# 目录

[1 引言 1](#_Toc515943504)

[2 相关理论知识介绍 2](#_Toc515943505)

[2.1 数据挖掘流程： 2](#_Toc515943506)

[2.2 人工神经网络算法 4](#_Toc515943507)

[2.2.1 人工神经网络简介 4](#_Toc515943508)

[2.2.2 人工神经网络模型： 4](#_Toc515943509)

[2.2.3 BP神经网络算法 6](#_Toc515943510)

[3 数据描述与准备 10](#_Toc515943511)

[3.1 样本数据描述 10](#_Toc515943512)

[3.1.1 数据来源 10](#_Toc515943513)

[3.1.2 用户行为分析 11](#_Toc515943514)

[3.1.3 歌曲信息分析 12](#_Toc515943515)

[3.2 数据探索 12](#_Toc515943516)

[3.2.1 质量分析 13](#_Toc515943517)

[3.2.2 数据特征分析 14](#_Toc515943518)

[3.3 数据预处理 15](#_Toc515943519)

[3.3.1 数据清洗 15](#_Toc515943520)

[3.3.2 数据集成 15](#_Toc515943521)

[3.3.3 数据变换 16](#_Toc515943522)

[3.4 本章小结 16](#_Toc515943523)

[4 人工神经网络算法在音乐播放量预测中的应用 18](#_Toc515943524)

[4.1 ANN算法选择 18](#_Toc515943525)

[4.2 数据准备 18](#_Toc515943526)

[4.3 确定输入元与输出元 19](#_Toc515943527)

[4.4 模型训练及隐层神经元确定 21](#_Toc515943528)

[4.5 实验预测结果 21](#_Toc515943529)

[4.6 本章小结 22](#_Toc515943530)

[5 结论 23](#_Toc515943531)

[致谢 24](#_Toc515943532)

[参考文献 25](#_Toc515943533)

[附录 26](#_Toc515943534)

# 1 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。

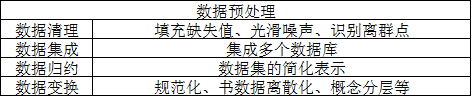
# 2 相关理论知识介绍

## 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：

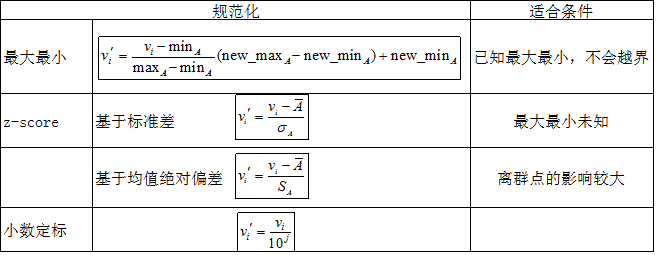
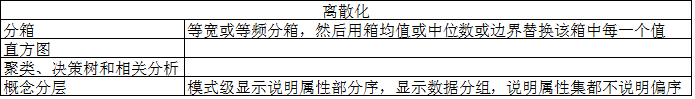
数据集选取：一般的数据集是已经存在的或者知道如何获取的，或者根据分析者自己根据任务的目的，从实际中构造自己需要的数据。例如访问某个数据库，网上抓取需要过滤的信息数据，问卷调查手动收集。

