数据挖掘分类算法综述

姓名： 李涛

专业： 软件工程

指导教师： 吴萍萍

字数:

**摘要**

随着数据库运用的日益发展，数据挖掘技术越来越多的被运用到了商务，医学，科学与工程等方面。在数据挖掘技术的众多分支中，分类算法是一个重要的方面。准确率、速度、强壮性、可伸缩性、可解释性是评估分类算法的五条标准，其中准确率最为重要。分类算法大体可分为传统分类算法和基于软计算的分类法两类。在分类算法中常用的算法有：决策树分类、神经网络分类、支持向量机分类、朴素贝叶斯分类等。本文将对数据挖掘中的分类算法以及分类算法中常用的算法做一个概述。

**关键字：**数据挖掘，分类算法，算法

**Abstract**

With the increasing development of database application, data mining technology has been applied to more and more business, medicine, science and engineering etc.. The classification is an important research branch in the data mining domain and so on. The rate of accuracy, speed, robustness, scalability and interpretability is to assess the five criteria of classification algorithm, the accuracy rate of one of the most important. The classification algorithms can be divided into traditional classification algorithm based on Soft Computing and classification of two types of. In the categorization algorithm，Commonly used algorithm: decision tree , neural network, support vector machine, naive Bayes and so on. This paper will give an overview of the classification algorithm in data mining and algorithms commonly used in classification algorithm.

**Keywords:** Data mining, Classification algorithm, Algorithm

目录

[一、绪论 5](#_Toc413680945)

[二、分类算法 5](#_Toc413680946)

[（一）简介 5](#_Toc413680947)

[（二）分类过程 5](#_Toc413680948)

[（四）分类目的 6](#_Toc413680949)

[（五）分类数据的预处理 7](#_Toc413680950)

[1、数据清理 7](#_Toc413680951)

[2、数据集成 7](#_Toc413680952)

[3、数据变换 7](#_Toc413680953)

[4、数据规约 8](#_Toc413680954)

[（六）分类算法常用模型 8](#_Toc413680955)

[（七） 分类算法常用模型优缺点对比 8](#_Toc413680956)

[1 贝叶斯分类器 8](#_Toc413680957)

[2 决策树分类器 8](#_Toc413680958)

[3 神经网络 8](#_Toc413680959)

[4 支持向量机 9](#_Toc413680960)

[5 K-最近邻 9](#_Toc413680961)

[（八）分类算法运用领域 9](#_Toc413680962)

[三、决策树分类 9](#_Toc413680963)

[（一）决策树概念 9](#_Toc413680964)

[（二）ID3算法 10](#_Toc413680965)

[（三）C4.5算法 10](#_Toc413680966)

[（四）SLIQ算法 11](#_Toc413680967)

[（五）SPRINT算法 12](#_Toc413680968)

[四、朴素贝叶斯分类 12](#_Toc413680969)

[（一）贝叶斯定理 12](#_Toc413680970)

[（二）朴素贝叶斯分类器 13](#_Toc413680971)

[（三）朴素贝叶斯分类器特点 13](#_Toc413680972)

[（四）朴素贝叶斯分类算法 14](#_Toc413680973)

[五、支持向量机分类 14](#_Toc413680974)

[（一）简介 14](#_Toc413680975)

[（二）logistic回归 14](#_Toc413680976)

[（三）形式化表示 16](#_Toc413680977)

[（四）函数间隔（functional margin）和几何间隔（geometric margin） 17](#_Toc413680978)

[（五）支持向量机的优缺点 19](#_Toc413680979)

[六、 人工神经网络分类 19](#_Toc413680980)

[（一）简介 19](#_Toc413680981)

[（三）主要解决问题 20](#_Toc413680982)

[（四）人工神经网络的特点 20](#_Toc413680983)

[（五）对神经网络分类器进行评估 20](#_Toc413680984)

[七、结语 20](#_Toc413680985)

# 一、绪论

数据挖掘是在大型数据存储库中获取有用知识和价值的过程。数据挖掘技术用来探察大型数据库，发现先前未知的有用模式。数据挖掘是一个交叉学科领域，受多个学科的影响，包括数据库系统、统计学、机器学习、可视化和信息科学等。数据挖掘实质上是一种决策支持过程，其主要技术手段是统计方法，包括数理统计方法、多元统计方法、计量经济学和时间序列分析方法等。数据挖掘基础是统计学与人工智能。人工智能是以自动机为手段，通过模拟人类宏观外显的思维行为，从而高效率地解决现实问题的技术。

分类算法是数据挖掘技术中一个重要组成部分。分类任务就是通过学习得到一个目标函数，该目标函数能把数据库中的数据项映射到给定类别中的某一个。

目标函数也称为分类模型，分类模型可运用于描述性建模和预测性建模。分类技术是一种根据输入数据建立分类模型的系统方法。分类算法包括决策树分类法、基于规则的分类法，神经网络，支持向量机和朴素贝叶斯算法等。这些技术都使用了一种学习算法来确定分类模型 ，这样的模型能够很好地拟合输入数据中类标号和属性之间的联系。本文将对数分类算法中常用的决策树分类、朴素贝叶斯分类、人工神经网络分类、支持向量机分类做一个概述。

