题目：利用数据挖掘方法预测音乐流行趋势

摘要：

Abstract：

关键词：

目录：

文本主体（15000）：

引言：

数据是知识的源泉，然而，拥有大量的数据和拥有大量有用的知识是两个不同的概念。在过去几年中，关于从数据库发现知识的技术发展的越来越快，这一知识领域的内容也不断被填充和扩大，从而飞速发展。计算机技术和数据收集技术的进步使得我们可以从宽广的范围和以前不可及的速度进行收集和存储信息。收集数据是为了得到信息，而收集到的大量的信息并不意味着信息。海量数据的不断产生，随之而来的问题是这么多的数据让人难以消化，并不能从表面上看出它们所蕴含的信息，而数据挖掘技术正是伴随着这种需求从研究走向应用。

随着电子音乐的不断发展，在空闲的时间听音乐是越来越多人的选择，而根据中国互联网络信息中心的统计，网络音乐应用率已经成为了全国第一网络应用，随着我国人民生活水平的提高越来越多的人喜欢听音乐，对网络与音乐的结合的要求与需求也越来越高。应运而生的，是各种各样的音乐平台，包括移动端与PC端。这些平台为平台用户们提供各种各样的个性化服务，诸如音乐推荐，社交网络分享等等。越来越多的原创歌手愿意把自己的原创歌曲上传分享到平台上供其他人品味。随着这些行为，每天都会有成千上百万的用户生产数以亿计的诸如试听，下载，收藏，分享等行为记录。由音乐平台产生的巨量数据对音乐的流行趋势具有极其重要的指引作用，而运营者所需要的信息则蕴含在这些数据中。2015年国家针对音乐版权问题发布文件，音乐行业随着文件的发布有迎来了一次颠覆，而能否准确预测音乐的流行趋势，判断出即将出现的黑马，是各个音乐平台吸引用户，留存用户，对用户进行附加服务的关键。这就需要数据挖掘技术在音乐领域大展宏图了。而机器学习则是数据挖掘技术中的重中之重，是数据挖掘技术上的瑰宝。

而在本次课题研究中，我通过分析电子音乐平台的用户产生的播放，下载，收藏歌曲的行为记录，通过对目标时间内艺人歌曲播放量的预测，挖掘出即将成为潮流的艺人，从而实现对目标时间段音乐流行趋势的把控。具体做法是分析用户行为信息结合艺人及其歌曲信息，进行数据预处理，基于机器学习构建预测模型，采用人工神经网络和支持向量机分别预测每日的播放量。

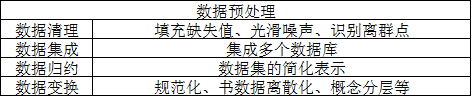
**正文：**

**第二章 相关理论知识介绍：**

**2.1数据挖掘流程：**

数据挖掘是从大量的数据中挖掘出有趣模式和知识的过程。数据源一般是数据库，数据仓库，Web等等，得到的数据称之为数据集（dataset）。其中数据仓库是数据挖掘的独有内容，是从多个数据源收集的信息存储库。“数据仓库是一个面向主题的、集成的、事变的、非易失的数据集合，支持管理者的决策过程”。

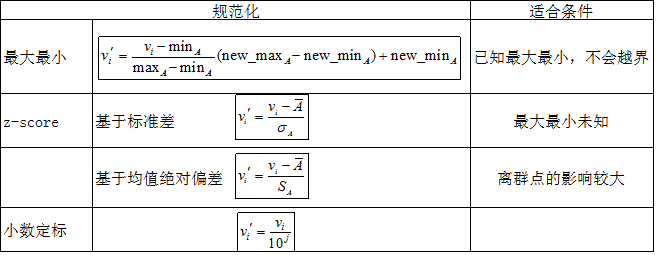
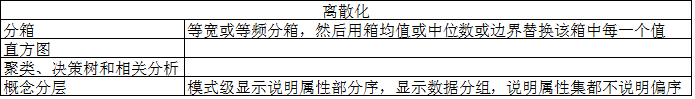
而数据挖掘的一般过程主要有以下几点：

1. 数据集选取：一般的数据集是已经存在的或者知道如何获取的，或者根据分析者自己根据任务的目的，从实际中构造自己需要的数据。例如访问某个数据库，网上抓取需要过滤的信息数据，问卷调查手动收集。
2. 数据预处理：数据预处理的目的是提高数据的质量：准确性，一致性与完整性。包括下面列表内的四个步骤：

