**机器学习与数据挖掘报告**

**基于UCI数据集的鸢尾属植物分类**

**学院：**

**班级：**

**学号：**

**姓名：**

**日期：2016年1月4日**

目录

[一、 问题描述 3](#_Toc440537931)

[二、 实验环境 3](#_Toc440537932)

[三、 算法综述 4](#_Toc440537933)

[1. VSM算法 4](#_Toc440537934)

[2. 决策树 4](#_Toc440537935)

[3. 模式匹配 5](#_Toc440537936)

[四、 实验原理 6](#_Toc440537937)

[1. Bp神经网络算法 6](#_Toc440537938)

[2. UCI数据集 8](#_Toc440537939)

[五、 实验过程与结果分析 9](#_Toc440537940)

[1. 实验过程 9](#_Toc440537941)

[2. 结果分析 10](#_Toc440537942)

[六、 总结讨论 12](#_Toc440537943)

[七、 参考文献 12](#_Toc440537944)

**摘要**

分类是数据挖掘中的一个重要课题。分类的目的是学会一个分类函数或分类模型（也常常称作分类器），该模型能把数据库中的数据项映射到给定类别中的某一个。分类可用于提取描述重要数据类的模型或预测未来的数据趋势。如今分类应用在生活生产的方方面面。本文使用的数据集是UCI数据集的Iris数据集， 通过BP神经网络算法进行分类，测试BP神经网络的分类性能。

**关键字：**数据挖掘、分类算法、BP神经网络算法、Iris数据集

## 问题描述

分类可描述如下：输入数据，或称训练集（training set）是一条条记录组成的。每一条记录包含若干条属性（attribute），组成一个特征向量。训练集的每条记录还有一个特定的类标签（类标签）与之对应。该类标签是系统的输入，通常是以往的一些经验数据。一个具体样本的形式可为样本向量：(v1,v2,…,…vn:c)。在这里vi表示字段值，c表示类别。

分类的目的是：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。这种描述常常用谓词表示。由此生成的类描述用来对未来的测试数据进行分类。尽管这些未来的测试数据的类标签是未知的，我们仍可以由此预测这些新数据所属的类。注意是预测，而不能肯定。我们也可以由此对数据中的每一个类有更好的理解。也就是说：我们获得了对这个类的知识。

分类技术在很多领域都有应用，例如可以通过客户分类构造一个分类模型来对银行贷款进行风险评估；当前的市场营销中很重要的一个特点是强调客户细分。客户类别分析的功能也在于此，采用数据挖掘中的分类技术，可以将客户分成不同的类别，比如呼叫中心设计时可以分为：呼叫频繁的客户、偶然大量呼叫的客户、稳定呼叫的客户、其他，帮助呼叫中心寻找出这些不同种类客户之间的特征，这样的分类模型可以让用户了解不同行为类别客户的分布特征；其他分类应用如文献检索和搜索引擎中的自动文本分类技术；安全领域有基于分类技术的入侵检测等等。

## 实验环境

一台win8系统的台式机，Matlab R2015， UCI数据集。

## 算法综述

### VSM算法

VSM法即向量空间模型(Vector Space Model)法，其基本思想是将文档表示为加权的特征向量：D=D(T1，W1；T2，W2；…；Tn，Wn)，然后通过计算文本相似度的方法来确定待分样本的类别。当文本被表示为空间向量模型的时候，文本的相似度就可以借助特征向量之间的内积来表示。

在实际应用中，VSM法一般事先依据语料库中的训练样本和分类体系建立类别向量空间。当需要对一篇待分样本进行分类的时候，只需要计算待分样本和每一个类别向量的相似度即内积，然后选取相似度最大的类别作为该待分样本所对应的类别。

由于VSM法中需要事先计算类别的空间向量，而该空间向量的建立又很大程度的依赖于该类别向量中所包含的特征项。根据研究发现，类别中所包含的非零特征项越多，其包含的每个特征项对于类别的表达能力越弱。因此，VSM法相对其他分类方法而言，更适合于专业文献的分类。