# 二、分类算法

## （一）简介

分类是事先定义好类别 ，类别数不变 。分类器需要由人工标注的分类训练语料训练得到，属于有指导学习范畴。

## （二）分类过程

1、建立一个模型，描述预定数据集和概念集，如图1所示。

通过对数据库中各数据行内容的分析而获得该模型。假定每个元组属于一个预定义的类，由一个类标号属性确定，其类别值是由一个属性描述。分类学习方法所使用的数据集叫做训练样本集合，因此分类学习也可叫做有指导学习(learning by example)。它是在已知训练样本类别情况下，通过学习建立相应的模型，无指导学习则是在训练样本的类别与类别个数均未知的情况下进行的。一般情况下分类学习所获得的模型可以表示为分类规则形式、决策树形式或数学公式形式。

训练数据集：由为建立模型而被分析的数据元组形成

训练样本：训练数据集中的单个样本（元组）

学习模型可以用分类规则、判定树或数学公式的形式提供



图1 数据分类过程中的学习建模

2、使用模型，对将来的或位置的对象进行分类，如图2，利用学习获得的分类规则。对已知测试数据进行模型准确率评估，以及对未知类别的新数据进行分类预测。

首先评估模型的预测准确性率，例如使用保持(holdout)方法。然后对每个测试样本，将已知的类标号和该样本的学习模型类预测比较，模型在给定测试集上的准确率是正确被模型分类的测试样本的百分比，测试集要独立于训练样本集，否则会出现“过分适应数据”的情况。



图2 数据分类过程中的分类测试

## （四）分类目的

1、描述性建模，分析输入的数据，通过在训练集中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，并将隐含函数表示出来。

2、预测性建模，通过描述性建模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。

## （五）分类数据的预处理

按照不同的源数据来进行预处理有4个基本功能，分别是：数据清理，数据集成，数据变换，数据规约

### 1、数据清理

数据清理主要目的是去除数据源集中的噪声和无关的数据信息，处理缺失的数据和清洗脏数据，空缺值等。

1. 噪音：噪声是一个测量变量中的随即错误和偏差。我们对噪音数据的操作是发现重复数据、查找错误的属性值以及数据平滑。
2. 空缺值的处理：在处理缺失数据时，因为在训练阶段和分类过程中，缺失数据值会导致一些问题，训练数据中的缺失值会会导致结果不准确，因此必须进行处理操作。最常用的方法是使用确定的值填充空缺值，例如用一个全局常量替换空缺值，使用属性的平均值填充空缺值或着按照某些属性将所有的元组分类，然后用同一类中属性的平均值填充空缺值。
3. 清洗脏数据：不同结构数据源数据库中的数据不一定都是正确的，也可能存在着不完整，不一致，不精确和重复的数据，这样的数据我们叫做“脏数据”。脏数据能使数据挖掘过程陷入混乱，可能使输出结果不可靠。

对于脏数据我们可以采用以下方法来清理：

a)手工清理；

b)编写特定的清理应用程序来清理；

c)利用概率统计学远离查找数值异常的记录；

d)检测和删除重复的记录；

### 2、数据集成

在数据集成时候，许多属性可能与分类任务不一定匹配，如果包含这些不匹配的属性有可能会减慢或误导学习过程。

数据集成时会导致数据冗余，对于属性间冗余我们可以先分析检测，然后再删除。

### 3、数据变换

数据变换主要目的是找到数据的特征表示，用规格化，规约，却换，旋转，投影等操作减少有效变量的数目或找到数据的不变式。

### 4、数据规约

数据规约是指将数据库中的大量数据进行规约，规约后的数据尽可能的保持原数据的完整性，但数据量比要不规约前要少得多，这样能提高挖掘的性能，并且挖掘效率会得到很大的提高。

数据规约的策略主要有：数据立方体聚集、数据压缩、数值压缩、维规约、离散化和概念分层。

## （六）分类算法常用模型

决策树分类、神经网络分类、支持向量机分类、朴素贝叶斯分类、K-最近邻、基于关联规则的分类等

## （七） [分类算法常用模型优缺点对比](http://www.cnblogs.com/94julia/p/3294464.html)

### 1 贝叶斯分类器

　　优点：接受大量数据训练和查询时所具备的高速度，支持增量式训练；对分类器实际学习的解释相对简单

　　缺点：无法处理基于特征组合所产生的变化结果

### 2 决策树分类器

　　优点：很容易解释一个受训模型，而且算法将最为重要的判断因素都很好的安排在了靠近树的根部位置；能够同时处理分类数据和数值数据；很容易处理变量之间的相互影响；适合小规模数据