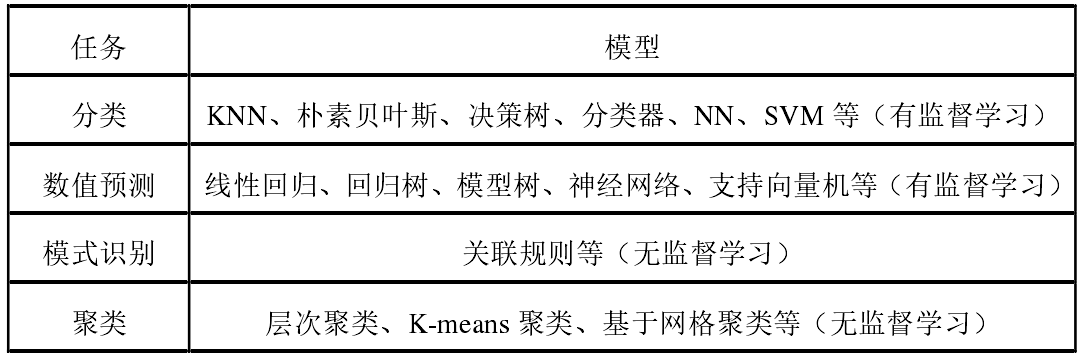
1.数据清理：填充缺失值，光滑噪声，识别离群点

2.数据集成：实体识别，冗余和相关分析

3.数据规约：维规约（小波变换 主成分分析）；数量规约（较小的数据替代原始数据）；数据压缩（有损和无损两种，尤其是图像视频等多媒体常用）

4.数据变换和数据离散化：数据变换：光滑，属性构造，聚集，规范化，离散化，概念分层

（3）特征转化：将预处理过的数据进行特征提取，特征应当尽可能的描述数据，特征的提取对模型优化至关重要。合适的特征才可以使机器学习算法达到最优。

（4）数据建模：根据机器学习模型的优缺点，选择适宜该数据任务的最佳模型，再进行测试数据的误差对比，选择误差小的模型，目前主流机器学习算法优缺点如下：

有监督的机器学习：有监督的学习可分为“回归”和“分类”问题。在回归问题中，我们会预测一个连续值；而在分类问题中，我们会预测一个离散值。

无监督的机器学习：在无监督学习中，我们可以通过聚类的方式从数据中提取一个特殊结构。在无监督学习中给定的数据和监督学习中给定的是不一样的。在无监督学习中给定的数据没有任何标签或者只有同一种标签。