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：

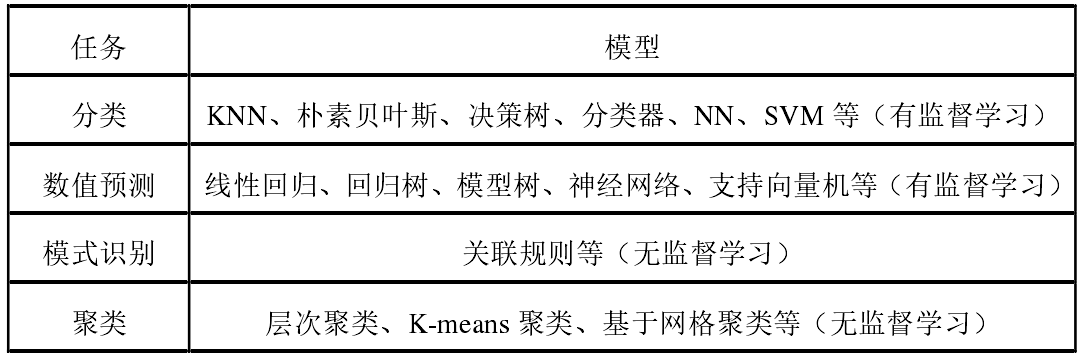
1.数据清理：填充缺失值，光滑噪声，识别离群点

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

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。

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。

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。

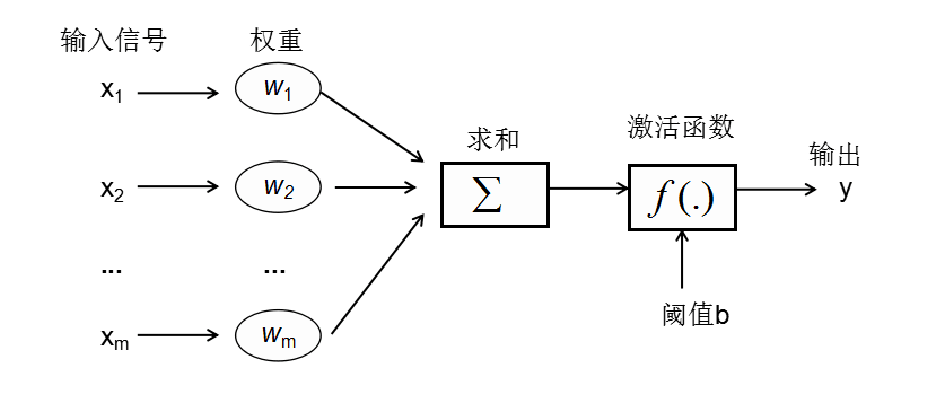
## 2.2 人工神经网络算法

### 2.2.1 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。

### 2.2.2 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。

### 2.2.3 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：

2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：

3D:\desktop\论文图片\gif (2).gif输出层的输出：

4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

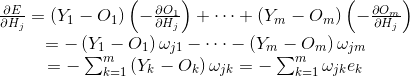
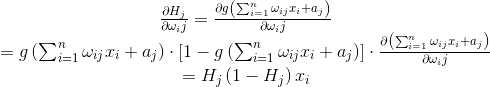
5权值的更新：

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：

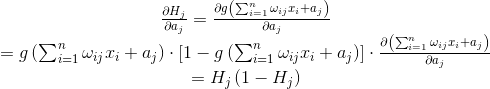
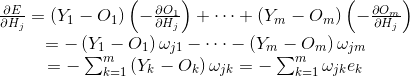
6偏置的更新：

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：



# 3 数据描述与准备

随着越来越多的用户依赖于网络音乐平台，这些音乐平台也吸引着各种各样的原创音乐人。目前，每天有数亿计及的用户活跃在平台上，同时产生了超高数量级人次的用户试听、收藏等行为。在原创艺人和作品方面，数十万的独立音乐人，每月上传上万个原创作品，形成超过百万首曲目的原创作品库，如此庞大的数据资源库对于音乐流行趋势的把握有着极为重要的指引作用。音乐的流行趋势可以根据目前的流行音乐艺人来确定。而判断某个艺人是否为当前的流行音乐艺人则需要根据该艺人最近一段时间的音乐播放量来进行判断。我们可以通过对网络音乐平台用户对不同音乐的操作（播放、下载、收藏）的行为记录，预测艺人接下来一个时期的每天的音乐试听量，进而分析那些歌曲表演者将成为下一阶段的潮流艺人。而这些被预测出的潮流艺人即代表着下一阶段的音乐流行趋势。