　　缺点：不擅长对数值结果进行预测；不支持增量式训练

### 3 神经网络

　　优点：能够处理复杂的非线性函数，并且能发现不同输入间的依赖关系；支持增量式训练

　　缺点：黑盒方法，无法确定推导过程；选择训练数据的比率与问题相适应的网络规模方面，没有明确的规则可以遵循，选择过高的训练数据比率有可能导致网络对噪声数据产生过渡归纳的现象，而选择过低的训练比率，则意味着除了已知数据，网络有可能不会再进一步学习了。

### 4 支持向量机

　　优点：通过将分类输入转化成数值输入，可以令支持向量同时支持分类数据和数值数据；适合大规模数据

　　缺点：针对每个数据集的最佳核变函数及其相应的参数都是不一样的，而且每当遇到新的数据集都必须重新确定这些函数及其参数；黑盒技术，由于存在高维空间的变换，SVM的分类过程更加难以解释。

### 5 K-最近邻

　　优点：能够利用复杂函数进行数值预测，同时又保持简单易懂的特点；合理的数据缩放量；在线（online）技术，新数据可以随时被添加进来，而不必重新进行训练。

　　缺点：要求所有训练数据都必须缺一不可；寻找合理的缩放因子比较乏味

## （八）分类算法运用领域

分类技术在很多领域都有应用，例如：通过对客户的分类来构造一个分类模型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。机器学习、专家系统、统计学和神经网络等领域的研究人员已经提出了许多具体的分类预测方法。

# 三、决策树分类

## （一）决策树概念

决策树是一种非常经典的分类算法，经过数据集的训练，能够高效的判断出一个数据项所属的类别。决策树技术主要用于分类和预测，是一种有监督的学习，能提前给定一定类别和数据集合。决策树是一颗树形的数据结构，它即可以是多叉树也可以是二叉树。决策树实际上是一种基于贪心策略构造的，每次选择的都是最优的属性进行分裂。它采用自顶向下的递归方式，在决策树的内部节点进行属性值的比较并根据不同属性判断从该节点向下的分支，在决策树的叶节点得到结论。

常见的决策树分类算法有ID3、C4.5（C5.0）、CART、PUBLIC、SLIQ、CART、CN2、SPRINT等。

## （二）ID3算法

ID3算法是1986年Quinlan提出的一个著名决策树生成方法，是目前引用率很高的一种算法。它通过选择窗口来形成决策树，是利用信息论中的互信息（信息增益）寻找数据库中具有最大信息量的属性字段，建立决策树的一个节点，再根据该属性字段的不同取值建立树的分支；在每个分支子集中重复建立树的下层节点和分支过程。这种方法的优点是描述简单，分类速度快，特别适合大规模的数据处理。缺点是：

1）互信息的计算依赖于属性取值的数目较多的特征，而属性取值较多的属性不一定最优；

2）是非递增学习算法；

3）抗噪性差，训练例子中正例和反例较难控制。

  从[信息论](http://en.wikipedia.org/wiki/Information_theory)知识中我们直到，期望信息越小，[信息增益](http://en.wikipedia.org/wiki/Information_gain)越大，从而纯度越高。所以ID3算法的核心思想就是以信息增益度量属性选择，选择分裂后信息增益最大的属性进行分裂。下面先定义几个要用到的概念。

      设D为用类别对训练元组进行的划分，则D的[熵](http://en.wikipedia.org/wiki/Entropy)（entropy）表示为：

http://latex.codecogs.com/gif.latex?info(D)=-\sum%20%5em_%7bi=1%7dp_ilog_2(p_i)

      其中pi表示第i个类别在整个训练元组中出现的概率，可以用属于此类别元素的数量除以训练元组元素总数量作为估计。熵的实际意义表示是D中元组的类标号所需要的平均信息量。

      现在我们假设将训练元组D按属性A进行划分，则A对D划分的期望信息为：

http://latex.codecogs.com/gif.latex?info_A(D)=\sum%20%5ev_%7bj=1%7d\frac%7b|D_j|%7d%7b|D|%7dinfo(D_j)

      而信息增益即为两者的差值：

http://latex.codecogs.com/gif.latex?gain(A)=info(D)-info_A(D)

## （三）C4.5算法

C4.5 ( Classsification 4. 5 )算法是是 Quinlan 在1993 年提出的。ID3算法的改进算 法，也是之后诸多算法的基础 。

假设T为训练集，为T构造决策树时，根据Information Gain值选择作为分裂结点的属性及标准，按照此标准将T分成n个子集。若第i个子集Ti含有的元组的类一致，该结点就成为决策树的叶子结点而停止分裂。对于不满足此条件的T的子集，按照上述方法继续分裂直至所有子集所含元组都属于一个类别为止。

C4.5算法继承了ID3算法的优点，并在ID3算法的基础上进行了改进：

1) 使用信息增益率来选择属性，这样克服了使用信息增益选择属性时偏向选择取值多的属性的不足；

2) 在树构造过程中进行剪枝；

3) 能对连续属性的离散化处理；

4) 能对不完整数据进行处理。

C4.5算法的优点有：产生的分类规则易于理解，准确率较高。C4.5的缺点是：在构造树过程中，需对数据集进行多次的顺序扫描和排序，这样导致算法的低效。C4.5只适用于能够驻留于内存的数据集，当训练集大得无法在内存容纳时程序无法运行。