### 决策树

决策树也称为判定树，是一种有监督的学习方法。决策树代表着决策树的树形结构，可以根据训练集数据构造出决策树。如果该树不能对所有对象给出正确的分类，就选择一些例外加入到训练集数据中。重复该过程，直到形成正确的决策集。决策树方法首先对数据进行处理，利用归纳算法生成可读的规则和决策树，然后使用决策树对新数据进行分析，本质上是通过一系列规则对数据进行分类的过程。决策树的典型算法有ID3，C4.5，CART等，本次采用ID3测试集的方法，将样本集合分为训练集与测试集。根据训练集构建决策树，决策树中的结点逐层展开。每展开一层子结点，并将其设为叶结点，就得到一棵决策树，然后采用测试集对所得决策树的分类性能进行统计。重复上述过程，可以得到决策树在测试集上的学习曲线。根据学习曲线，选择在测试集上性能最佳的决策树为最终的决策树。与决策树相关得到主要算法有：CLS，ID3，C4.5，CART。

CLS基本思想：从一棵空决策树开始，选择某一属性（分类属性）作为测试属性。该测试属性对应决策树中的决策结点。根据该属性的值的不同，可将训练样本分成相应的子集，如果该子集为空，或该子集中的样本属于同一个类，则该子集为叶结点，否则该子集对应于决策树的内部结点，即测试结点，需要选择一个新的分类属性对该子集进行划分，直到所有的子集都为空或者属于同一类。

### 模式匹配

模式识别又常称作模式分类，常用的方法有：决策理论方法和句法方法。

1. **决策理论方法**

又称统计方法，是发展较早也比较成熟的一种方法。被识别对象首先数字化，变换为适于计算机处理的数字信息。一个模式常常要用很大的信息量来表示。许多模式识别系统在数字化环节之后还进行预处理，用于除去混入的干扰信息并减少某些变形和失真。随后是进行特征抽取，即从数字化后或预处理后的输入模式中抽取一组特征。特征抽取过程将输入模式从对象空间映射到特征空间。这时，模式可用特征空间中的一个点或一个特征矢量表示。这种映射不仅压缩了信息量，而且易于分类。在决策理论方法中，特征抽取占有重要的地位，但尚无通用的理论指导，只能通过分析具体识别对象决定选取何种特征。特征抽取后可进行分类，即从特征空间再映射到决策空间。为此而引入鉴别函数，由特征矢量计算出相应于各类别的鉴别函数值，通过鉴别函数值的比较实行分类。

1. **句法方法**

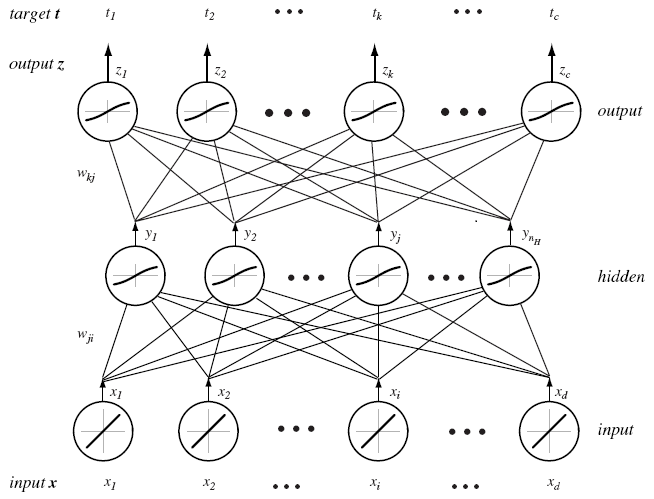
又称结构方法或语言学方法。其基本思想是把一个模式描述为较简单的子模式的组合，子模式又可描述为更简单的子模式的组合，最终得到一个树形的结构描述，在底层的最简单的子模式称为模式基元。在句法方法中选取基元的问题相当于在决策理论方法中选取特征的问题。通常要求所选的基元能对模式提供一个紧凑的反映其结构关系的描述，又要易于用非句法方法加以抽取。显然，基元本身不应该含有重要的结构信息。模式以一组基元和它们的组合关系来描述，称为模式描述语句，这相当于在语言中，句子和短语用词组合，词用字符组合一样。基元组合成模式的规则，由所谓语法来指定。一旦基元被鉴别，识别过程可通过句法分析进行，即分析给定的模式语句是否符合指定的语法，满足某类语法的即被分入该类。