## 3.1 样本数据描述

### 3.1.1 数据来源

本文数据来源于阿里云音乐平台。本文所采用的数据集包括了该平台上六个月用户的行为信息和歌曲信息，数据所涵盖的歌曲通过统计初步的播放总热度为二十六万余，代表了其历史播放人次。用户的播放行为共计4783603次，这是我们实验数据的主体部分。数据包含两个部分：一是记录歌曲的信息数据；二是记录用户的操作的数据；两个表格的具体信息如下：

用户行为表

|  |  |  |
| --- | --- | --- |
| 列名 | 类型 | 示例 |
| user\_id | String | 51cecc193b9b7d0478904cd99999fcb4 |
| song\_id | String | 8a103bd3a3295fbf9b3c3bf7972db299 |
| gmt\_create | String | 1426381200 |
| action\_type | String | 2 |
| Ds | String | 20150315 |

表中：

user\_id：该操作的用户id，是用户的唯一标识符。

song\_id：被操作的歌曲的id，是歌曲的唯一标识符。

gmt\_create：用户的播放时间（unix时间戳表示），精确到小时。

action\_type：用户的操作类型1表示播放，2表示下载，3表示收藏。

Ds：该记录产生的日期。

该表记录了所有用户在2015年3月1日到2015年8月30日期间对50名歌曲艺人的歌曲所进行的所有操作（试听，下载，收藏）的记录，共有5652232条用户操作记录。

歌曲艺人表

|  |  |  |
| --- | --- | --- |
| 列名 | 类型 | 示例 |
| song\_id | String | e84ba26cd51b4fcd0811a89957f7bec4 |
| artist\_id | String | 0c80008b0a28d356026f4b1097041689 |
| publish\_time | String | 20121224 |
| song\_init\_plays | String | 2151 |
| Language | String | 2 |
| Gender | String | 1 |

表中：

song\_id：歌曲id，歌曲唯一标识符

artist\_id：歌曲所属的艺人id。

publish\_time：歌曲发行的日趋，精确到天。

song\_init\_plays：歌曲的初始播放数量，表面该歌曲的初始热度。

Language：表示该首歌曲所使用的语种。

Gender：歌曲演唱者性别。

该表记录了50位艺人所出版发行的所有歌曲，共有10842条记录。每一条记录包含了歌曲id，歌曲发布时间，初始播放量，语言等基本信息。初始热度总和为26107万，表示初始播放量综合。

### 3.1.2 用户行为分析

在日常活动中，我结合自身以及周边用户对该音乐平台的使用情况，可以得到一些用户对平台歌曲的操作以及对喜爱歌手的收听习惯。我们在平台中进行的播放，下载，收藏等行为大致表现出了我们对某些艺人的关注程度，抑或是某一类型歌曲的喜爱程度。播放行为大致主要为一下几类：播放日常喜爱听到的老歌，喜爱歌手的热门歌曲怀旧经典歌曲；播放喜爱歌手的新推出歌手，其后的播放量由对歌曲的喜爱程度决定；播放随机歌单或者热门榜单内的歌曲，紧跟潮流播放引发现象级流行的歌曲；寻找新歌播放同类型的歌曲等。然而对于有些用户，主要是追星族，可能会为了给自己喜爱的歌手“打榜”而连续点击一首歌曲引起播放量的暴增（在许多粉丝同时操作的情况下），这样就会对我们进行数据分析造成很大的困扰，于是在分析前我们应注意这些数据进行针对性的数据清洗。

由于本文的预测是关于歌手音乐播放量的，所以针对用户的行为来说我首先想到的是是否可以针对用户是否在预测日期内会播放该艺人的歌曲，所以观察用户播放歌曲的规律。但是在对用户行为表的处理后发现，用户播放的曲目相对分散，并没有发现规律。大多数用户并不会针对某位歌手进行播放，或者集中听排行榜前列的歌曲，只有一部分用户有规律的听歌习惯，这些用户应该是某位歌手的粉丝，或者只会播放热门歌曲。所以我们通过用户与歌曲之间的关系来进行迂回的预测播放量难度较大而且准确率较低。