另外，我们还需要调整模型的参数，使得模型的表现尽可能最最优。调节参数的方法主要有：手动调优，随机搜索，网格搜索，贝叶斯方法等等。

（5）结果表达与解释：我们将数据挖掘分析的结果进行分析和评价，分析说明该模型是否有效，得到了可以应用于工业产品的知识积累。

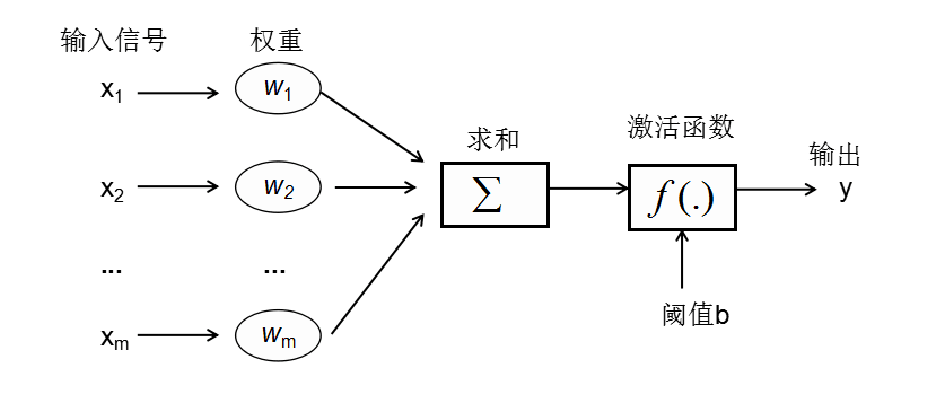
**人工神经网络算法**

**人工神经网络简介：**

人脑是一个非常非常复杂而且非常高效的智能系统，它具备非常优越的信息处理能力，学习能力，信息的存储能力。正是由于大脑的这些功能，人类百万年来增加的闹容量，使得人类从万万千千物种中脱颖而出，在历史长河中不断进步，不断地改善我们的生活环境。所以，古今学者们不断地对人脑进行研究，其他得到一种模拟人类大脑思维方式的方法。最终，著名的人工神经网络（Artificial Neural Network，即ANN）理论诞生了。神经网络是一种应用类似于大脑神经突触连接结构进行信息处理的数学模型。它实在人类对自身大脑组织结合和思维机制的认识理解基础上模拟出来的，它是根植于神经科学，数学，思维科学，人工智能，统计学，物理学，计算机科学以及工程科学的一门技术。

人工神经网络模型最早是由心理学家McCulloch 与数理逻辑学家Pitts于1943年共同提出，即为M-P模型，此模型虽然比较简单，但是意义重大。在模型中，通过把神经元看作个功能逻辑器件来实现算法，从此开创了视神经网络模型的理论研究。而神经网络模型与人脑的相似之处主要有两点：一是它们都从外界获取知识并且学习；二是每个神经元都能在内部存储知识，并且向下一个神经元发出相应信号。

人工神经网络模型：



上图为人工神经元的模型图，它定义了输入信号x和输出信号y之间的关系。对于每一个输入的信号xm加权后求它们的和，权重则为上层神经元对此神经元连接的权值。然后通过f（）标识的激活函数进行传递，最终得到输出y。目前已经发展出了很多种神经网络的模型，这些发展出来的模型都可以由激活函数、网络拓扑、训练算法这三个特征来构建。

激活函数：激活函数对于人工神经网络模型去学习理解非常复杂和非线性的函数来说具有非常重要的作用。它们将非线性引入我们的网络中。其中主要目的是将ANN模型中的一个输入信号转换成一个输出信号，该输出信号被用作堆叠中下一个层的输入。它会判断输入的求和是否激活阈值，若满足则传递，否则不进行操作。激活函数的作用简单来说，无论有多少层隐含层，没有激活函数的话最终的结果还是原始输入的线性变化。它给是数据的提供非线性的变化。典型的激活函数有Sigmoid函数，Tanh-Hyperbolic tangent （双曲线正切函数），ReLu-Rectified linear units（线性修正单元），高斯激活函数，单位跳跃激活函数等。

网络拓扑：网络拓扑是描述神经元的层数、数量、连接方式的结构解释。网络拓扑结构对神经网络的学习能力和复杂度起到了决定性的作用。所有的神经网络模型结构可以通过三个特征来区分：

层的数目：输入的节点称之为输入层；输出的节点称之为输出层。网络的复杂程度与层数息息相关。在输入层与输出层之间，我们通常会增加一个或者多个隐层。

传播方向：传播方向即为神经网络中箭头的方向。前馈网络中的输入信号沿一个方向连续传递直接抵达输出层。它可以应用多个隐层，因为它的层数，层内的节点数都可改变。而反馈网络（递归网络）可以使信号在两个方向上传递，更加体现了生物大脑的工作方式。但是反馈网络的实现目前还较少。

 网络中每一层的节点数：中间层的节点数目是改变网络复杂性的关键，因为输入节点与输出节点是固定的。现在并没有计算最优隐层节点数目的方法与规则，有很多因素会影响最优节点数，如输入节点数目、接受数据的数量、噪声数据等等。下图表示了一个通用的神经网络模型结构：

训练算法：随着神经网络算法的不断改进，对于很多不同领域的应用，学者们生产了许多不同的算法模型。目前已经有百种圣经网络模型，它们可以应用在不同领域中。

在本论文中，我将使用BP（Back Propagation）神经网络模型结构及算法进行音乐播放量的预测。BP是算法是最精确的建模方法之一，它可以用于分类和预测问题，它是一种应用广泛的前向型神经网络模型，能根据不同的输入和输出尽可能满足所有给定的训练数据集的神经网络模型。BP神经网络由三层及以上的神经元组成，同层之间并没有连接。BP神经网络结构与通用神经网络结构一致，均由输入层，1-n个隐层，输出层组成。目前已有理论证明只含有一个隐层的BP神经网络具有逼近任意非线性连续函数的能力。所以一般使用BP神经网络算法时基本只是用一层隐层。