这两种方法不能截然分开，在句法方法中，基元本身就是用决策理论方法抽取的。在应用中，将这两种方法结合起来分别施加于不同的层次，常能收到较好的效果。

## 实验原理

### Bp神经网络算法

BP（Back Propagation）是一种按误差逆传播算法训练的多层前馈网络。BP网络能学习和存贮大量的输入-输出模式映射关系，而无需事前揭示描述这种映射关系的数学方程。它的学习规则是使用最速下降法，通过反向传播来不断调整网络的权值和阈值，使网络的误差平方和最小。BP神经网络模型拓扑结构包括输入层（input）、隐层(hide layer)和输出层(output layer)。



输入层：单元i的输入： xi 单元数量：d

单元i的输出： xi

单元i的激活函数：线性函数

隐 层：单元j的净输入：netj 单元数量： nH

单元j的输出：单元j的激活函数：非线性函数输出层：单元k的净输入： 单元数量：c

单元k的输出：单元k的激活函数：非线性函数

1. 权值的调整

输入一个样本(x,t)，计算网络的输出z，根据z与t的差距调整所有的权系数，使z与t尽可能接近，即使近可能小。

权系数的调整：

1. 激活函数

激活函数f应满足如下基本条件: 非线性、有界、连续、可微，用的激活函数：Sigmoid函数

f(net) = 1/(1+e-net)

(1-e-bnet)/(1+e-bnet)

在训练过程开始前，权系数通常采用随机化方法进行随机赋值，且初始权值不宜过大。典型的方法是：对于每一个权系数，按照均匀分布在区间内随机选择一个数值进行赋值。对于每一个权系数，按照均匀分布在区间内随机选择一个数值进行赋值。

1. 学习速率

学习速率 直接影响权系数调整时的步长。学习速率过小，导致算法收敛速度缓慢。学习速率过大，导致算法不收敛。学习速率的典型取值 。另外学习速率可变。误差函数的局部极小值调整权系数的目标是使误差函数取得最小值。但是，采用梯度下降法(Gradient Descent Procedure)不能保证获得最小值，而只能保证得到一个极小值。如果训练过程无法使误差函数降低到预期的程度，一种常用的方法是：再一次对权系数进行随机初始化，并重新训练网络。

1. 学习曲线

样本集的划分：一般情况下，可把已知的样本集划分为三个子集：练集、确认集、测试集。本次实验只划分为两个子集，即训练集、测试集。

训练集：用来调整权系数，使误差函数尽可能变小。

确认集：用来初步验证神经网络对未来新样本的分类能力，并据此确定最佳的权系数。神经网络的训练过程需要采用训练集及确认集共同完成。

测试集：在训练过程最终结束后，再采用测试集对网络的分类性能进行最后测试，用以评估神经网络在实际应用中的实际性能。

### UCI数据集

UCI数据库是加州大学欧文分校(University of CaliforniaIrvine)提出的用于机器学习的数据库，这个数据库目前共有335个数据集，其数目还在不断增加，UCI数据集是一个常用的标准测试数据集。

UCI数据可以使用matlab的dlmread（或textread或者利用matlab的导入数据）读取，不过，需要先将不是数字的类别用数字，比如1/2/3等替换，否则读入不了数值，当字符了。每个数据文件（\*.data）包含以“属性-值”对形式描述的很多个体样本的记录。作为数据集和领域知识的补充，在utilities目录里包含了一些在使用这一数据集时的有用资料。

本次实验中通过对UCI中的Iris数据进行分类测试的。Iris data set，也称鸢尾花卉数据集，是一类多重变量分析的数据集。其数据集包含了150个样本，都属于鸢尾属下的三个亚属，分别是山鸢尾 (Iris setosa)，变色鸢尾(Iris versicolor)和维吉尼亚鸢尾(Iris virginica)。四个特征被用作样本的定量分析，分别是花萼和花瓣的长度和宽度。实验中所用的数据集已经分为三类，第一组为山鸢尾数据，第二组为变色鸢尾数据，第三组为维吉尼亚鸢尾数据。