ID3算法存在一个问题，就是偏向于多值属性，例如，如果存在唯一标识属性ID，则ID3会选择它作为分裂属性，这样虽然使得划分充分纯净，但这种划分对分类几乎毫无用处。ID3的后继算法C4.5使用[增益率](http://en.wikipedia.org/wiki/Information_gain_ratio)（gain ratio）的信息增益扩充，试图克服这个偏倚。

      C4.5算法首先定义了“分裂信息”，其定义可以表示成：

http://latex.codecogs.com/gif.latex?split\_info_A(D)=-\sum%20%5ev_%7bj=1%7d\frac%7b|D_j|%7d%7b|D|%7dlog_2(\frac%7b|D_j|%7d%7b|D|%7d)

      其中各符号意义与ID3算法相同，然后，增益率被定义为：

http://latex.codecogs.com/gif.latex?gain\_ratio(A)=\frac%7bgain(A)%7d%7bsplit\_info(A)%7d

      C4.5选择具有最大增益率的属性作为分裂属性，其具体应用与ID3类似，不再赘述。

### （四）SLIQ算法

SLIQ算法是对c4.5算法的实现算法进行了该进，在决策树的构造过程中采用了“预排序”和“广度优先”技术。在一般决策树中，使用信息量作为评价节点分裂质量的参数。

SLIQ算法能处理更多的训练集，在一定程度上具有良好的随记录个数和属性个数增长的可拓展性。但它也存在如下缺点：

1. 由于需要将类别列表存放于内存，而类别列表的长度与训练集的长度是相同的，这就一定程度上限制了可以处理的数据集的大小。
2. 由于采用了预排序技术，而排序算法的复杂度本身并不是与记录个数成线性关系，因此使得!"#$算法不可能达到随记录数目增长的线性可扩展性。

### （五）SPRINT算法

为了减少需要驻留于内存的数据量，SPRINT算法进一步改进了决策树算法实现的据结构，去掉在SLIQ中需要驻留于内存的类别列表，将它的类别列合并到每个属性列表中。这样在，遍历每个属性列表寻找当前结点的最优分裂标准时，不必参照其他信息。而对结点的分裂表现在对属性列表的分裂，即将每个属性列表分成两个，分别存放属于各个结点的记录。其优点是在寻找每个结点的最优分裂标准时变得相对简单一些，但是其缺点是对非分裂属性的属性列表进行分裂变得很困难。解决的办法是对分裂属性进行分裂时用哈希表表记录下每个记录属于哪个孩子结点，若内存能够容纳下整个哈希表，其他属性列表的分裂只需参照该哈希表即可。由于哈希表的大小与训练集的大小成正比，当训练集很大时，哈希表可能无法在内存容纳，此时分裂只能分批执行，这使得SPRINT算法的可扩展性仍然不是很好。

# 四、朴素贝叶斯分类

## （一）贝叶斯定理

贝叶斯定理（Bayes' theorem）也称贝叶斯推理，在18世纪，由英国学者[贝叶斯](http://baike.baidu.com/view/77778.htm)(1702～1763)提出。贝叶斯定理是[概率论](http://zh.wikipedia.org/wiki/%E6%A6%82%E7%8E%87%E8%AE%BA)中的一个结论，它跟[随机变量](http://zh.wikipedia.org/wiki/%E9%9A%8F%E6%9C%BA%E5%8F%98%E9%87%8F)的[条件概率](http://zh.wikipedia.org/wiki/%E6%9D%A1%E4%BB%B6%E6%A6%82%E7%8E%87)以及[边缘概率分布](http://zh.wikipedia.org/wiki/%E8%81%94%E5%90%88%E5%88%86%E5%B8%83)有关。

通常，事件A在事件B（发生）的条件下的概率，与事件B在事件A的条件下的概率是不一样的；然而，这两者是有确定的关系，贝叶斯定理就是这种关系的陈述。贝叶斯公式的用途在于通过己知三个概率函数推出第四个。它的内容是:在B出现的前提下,A出现的概率等于A出现的前提下B出现的概率乘以A出现的概率再除以B出现的概率。通过联系A与B,计算从一个事件产生另一事件的概率。

贝叶斯公式：

贝叶斯（Bayes）分类算法是一类利用概率统计知识进行分类的算法，如朴素贝叶斯（Naive Bayes）算法。这些算法主要利用Bayes定理来预测一个未知类别的样本属于各个类别的可能性，选择其中可能性最大的一个类别作为该样本的最终类别。由于贝叶斯定理的成立本身需要一个很强的条件独立性假设前提，而此假设在实际情况中经常是不成立的，因而其分类准确性就会下降。为此就出现了许多降低独立性假设的贝叶斯分类算法，如TAN（Tree Augmented Na?ve Bayes)算法，它是在贝叶斯网络结构的基础上增加属性对之间的关联来实现的。

