

BP算法的优点：预测精度总的来说较高、健壮性好、训练样本中包含错误时也可以正常工作；输出可能是离散值，连续值或为离散或量化属性的向量值；对目标进行快速分类等优点，所以BP 算法很实用,在工业控制如DC - DC 变换器的智能控制、语音识别、图像处理如手写体识别和图像压缩等方面已成功获得应用[6]。

BP 算法的缺点：训练学习时间长；蕴含在学习的权中的符号含义比较难理解；很难与专业领域知识相整合；属于非线性优化法,存在局部最小值问题；收敛速度慢,通常需要几千步迭代；网络运行只是单向传播,无反馈；网络的隐节点数无理论上指导,凭经验选取；对新加入的样本要影响到已经学完的样本、刻划每个输入的特征数目必须相同。

2.4 神经网络聚类方法

2.4.1 聚类分析

将物理的或抽象的对象的集合分组成为由类似的对象组成的多个类的过程被称为聚类[2]。一个聚类就是由彼此相似的一组对象所构成的集合；不同聚类中对象是不相似的。聚类分析为无导师的学习方法。其主要算法有：基于划分的聚类，基于层次的聚类，基于密度的聚类，基于网格的聚类以及基于模型的聚类，其中这部分主要介绍基于模型的聚类中的神经网络方法。

2.4.2 神经网络方法

神经网络的聚类有三个比较重要的方法：一个是竞争学习，一个是自组织特征映射，一个是自适应谐振网。在本部分，主要讨论竞争学习，涉及到竞争的神经单元。

2.4.2.1 竞争学习算法介绍

竞争神经网络属于一种循环网络，是以无指导学习算法为基础的。在竞争学习中，神经网络的输出神经在它们自己当中竞争以便被激活。在竞争学习中，无论何时都只有一个输出神经被激活。竞争学习的神经网络的标准技术，必需满足如下三个基本元素[5]：

1. 具有相同结构，且与初始随机选择的权重连接的一组神经。因此，神经可以不同的响应一组被给定的输入样本；
2. 决定每根神经强度的极限值；
3. 允许神经争取响应一组给定的输入子集权利的机制，这样每次只有一个输出神经被激活，赢得竞争的神经被称为胜者全获神经。

简单的竞争学习形式中，人工神经网络具有一个输出神经单层，每个单层完全和输入节点相连，网络可以包括神经中的反馈连接。如图四所示，在此描述的网络结构中，反馈连接执行侧面抑制，每根神经倾向该神经至它被侧面连接处。

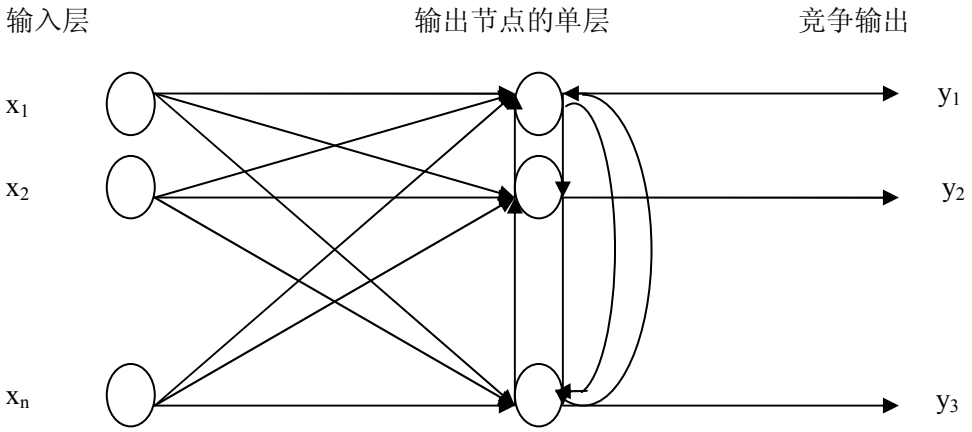

图4　简单竞争性网络结构图

对于输入样本 $X=\{x_1,x_2,\ldots x_n\}$，其网络值 net_k 为网络所有神经中最大的，则神经 k 为获胜神经。可以把获胜神经 k 输出信号设为 1，其他神经输出信号设为 0,可以记为：

$$y_k=\begin{cases}1 & net_k>net_j \text{ 对所有的 } j,j<>k\\ 0 & \text{否则}\end{cases}$$