## 实验过程与结果分析

### 实验过程

1. 将下载的iris数据集转成（.csv）格式的数据文件。
2. 读取Iris数据文件，获取训练样本。

M=csvread('Iris.csv');

M11=M(1:20,2);M12=M(51:70,2);M13=M(101:120,2);

M21=M(1:20,4);M22=M(51:70,4);M23=M(101:120,4);

M1=[M11' M12' M13']';

M2=[M21' M22' M23']';

SI=[M1 M2]';

[xx,SN]=size(SI);

1. 设置神经网络的训练参数。

HUN=20;%隐含层节点个数

Maxsteps=20000;%最大学习次数

lrate=0.1;%学习速率

E0=0.01;%目标误差

1. 创建网络，训练数据

HO=logsig(W1Ex\*SIE);

HOE=[HO' ones(SN,1)]';

NetworkO=logsig(W2Ex\*HOE);

Error=SO-NetworkO;

SSE=sumsqr(Error)

ErrRecord=[ErrRecord SSE];

if SSE<E0,break,end

Delta2=2\*lrate\*Error.\*NetworkO.\*(1-NetworkO);

Delta1=W2'\*Delta2.\*HO.\*(1-HO);

dW2Ex=Delta2\*HOE';

dW1Ex=Delta1\*SIE';

%采用动量BP学习算法

if i>1

beta=0.5;%动量参数

dW2Ex=beta\*LdW2Ex+(1-beta)\*dW2Ex;

dW1Ex=beta\*LdW1Ex+(1-beta)\*dW1Ex;

LdW2Ex=dW2Ex;

LdW1Ex=dW1Ex;

else

LdW1Ex=dW1Ex;

LdW2Ex=dW2Ex;

end

W1Ex=W1Ex+dW1Ex;

W2Ex=W2Ex+dW2Ex;

W2=W2Ex(:,1:HUN);

end

1. 获取测试数据并测试

%获取测试样本

TM11=M(21:50,2);TM12=M(71:100,2);TM13=M(121:150,2);

TM21=M(21:50,4);TM22=M(71:100,4);TM23=M(121:150,4);

TM1=[TM11' TM12' TM13']';

TM2=[TM21' TM22' TM23']';

TestSI=[TM1 TM2]';

[xx,TestSN]=size(TestSI);

%测试

TestHO=logsig(W1\*TestSI+repmat(B1,1,TestSN));

TestNetworkO=logsig(W2\*TestHO+repmat(B2,1,TestSN));

[Val,NNClass]=max(TestNetworkO);