## （二）朴素贝叶斯分类器

朴素贝叶斯分类器(Naive Bayes Classifier,或 NBC)发源于古典数学理论，有着坚实的数学基础，以及稳定的分类效率。Bayes分类算法是一类利用概率统计知识进行分类的算法。NBC模型所需估计的参数很少，对缺失数据不太敏感，算法也比较简单。理论上，NBC模型与其他分类方法相比具有最小的误差率。但是实际上并非总是如此，这是因为NBC模型假设属性之间相互独立，这个假设在实际应用中往往是不成立的，这给NBC模型的正确分类带来了一定影响。

给定类标号y，朴素贝叶斯分类器在估计类条件概率时假设属性之间条件独立。条件独立假设可形式化地表述如下：

其中每个属性集包含d个属性。

## （三）朴素贝叶斯分类器特点

1）需要知道先验概率

先验概率是计算后验概率的基础。在传统的概率理论中，先验概率可以由大量 的重复实验所获得的各类样本出现的频率来近似获得，其基础是“大数定律”， 这一思想称为“频率主义”。而在称为“贝叶斯主义”的数理统计学派中，他 们认为时间是单向的，许多事件的发生不具有可重复性，因此先验概率只能根 据对置信度的主观判定来给出，也可以说由“信仰”来确定。

2) 按照获得的信息对先验概率进行修正

在没有获得任何信息的时候，如果要进行分类判别，只能依据各类存在的先验 概率，将样本划分到先验概率大的一类中。而在获得了更多关于样本特征的信 息后，可以依照贝叶斯公式对先验概率进行修正，得到后验概率，提高分类决 策的准确性和置信度。

3) 分类决策存在错误率

由于贝叶斯分类是在样本取得某特征值时对它属于各类的概率进行推测，并无 法获得样本真实的类别归属情况，所以分类决策一定存在错误率，即使错误率 很低，分类错误的情况也可能发生。

## （四）朴素贝叶斯分类算法

朴素贝叶斯分类算法采用了变量独立假设的最初形式，也是限制最多的一种形式，它假设每个特征变量Xi在给定类别变量C 下都是独立的，则有:

在计算过程中，取训练样本中xi的最大似然估计作为给定ck下的P（Xi = xi），即：

其中，nck是类别ck的样本数, 是类别ck中特征变量Xi值等于xi 的样本数。

在计算时经常等于0，故采用取对数的方法将其化为：

# 五、支持向量机分类

## （一）简介

在1995年，Corinna Cortes和Vapnik等人首先提出了支持向量机( SVM, support vector machine )的概念。SVN是通过最优化方法来解决机器学习问题的新工具。

支持向量机是建立在统计学习理论的[VC维](http://baike.baidu.com/view/3685811.htm)理论和结构风险最小原理基础之上的，根据有限的样本信息在模型对特定训练样本的学习精度和学习能力之间寻求折中。

支持向量：指在间隔区边缘的训练样本点。

机（machine，机器）：是一个算法。在机器学习领域，我们通常把算法看做是一个机器。

## （二）logistic回归

Logistic回归目的是从特征学习出一个0/1分类模型，而这个模型是将特性的线性组合作为自变量，由于自变量的取值范围是负无穷到正无穷。因此，使用logistic函数（或称作sigmoid函数）将自变量映射到(0,1)上，映射后的值被认为是属于y=1的概率。

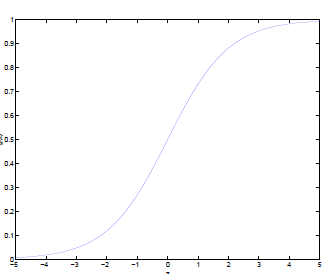
形式化表示就是

假设函数

[clip_image001](http://images.cnblogs.com/cnblogs_com/jerrylead/201103/201103131131373510.png)

其中x是n维特征向量，函数g就是logistic函数。

[clip_image002](http://images.cnblogs.com/cnblogs_com/jerrylead/201103/20110313113137969.png)的图像是

[](http://images.cnblogs.com/cnblogs_com/jerrylead/201103/201103131131375604.png)

可以看到，将无穷映射到了(0,1)。

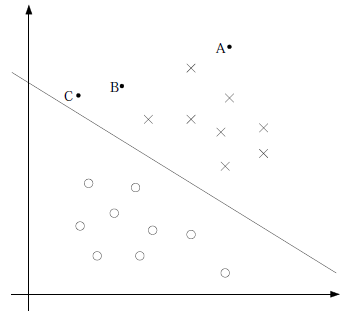
而假设函数就是特征属于y=1的概率。

[clip_image004](http://images.cnblogs.com/cnblogs_com/jerrylead/201103/201103131131385015.png)

当我们要判别一个新来的特征属于哪个类时，只需求[clip_image006](http://images.cnblogs.com/cnblogs_com/jerrylead/201103/201103131131389409.png)，若大于0.5就是y=1的类，反之属于y=0类。