回到问题本质，对播放量的预测问题实际上是一个时间序列问题。所以正确而直观的做法应该在播放量与日期时间的关系上着重考虑。数据提供了用户播放，收藏，下载的行为，然而收藏与下载的数据少量并且空缺值很多。纵使作为特征进行考虑也不能起到很好的效果，反而有可能很大程度上影响预测结果的准确性。所以在后续的分析中决定摒弃这两类行为信息，针对播放行为进行针对性的特征提取。

### 3.1.3 歌曲信息分析

在针对用户行为表中的用户行为进行初步分析后，我们得出有许多用户操作是不具有分析意义的，而歌曲信息表中，绝大多数数据项都是有用的。publish\_time代表了歌曲的发行日期，这个信息可以给歌曲打上新歌或者是老歌的标签；而song\_init\_plays代表歌曲的初始播放数量，即为从发布到3月1日的总共播放数，代表了歌曲的热度，也可以将歌曲按照热度进行分类，而有些歌手的新歌的初始热度是根据该歌手其他歌曲的热度来赋值的；而language和gender两项在后续操作中发现与播放数相关系数较小，虽然也影响着播放数，但两个属性大致跟歌手是贴合的，同一个歌手的歌曲，language与gender属性一般不会有变化。

对与歌曲信息的分析使得对用户-歌曲之间的关系观看地更加清晰，同时间接的可以对用户-歌手的关系打上标签。

## 3.2 数据探索

数据探索的目的是让研究者最大化对数据的直觉，对要进行分析的数据更加有感觉。对本文数据的探索，不仅需要知道数据里面有什么，还需要知道数据里面没有什么；数据探索的方法应当是利用各种可用的工具，结合各种统计学的图形把数据以各种形式呈现出来，最大化的感受数据的质量和特征。