### 结果分析

取60个训练样本与60个测试样本分类结果如下：

Test\_C1\_C1 = 20

Test\_C1\_C2 = 0

Test\_C1\_C3 = 0

Test\_C2\_C1 = 0

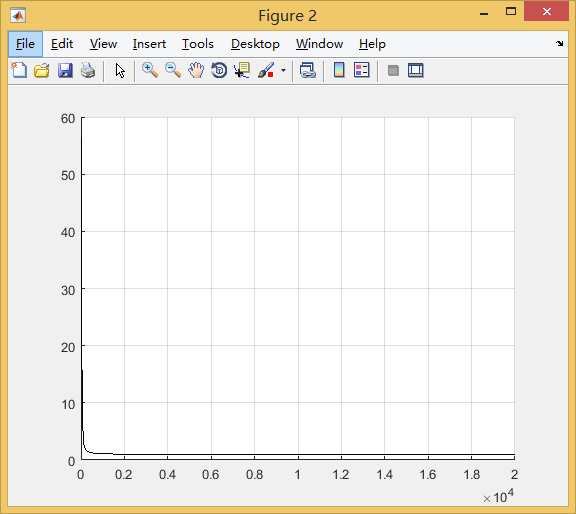
Test\_C2\_C2 = 18

Test\_C2\_C3 = 3

Test\_C3\_C1 = 0

Test\_C3\_C2 = 2

Test\_C3\_C3 = 17



1. 60样本的学习误差曲线

Test\_x\_y其中x表示样本的类别，y表示被分类的结果。有上面的数据可得分类的准确率为91.67%。

取90个训练样本与60个测试得到的分类结果如下：

Test\_C1\_C1 = 20

Test\_C1\_C2 = 0

Test\_C1\_C3 = 0

Test\_C2\_C1 = 0

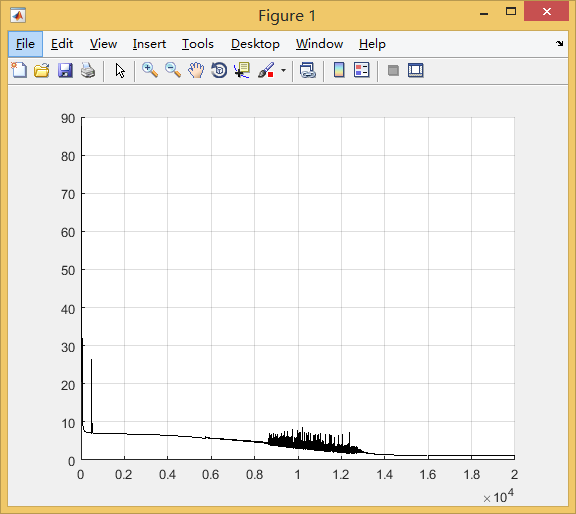
Test\_C2\_C2 = 19

Test\_C2\_C3 = 1

Test\_C3\_C1 = 0

Test\_C3\_C2 = 1

Test\_C3\_C3 = 19



1. 90样本的学习误差曲线

有上面的数据可得分类的准确率为96.67%。

通过两组数据的对比，可以直观的看出，在其他条件不变，单独增加测试的样本的数量的时候，可以提高神经网络BP算法分类的准确率。而且神经网络在数据挖掘分类方面的准确率很高。

## 总结讨论

由实验结果可知，神经网络的BP算法的预测分类的准确性很高，当样本的总数增加时，或者学习速率增大时，分类的准确性也会增大，这是因为当学习速率增大时，会加快算法的收敛速度。

BP算法的优点：预测精度较高、健壮性好、训练样本中包含错误时也可以正常工作；进行快速分类等优点，所以BP 算法很实用,在工业控制如语音识别、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。 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；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。

除了神经网络，还有支持向量机，决策树，贝叶斯等算法都有可以和好的应用在数据挖掘的分类上。每个的算法的优势都不相同。因此，实际的应用中，对于不同类型的问题，实际操作时应结合实际问题综合考虑，而不是单一的使用哪种方式。

经过本次的实验过程，在查阅了大量关于数据挖掘分类计数的文献后，对于数据挖掘这门技术有了更加深刻的认识与理解，明白了数据挖掘的过程以及每个过程中使用的算法。学习到了许多的知识，锻炼了自己，也充实了自己。最后，感谢老师的悉心教导。

## 参考文献

1. 朱凯. 精通MATLAB神经网络. 北京：电子工业出版社，2010.
2. 张德丰. MATLAB神经网络应用设计. 北京：机械工业出版社，2009
3. 李雄飞 李军. 数据挖掘与知识发现. 北京：高等教育出版社，2003。
4. 毛国君, 石云. 数据挖掘原理与算法. 北京：清华大学出版社，2005.
5. 苏新宁等. 数据仓库和数据挖掘.北京：清华大学出版社，2006
6. Jiawei Han Micheline Kamber著. 范明 孟晓峰等译. 数据挖掘概念与技术. 北京： 机械工业出版社， 2001.
7. 张青贵. 人工神经网络导论. 北京：中国水利水电出版社，2004.
8. Mehmed Kantardzic著.闪四清等译. 数据挖掘-概念、模型、方法和算法. 北京：清华大学出版社，2001.
9. 单潮龙. BP人工神经网络的应用及其实现技术. 武汉：海军工程大学学报，2000年第4期.