再审视一下[clip_image006[1]](http://images.cnblogs.com/cnblogs_com/jerrylead/201103/201103131131392374.png)，发现[clip_image006[2]](http://images.cnblogs.com/cnblogs_com/jerrylead/201103/201103131131403704.png)只和[clip_image008](http://images.cnblogs.com/cnblogs_com/jerrylead/201103/201103131131416669.png)有关，[clip_image008[1]](http://images.cnblogs.com/cnblogs_com/jerrylead/201103/201103131131416047.png)>0，那么[clip_image010](http://images.cnblogs.com/cnblogs_com/jerrylead/201103/20110313113142997.png)，g(z)只不过是用来映射，真实的类别决定权还在[clip_image008[2]](http://images.cnblogs.com/cnblogs_com/jerrylead/201103/201103131131425948.png)。还有当[clip_image012](http://images.cnblogs.com/cnblogs_com/jerrylead/201103/20110313113143898.png)时，[clip_image006[3]](http://images.cnblogs.com/cnblogs_com/jerrylead/201103/201103131131433340.png)=1，反之[clip_image006[4]](http://images.cnblogs.com/cnblogs_com/jerrylead/201103/201103131131442402.png)=0。如果我们只从[clip_image008[3]](http://images.cnblogs.com/cnblogs_com/jerrylead/201103/201103131131456240.png)出发，希望模型达到的目标无非就是让训练数据中y=1的特征[clip_image012[1]](http://images.cnblogs.com/cnblogs_com/jerrylead/201103/201103131131458715.png)，而是y=0的特征[clip_image014](http://images.cnblogs.com/cnblogs_com/jerrylead/201103/201103131131455617.png)。Logistic回归就是要学习得到[clip_image016](http://images.cnblogs.com/cnblogs_com/jerrylead/201103/201103131131464679.png)，使得正例的特征远大于0，负例的特征远小于0，强调在全部训练实例上达到这个目标。

图形化表示如下：

[](http://images.cnblogs.com/cnblogs_com/jerrylead/201103/201103131131479073.png)

中间那条线是[clip_image019](http://images.cnblogs.com/cnblogs_com/jerrylead/201103/201103131131479879.png)，logistic回顾强调所有点尽可能地远离中间那条线。学习出的结果也就中间那条线。考虑上面3个点A、B和C。从图中我们可以确定A是×类别的，然而C我们是不太确定的，B还算能够确定。这样我们可以得出结论，我们更应该关心靠近中间分割线的点，让他们尽可能地远离中间线，而不是在所有点上达到最优。因为那样的话，要使得一部分点靠近中间线来换取另外一部分点更加远离中间线。我想这就是支持向量机的思路和logistic回归的不同点，一个考虑局部（不关心已经确定远离的点），一个考虑全局（已经远离的点可能通过调整中间线使其能够更加远离）。

## （三）形式化表示

我们这次使用的结果标签是y=-1,y=1，替换在logistic回归中使用的y=0和y=1。同时将[clip_image016[1]](http://images.cnblogs.com/cnblogs_com/jerrylead/201103/201103131131482006.png)替换成w和b。以前的[clip_image021](http://images.cnblogs.com/cnblogs_com/jerrylead/201103/201103131131483368.png)，其中认为[clip_image023](http://images.cnblogs.com/cnblogs_com/jerrylead/201103/201103131131482256.png)。现在我们替换[clip_image025](http://images.cnblogs.com/cnblogs_com/jerrylead/201103/20110313113149794.png)为b，后面替换[clip_image027](http://images.cnblogs.com/cnblogs_com/jerrylead/201103/201103131131492157.png)为[clip_image029](http://images.cnblogs.com/cnblogs_com/jerrylead/201103/201103131131503760.png)（即[clip_image031](http://images.cnblogs.com/cnblogs_com/jerrylead/201103/20110313113150662.png)）。这样，我们让[clip_image033](http://images.cnblogs.com/cnblogs_com/jerrylead/201103/201103131131511152.png)，进一步[clip_image035](http://images.cnblogs.com/cnblogs_com/jerrylead/201103/201103131131513628.png)。也就是说除了y由y=0变为y=-1，只是标记不同外，与logistic回归的形式化表示没区别。再明确下假设函数

[clip_image037](http://images.cnblogs.com/cnblogs_com/jerrylead/201103/201103131131518055.png)

上一节提到过我们只需考虑[clip_image008[4]](http://images.cnblogs.com/cnblogs_com/jerrylead/201103/201103131131526593.png)的正负问题，而不用关心g(z)，因此我们这里将g(z)做一个简化，将其简单映射到y=-1和y=1上。映射关系如下：

[clip_image039](http://images.cnblogs.com/cnblogs_com/jerrylead/201103/201103131131529908.png)

## （四）函数间隔（functional margin）和几何间隔（geometric margin）

给定一个训练样本[clip_image041](http://images.cnblogs.com/cnblogs_com/jerrylead/201103/201103131131525971.png)，x是特征，y是结果标签。i表示第i个样本。我们定义函数间隔如下：