### 3.2.1 质量分析

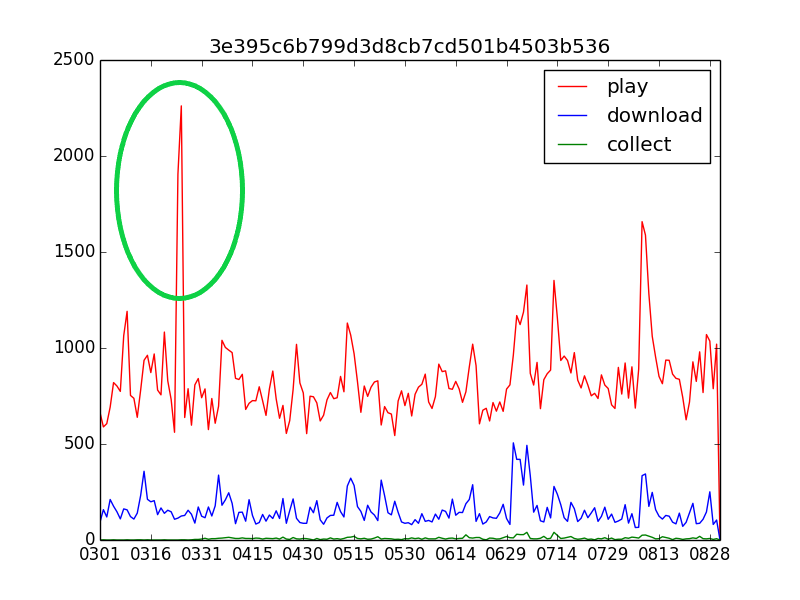
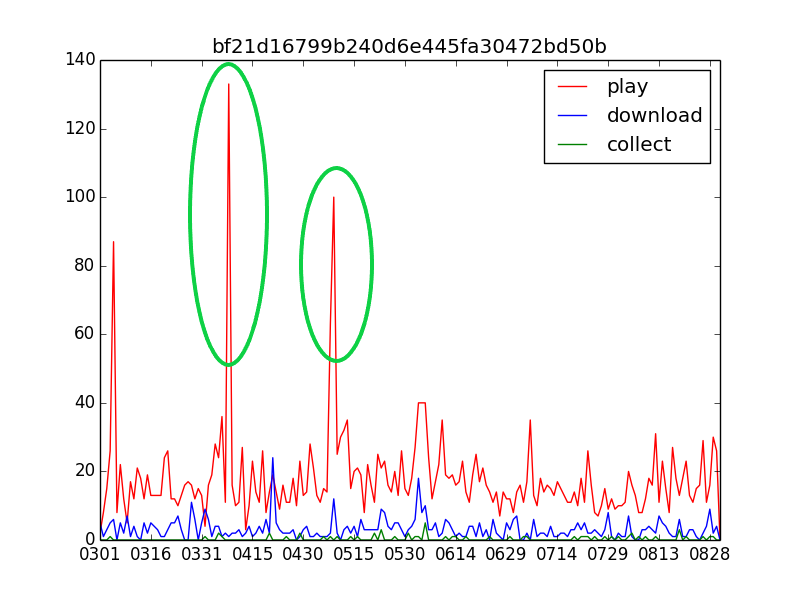
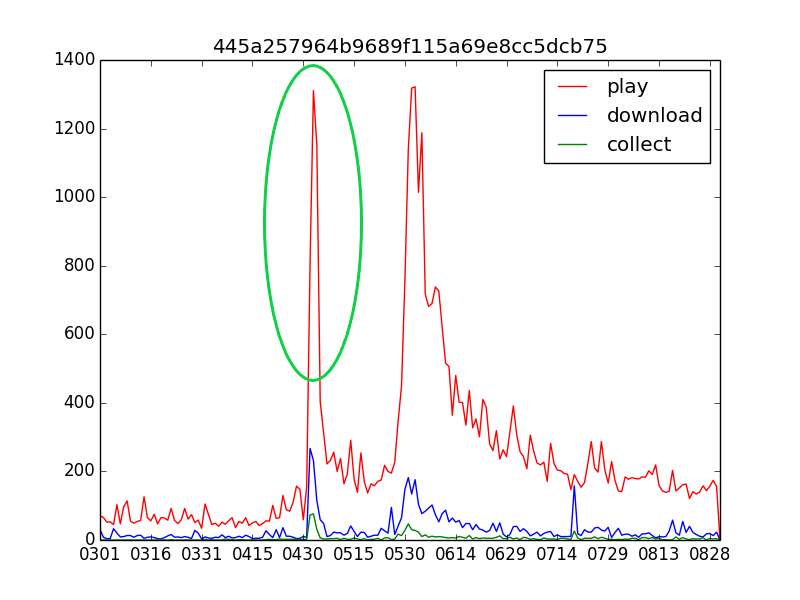
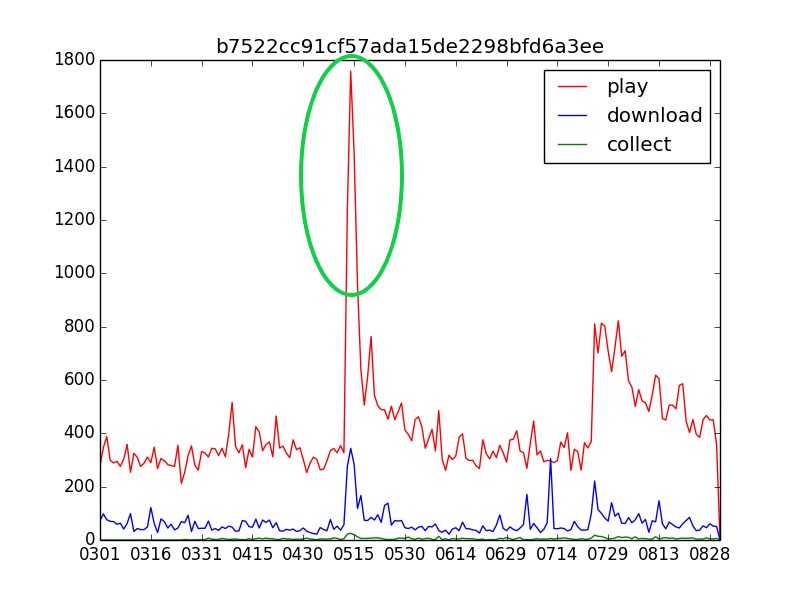
数据质量分析是数据挖掘和数据分析的重要开头，正确有效的数据才能挖掘出真正隐藏的信息，否则会导致很严重的损失。数据质量分析的目的是保证数据的正确性和有效性。有句习语是“Garbage in, garbage out.”，意思是输入的垃圾，得到的也是垃圾。说明了输出质量是由输入质量决定的，直接反映了数据质量分析的重要性。