BP神经网络模型的使用主要包括两个过程：首先是根据输入的向量正向传播计算出输出的向量；然后根据输出向量与期望向量的误差反复更新输入层到隐层，隐层到输出层之间的权重和阈值。标准的BP神经网络算法采用误差函数按照梯度下降的方法学习，直到达到预期的误差。

**BP神经网络算法:**在得到了BP神经网络的特点之后，我们依据信号的向前传播和误差的反向传播来构建整个网络。具体步骤如下：

1. 网络初始化：

D:\softwarefile\1826239165\FileRecv\gif.gif假设输入层的节点个数为n，隐含层节点个数为l，输出层的节点个数为m。输入层到隐含层的权重wij，隐含层到输出层的权重为wjk，输入层到隐含层的偏置为aj，隐含层到输出层的偏置为bk。学习速率为η，激励函数为g(x)。其中激励函数为g(x)取Sigmoid函数。形式为：

1. 隐含层的输出：

D:\desktop\论文图片\gif (1).gif根据上面展示的神经网络结构展示，隐含层的输出Hj为：

1. D:\desktop\论文图片\gif (2).gif输出层的输出
2. 误差的计算：

D:\desktop\论文图片\gif (3).gif我们的误差公式为：

D:\desktop\论文图片\gif (4).gif其中YK为期望输出。记Yk– Ok = ek ，则E 可以表示为：

上述公式中，i=1···n j=1···l k=1···m

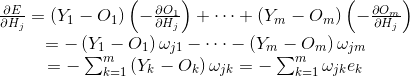
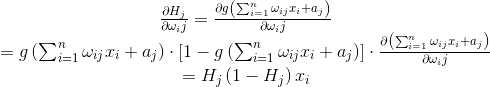
1. 权值的更新：

D:\desktop\论文图片\gif (5).gif权值更新公式为：

D:\desktop\论文图片\gif (6).gif该公式由梯度下降法得来，在误差反向传播的过程中，我们的目标是是的误差函数达到最小值，即minE,首先得出隐含层到输出层的权重更新：

D:\desktop\论文图片\gif (7).gif则权重的更新公式为：

D:\desktop\论文图片\gif (8).gif然后得到输入层到隐含层的权重更新：

其中：

D:\desktop\论文图片\gif (11).gif则权重的更新公式为：

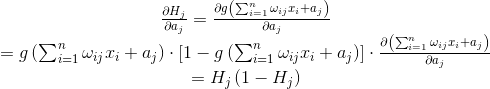
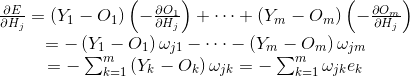
1. 偏置的更新：

D:\softwarefile\1826239165\FileRecv\gif11.gif偏执的更新公式为：

D:\desktop\论文图片\gif (12).gif隐含层到输出层的偏置更新：

偏置更新公式为： bk = bk + ηek

D:\desktop\论文图片\gif (13).gif输入层到隐含层的偏置更新：

其中：

D:\desktop\论文图片\gif (16).gif则更新的偏置公式为：

7、判断算法迭代是否结束：判断神经网络模型训练是否可以结束根据一下几个条件进行判断：

（1）误差E小于给定的误差最大值；

（2）迭代次数大于设定的参数值；

（3）训练误差基本达到了最小值，即连续多次误差E基本上不发生改变。

8、验证算法可行性：用得到的算法测试划分好的数据集，得到模型的输出，对比实际输出，计算预测误差分析模型的构建是否合理。

9、使用模型：结束算法的训练，解决实际的目标问题。

神经网络模型的使用流程图大致如下：



**第三章 数据描述及准备**

**3.1 样本数据描述**

**3.1.1 数据集来源**

**3.1.2 用户行为分析**

**3.1.3 歌曲信息分析**

**3.2 数据探索**

**3.2.1 数据质量分析**

**3.2.2 数据特征分析**

**3.3 数据预处理**

**3.3.1 数据清洗**

**3.3.2 数据集成**

**3.3.3 数据变换**

**样本描述及数据预处理：**

**数值预测模型构建与分析：**

**Ann BP**

**SVM**

**结论：**

**致谢：**

**参考文献：**