[clip_image043](http://images.cnblogs.com/cnblogs_com/jerrylead/201103/201103131131534858.png)

可想而知，当[clip_image045](http://images.cnblogs.com/cnblogs_com/jerrylead/201103/201103131131539808.png)时，在我们的g(z)定义中，[clip_image047](http://images.cnblogs.com/cnblogs_com/jerrylead/201103/201103131131545249.png)，[clip_image049](http://images.cnblogs.com/cnblogs_com/jerrylead/201103/20110313113155200.png)的值实际上就是[clip_image051](http://images.cnblogs.com/cnblogs_com/jerrylead/201103/201103131131552675.png)。反之亦然。为了使函数间隔最大（更大的信心确定该例是正例还是反例），当[clip_image045[1]](http://images.cnblogs.com/cnblogs_com/jerrylead/201103/201103131131559577.png)时，[clip_image053](http://images.cnblogs.com/cnblogs_com/jerrylead/201103/201103131131564528.png)应该是个大正数，反之是个大负数。因此函数间隔代表了我们认为特征是正例还是反例的确信度。

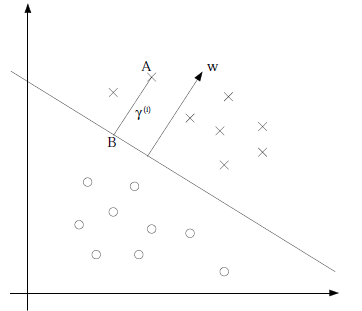
继续考虑w和b，如果同时加大w和b，比如在[clip_image055](http://images.cnblogs.com/cnblogs_com/jerrylead/201103/201103131131568639.png)前面乘个系数比如2，那么所有点的函数间隔都会增大二倍，这个对求解问题来说不应该有影响，因为我们要求解的是[clip_image057](http://images.cnblogs.com/cnblogs_com/jerrylead/201103/201103131131569478.png)，同时扩大w和b对结果是无影响的。这样，我们为了限制w和b，可能需要加入归一化条件，毕竟求解的目标是确定唯一一个w和b，而不是多组线性相关的向量。这个归一化一会再考虑。

刚刚我们定义的函数间隔是针对某一个样本的，现在我们定义全局样本上的函数间隔

[clip_image058](http://images.cnblogs.com/cnblogs_com/jerrylead/201103/201103131131573905.png)

说白了就是在训练样本上分类正例和负例确信度最小那个函数间隔。

接下来定义几何间隔，先看图

[](http://images.cnblogs.com/cnblogs_com/jerrylead/201103/201103131131571364.png)

假设我们有了B点所在的[clip_image057[1]](http://images.cnblogs.com/cnblogs_com/jerrylead/201103/20110313113158426.png)分割面。任何其他一点，比如A到该面的距离以[clip_image061](http://images.cnblogs.com/cnblogs_com/jerrylead/201103/201103131131584820.png)表示，假设B就是A在分割面上的投影。我们知道向量BA的方向是[clip_image063](http://images.cnblogs.com/cnblogs_com/jerrylead/201103/201103131131593359.png)（分割面的梯度），单位向量是[clip_image065](http://images.cnblogs.com/cnblogs_com/jerrylead/201103/201103131131593608.png)。A点是[clip_image041[1]](http://images.cnblogs.com/cnblogs_com/jerrylead/201103/201103131132006050.png)，所以B点是x=[clip_image067](http://images.cnblogs.com/cnblogs_com/jerrylead/201103/201103131132002114.png)（利用初中的几何知识），带入[clip_image057[2]](http://images.cnblogs.com/cnblogs_com/jerrylead/201103/201103131132002637.png)得，

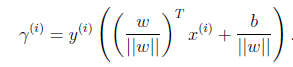
[clip_image069](http://images.cnblogs.com/cnblogs_com/jerrylead/201103/201103131132011524.png)

进一步得到

[clip_image070](http://images.cnblogs.com/cnblogs_com/jerrylead/201103/201103131132017588.png)

[clip_image061[1]](http://images.cnblogs.com/cnblogs_com/jerrylead/201103/201103131132018394.png)实际上就是点到平面距离。

再换种更加优雅的写法：

[](http://images.cnblogs.com/cnblogs_com/jerrylead/201103/20110313113202836.png)

当[clip_image073](http://images.cnblogs.com/cnblogs_com/jerrylead/201103/201103131132024914.png)时，不就是函数间隔吗？是的，前面提到的函数间隔归一化结果就是几何间隔。他们为什么会一样呢？因为函数间隔是我们定义的，在定义的时候就有几何间隔的色彩。同样，同时扩大w和b，w扩大几倍，[clip_image075](http://images.cnblogs.com/cnblogs_com/jerrylead/201103/201103131132037880.png)就扩大几倍，结果无影响。同样定义全局的几何间隔[clip_image076](http://images.cnblogs.com/cnblogs_com/jerrylead/201103/201103131132042274.png)