在数据正确性分析方面，主要包括以下几方面内容：  
 （1）缺失值：缺少数据包括空值或者编码为无意义的值（null值）。面对缺失值，我们一般有三种解决方法：删除记录；数据插补；不处理。而在本次实验数据中，缺失值产生的问题很多。有些信息无法获取，有些信息采集的时候就不完整，这种情况下的缺失值根本不可考究，也许根本不存在。在给定的数据中，song\_id与artist\_id是我们的重点数据，如果这两个关键数值缺失那么该条记录其他内容就没有意义。所以针对这种缺失我们直接删除记录。在删除了缺失artist\_id的条目下，歌曲信息中只剩下了9656条记录，删除掉了缺失关键值的数据条目。

（2）异常值：异常值分析是检测数据中是否存在着采集的错误，是否有违背了事实的数据，这些异常值会对实验产生巨大误差。在处理异常值的时候，最简单粗暴的方法就是直接删除异常值的记录。然而缺点也很明显，在数据集本来就缺乏的情况下，不但会造成严重的样本量不足，还极大影响了分析的客观性和正确性，也有可能改变变量的原有分布，产生对分析结果的误差。对异常值的处理通常有三种方法：

简单统计量分析：首先对变量进行描述性的分析；通常用最大值与最小值进行判断。在本文中，一首歌的平均时间在3-5分钟。那么一个用户在一个小时之内最多听20次同一首歌曲，一日内最多播放该歌曲480次。那么我们可以给用户-歌曲表中某日的播放量设定一个阈值。譬如2000，即max=2000，如果播放量超过了2000，那么我们就标记为异常值，进行删除处理。

3σ原则：服从正太分布的数据，在该原则下，异常值被定义为一组测定值中与平均值的偏差超过三倍标准差的值。因为在正太分布下，距离平均值3σ之外的值出现的概率为不超过0.003，属于个别的极小概率事件。在本文所采用的数据中，也存在许多播放量突然呈爆炸式增长，然后突然迅速回落的例子。虽然音乐播放量并不是严格遵循正太分布，同时产生这种爆炸式的增长并不一定就是异常值，但是即便不是异常值，产生这种增长曲线的原因也是不可预测的，是毫无规律可循的，我们可以将它们也当成异常值进行处理。下图为典型播放量异常增长的艺人示例（画圈部分为异常）：



箱型图分析：箱型图又称盒须图，盒式图，箱线图，是一种数据分散情况资料的图。而在本文实验目标中，预测主要以回归为主，所以并不适合用箱型图分析所给数据。

### 3.2.2 数据特征分析

质量分析结束后，我们得到了相对来说精确可信的数据，可以通过利用现有工具绘制图表，计算特征量的方法进行特征分析。利用python的pandas库及matplotlib库可以绘制出我们想要的图表。通常来讲，我们从两个方面来考虑选择特征。一、特征是否发散，如果一个特征不发散，例如方差接近与0，也就是说样本在这个特征上基本没有差异，这个特征对于样本的区分并没有什么帮助；二、特征与目标的相关性：这点显而易见，与目标相关性高的特征应当优先选择。