设 W_{kj} 表示输入节点 j 连接神经 k 的突触权重，于是，神经通过将突触权重的非活动输入节点移动到它的活动输入节点进行学习。如果一个神经赢得竞争，该神经的每个输入节点应放弃一定比例的突触权重，于是，被放弃的权重被分配到活动的输入节点中，运用突触权重 W_{kj} 替代值$\triangle W_{kj}$ 被定义为：

$$\triangle W_{kj}=\begin{cases}\eta(x_j-W_{kj}) & \text{如果神经 k 赢得竞争}\\ 0 & \text{如果神经 k 未赢得竞争}\end{cases}$$

则竞争后的 $W_{kj}= W_{kj}+\triangle W_{kj}$

在此, η 为学习率参数.规则的全部作用是将取胜的神经的突触权重移向输入模式 X。在竞争学习过程中，类似样本通过网络进行分组，并用输出上的单一人工神经网络描述。基于数据相关性的分组被自动完成。然而，此项功能以一种稳定的方式执行，输入样本必须分成十分明显的组，否则，网络可能不稳定。

2.4.2.2 竞争学习算法优缺点

在聚类结束时，每个聚类能够被认为是一个新的特征，它可以检测出对象中的规律。因此所获得的聚类可以看成是从低层特征到高层特征的一个映射。

神经网络聚类方法与脑处理具有较强的理论联系,但其需要较长的处理时间和数据的复杂性，因此要适合大型数据库的应用，还需要进行更多的研究。

总的说来，竞争学习可以很好的用于数据的聚类分析，但它也存在一些局限性：

1．学习率 η 的选择，必须在学习速度和最终权重因子稳定性之间作出选择；

2．竞争学习的稳定性问题还可能在神经的初始权重向量的定位不准而未赢得竞争，因此也未学习的情况下产生。

3．竞争学习过程总是有和它的输出神经一样多的类，这不适用于一些应用，特别是当

类的数量未知或如果很难事先估计时。

3．总结

通过对人工神经网络中的 BP 算法和竞争网络算法进行详细地介绍和分析，可以实现数据挖掘的分类和聚类两个处理过程。本文的新颖之处在于总结了神经网络算法中的 BP 算法和竞争学习算法在数据挖掘中的应用，并提出了两种算法中存在的优缺点和算法改进的方向，希望可以通过将算法运用到实际的分类聚类中，验证算法的正确性和可行性。

参考文献

[1] 苏新宁等。《数据仓库和数据挖掘》。[J]，2006，清华大学出版社：115－134
[2] Jiawei Han Micheline Kamber 著，范明 孟晓峰等译，《数据挖掘概念与技术》，[J]，2001， 机械工业出版社 ；
[3] 冯欢欣等，《基于神经网络计算的数据挖掘方法研究》， 大众科技，2006 年第 8 期；
[4] 张青贵，《人工神经网络导论》，[J]，2004，中国水利水电出版社；
[5] Mehmed Kantardzic 著，闪四清等译，《数据挖掘-概念、模型、方法和算法》，[J]，2001，清华大学出版社；
[6] 单潮龙等，《BP 人工神经网络的应用及其实现技术》，海军工程大学学报，2000 年第 4 期；
[7] 邵峰晶等，《数据挖掘原理与算法》。[J]，2003，中国水利水电出版社；
[8] 朱大奇等，《人工神经网络原理及应用》，[J]，2006，科学出版社；
[9] 高峰等，《基于神经网络的数据挖掘综述》， 信息与控制，第 28 卷增刊 1999 年 8 月；
[10] 张德锋等，《数据挖掘技术》航空计算技术 第 35 卷 第 3 期 2005 年 9 月；

The application of Artificial Neural Networks in Data Mining

Zhao Jinghong，Pan Weimin
Computer Application Technology centrol, Computer Science & Technology school，Beijing University of Posts and Telecommunications，Beijing (100876)

Abstract
Data mining is the process of finding useful information and knowledge in large relational database. In this paper, we present the concepts of Data mining, Artificial Neural Networks (ANN), *Back Propagation network algorithm* and *Competitive learning algorithm*. Then we research the later two *algorithms* in details and their applications in classification and clustering realms which are two parts of Data mining. Our contribute is indicating the advantages and disadvantages about these *algorithms* and finding a way to resolve it.We hope to continue our research about applying our idea to the practice problem of classification and clustering and achieve an ideal result.

Keywords：Data mining，Artificial Neural Networks，Back Propagation network algorithm，Competitive learning algorithm

作者简介：
赵婧宏，女，硕士研究生，主要研究方向为数据仓库与数据挖掘；
潘维民，男，博士，副教授，研究生导师，主要研究领域为计算逻辑学，数据仓库与数据挖掘技术，分析型应用系统技术，金融工程研究。