## （五）支持向量机的优缺点

支持向量机有以下特点：

（1）支持向量机方法的理论基础是：非线性映射, 支持向量机利用内积核函数代替向高维空间的非线性映射；

（2）支持向量机的目标是：对特征空间划分的最优超平面

（3）支持向量机方法的核心是：最大化分类边际的思想

（4）支持向量机的训练结果是：支持向量,

（5）在支持向量机分类决策中起决定作用的是支持向量；

（6）支持向量机是一种有坚实理论基础的新颖的小样本学习方法。

（7）支持向量机的最终决策函数只由少数的支持向量所确定,计算的复杂程度取决于支持向量的数目,而不是样本空间的维数。

（8）少数支持向量决定了最终结果,该方法不但算法简单,而且具有较好的“鲁棒”性，但不能帮助我们抓住关键样本、“剔除”大量冗余样本。

支持向量机有以下缺点：

(1) 支持向量机算法对大规模训练样本难以实施，因为支持向量机是借助二次规划来求解支持向量，而求解二次规划将涉及n阶矩阵的计算（n为样本的个数），当n数目很大时该矩阵的存储和计算时，会耗费大量的机器内存和运算时间。

改进的方法有：

J.Platt的SMO算法、C.J.C.Burges等的PCGC、、.Joachims的SVM、张学工的CSVM以及O.L.Mangasarian的SOR算法等

(2) 用支持向量机解决多分类问题存在困难 。经典的支持向量机算法只给出了二类分类的算法，但在数据挖掘的实际应用中，通常要解决多类的分类问题。

改进方法是通过多个二类支持向量机的组合来解决。

# 人工神经网络分类

## （一）简介

人工神经网络（ANN）是在现代[神经科学](http://baike.baidu.com/view/1542619.htm" \t "_blank)研究成果的基础上提出的。神经网络模拟人类的神经元功能，从结构上模仿生物的神经网络，经过输入层、隐藏层、输出层等，对数据进行调整、计算，最后得到结果。是一种通过训练来学习的非线性预测模型，可以完成分类、聚类、特征挖掘、回归分析等多种数据挖掘任务。

## （三）主要解决问题

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 。

## （四）人工神经网络的特点

1. 可以充分逼近任意复杂的非线性关系；

2、所有定量或定性的信息都等势分布贮存于网络内的各神经元，故有很强的鲁棒性和容错性；   
3、采用并行分布处理方法，使得快速进行大量运算成为可能；   
4、可学习和自适应不知道或不确定的系统；   
5、能够同时处理定量、定性知识。

## （五）对神经网络分类器进行评估

在神经网络分类器已定型并且生成了一组最佳权重和偏置后，下一步是确定对于测试数据而言，最终生成的模型的精确度（其中，模型意味着具有一组最佳权重的神经网络）。 尽管可以使用方差和或互熵误差之类的度量值，但精确度的合理度量值只不过是该模型生成的正确预测的百分比。

# 七、结语

随着信息技术的不断发展，国际互联网的全球化热潮使人类社会进入了一个新的信息时代。信息无处不在，我们说的一句话一个字都是信息。随着信息时代的来临，信息系统的广泛深入和发展，建立与开发管理信息系统已具有深刻意义。在浩瀚的信息海里有我们想要的信息，也有我们厌恶的信息。发现和挖掘好的信息，存储和保留信息。

参考文献

1. Pang-Ning Tan Michael Steinbach Vipin Kumar著:《数据挖掘导论（完整版）》，人民邮电出版社
2. 刘红岩，陈剑，陈国青：《数据挖掘中的数据分类算法综述》，《清华大学报（自然科学版）》2002年第42卷 第6期
3. 孙娟：《数据挖掘分类算法研究与讨论》，《电脑知识与科技》2008年12月第4卷 第8期
4. 王刚等：《数据挖掘分类算法研究综述》，《科技导报》2006年第12期 第24卷
5. 谈恒贵，王文杰，李游华：《数据挖掘分类算法综述》，《微型机与运用》2005年第2期
6. 李如平：《数据挖掘中决策树分类算法的研究》，《华东理工大学学报（自然科学版）》2010年6月 第33卷 第2期

致谢

历时将近两个月的时间终于将这篇论文写完，在论文的写作过程中遇到了无数的困难和障碍，都在同学和老师的帮助下度过了。尤其要强烈感谢我的论文指导老师—吴萍萍，她对我进行了无私的指导和帮助，不厌其烦的帮助进行论文的修改和改进。在此向帮助和指导过我的各位老师表示最中心的感谢！

感谢这篇论文所涉及到的各位学者。本文引用了数位学者的研究文献，如果没有各位学者的研究成果的帮助和启发，我将很难完成本篇论文的写作。感谢我的同学和朋友，在我写论文的过程中给予我了很多你问素材，还在论文的撰写和排版灯过程中提供热情的帮助。由于我的学术水平有限，所写论文难免有不足之处，恳请各位老师和学友批评和指正！