而根据特征形式，又可以将特征选择方法分为三类：

Filter（过滤法），按照发散性或者相关性对各个特征进行评分，设定阈值或者选择阈值个数进而选择特征；

Wrapper（包装法），根据目标函数，一般即为预测效果评分，每次选取若干特征，或排除若干特征；

Embedded（嵌入法），先使用某些机器学习算法和模型进行训练，得到各个特征的权值系数，根据从小到大选择特征，类似于Filter方法但是通过训练来确定特征对优劣。

## 3.3 数据预处理

### 3.3.1 数据清洗

根据数据质量分析的结果，我们要对原始数据中的脏数据进行积累成操作，主要为删除重复数据、清除无关数据、删除缺失值、修改异常值、平滑噪声数据。

对于重复数据与缺失值，在前文中分析过，我们可以用SQL语句进行删除，重复数据和缺失值数据都主要存在于歌曲信息表中的歌曲id与艺人id之间。

对数据进行初步的分析后，发现在用户行为表中虽然有三种操作，但是收藏操作有很多空值，对于后续的建模变量输入并没有意义，下载量虽然质量稍好，但空余值数量也很多，这两项歌曲操作对播放量的预测并不能起到加成作用反而有可能有反作用，所以当作无意义的无关数据处理。歌曲信息表中，由于gender（性别），language（语言）选项一般与艺人id直接关联，同时对于使用者是否会播放歌曲的相关性非常低，并没有价值，当作无关数据处理。

在质量分析中我们对某天播放量有疑问的数据当作异常值处理，我们设置一个数值为2000的阈值，如果当天播放量超过2000，那么数据将统一修正为2000。

### 3.3.2 数据集成

数据集成是将多个数据源的数据整合在一起，形成一致的数据存储，如将不同的数据库中的数据集成到一个数据仓库中存储，数据集成主要涉及冗余处理和冲突数据检测和处理。在本文提供的两个输出文件中，因为要预测每天的音乐播放量，所以我们统计用户操作表后得到歌曲每日操作表songsSta，

|  |  |  |  |
| --- | --- | --- | --- |
| 字段名称 | 类型 | 字段长度 | 说明 |
| Song\_id | Varchar | 255 | 歌曲id |
| Num | Int | 11 | 播放数量 |
| Ds | Varchar | 255 | 统计当天日期 |

我们需要对歌曲每日操作表和歌曲信息表两个表格进行整合，得到新的统计表，艺人歌曲播放量表，

|  |  |  |  |
| --- | --- | --- | --- |
| 字段名称 | 类型 | 字段长度 | 说明 |
| Artist\_id | Varchar | 255 | 艺人id |
| Num | Int | 11 | 播放数量 |
| Ds | Varchar | 255 | 日期 |
| Id | Int | 11 | 本表主键 |

而收藏，下载，歌曲的性别，语言属性则已经简化，否则将会产生冗余问题，使得数据扁平化，拉低了神经网络模型的预测效果。

### 3.3.3 数据变换

在对数据进行统计分析时，要求数据必须满足一定条件，本文的数据变换主要是为了适用于挖掘任务与算法的需要。通常数据变换的方法主要有简单函数变换，最大规范化，均值规范化，小数定标规范化，连续属性离散化，属性构造，小波变换等等；在本文中，我们主要采用了规范化与属性构造的方法。

D:\desktop\论文图片\min-max.gif 规范化：ANN（人工神经网络）要将数据进行预处理，使用min-max normalization方法，公式如下：

后续在构建训练集中会用到这一公式对输入层的样本数据进行归一化处理。因为这些输入的样本数据可能在数值或者单位上存在很大的差距，从而导致输入层与隐层之间的权值、阈值很难进行调整。

属性构造：原始数据中的自带属性有艺人的性别，播放量，初始播放量，是否发布新歌，同时观察大部分艺人歌曲播放量发现有周期性，所以可以用某一天T的前七天作为一个周期，同时影响预测的还有预测日期与当日的时间间隔D，连续五天的艺人歌曲播放量S1,S2,..S5，艺人歌曲播放量的最大值，最小值，平均值。预测日期是否是周一，周二，周三…周日，这是时间因素的属性。预测日期是否是节假日也是时间属性。艺人最近是否发布了新的歌曲，根据图像的结果如果艺人发布新歌会引起播放量的增加等等，在后续中会着重说明。

## 3.4 本章小结

这一章介绍了本文的两个具体的数据文件来源，内容，大小与属性，还进行了数据分析的准备工作，充分的了解了数据，对数据进行了详细的分析，并进行了包括清洗，集成，变换在内的数据预处理流程，构造了与下一章节息息相关的模型所需要的属性特征。

# 4 人工神经网络算法在音乐播放量预测中的应用

## 4.1 ANN算法选择

人工神经网络是通过模拟动物的中枢神经系统来进行工作的。而我们则通过对一组输入的信号和输出的信号中间进行建模的过程，将输入的信息数据转换为输出的信息数据，这个过程是黑盒模糊处理的，所以又称之为“黑箱过程”。而它在函数逼近，模式识别，建模仿真等领域的应用取得显著成效。如今神经网络模型众多，而且他们的应用范围都存在较大差别，有点适用于模式识别，有的则是图像处理，等等。目前流行的有感知器神经网络，前馈神经网络，反馈网络，BP神经网络，卷积神经网络等。

在用人工神经网络对音乐的播放量进行预测的问题上，我们选取神经网络算法需要考虑一下几点：

1．歌手的歌曲试听量问题是一个时间序列问题，需要先进行非线性拟合再进行预测。

2.为了保障预测的速度，选取的网络模型不应该太复杂，预测时间应在可控范围内。

综合考虑上述的条件和对比各个流行的人工神经网络算法，最终决定使用BP神经网络算法对播放量进行预测。

## 4.2 数据准备

由于特征值通常较大，我们将数据规范化可降低较大值的影响，使用上一章的min-max 公式可将数据规范化再0-1之间，数据文件都保存为csv的格式。

划分数据集是训练模型前必须的一步，本文将数据集按照7：3的比例划分为训练集与测试集，通过python的sklearn库中的cross\_validation里面的train\_test\_split函数，可以将数据集按照一定比例随机划分为训练集与测试集，具体代码如下：

from sklearn.cross\_validation import train\_test\_split

#dat为数据集,含有feature和label.

train, test = train\_test\_split(dat, test\_size = 0.3)

划分完数据集之后就可以进行特征对提取。

## 4.3 确定输入元与输出元

在对歌手歌曲播放总量的预测过程中，我们应当合理的选取对歌手歌曲播放总量有代表性的因素，如果胡乱的选取过多的因素，会导致神经网络的输入神经元增多，隐层神经元也增加，导致神经网络模型结构变得非常复杂，延长训练时间，同时有可能降低预测的精准度，而选取的元素太少则有可能是模型无法精准预测，在前文大体的数据分析下，确定对歌手某一天的歌曲播放总量的影响因素有一下几方面：

1.预测日期与已知的最后日期之间的时间间隔G，假设我们要预测当前已经观测到的数据的日期T1之后的第n天的数据，那么作为实验输入的数据只能选择T1之前的数据，而第n天的预测结果与T1的日期间隔对预测的影响是很重要的，所以用时间间隔G作为一个特征和影响因素，定义G的值与范围如下：

若时间间隔在五天及五天以内，G=0.75；若时间间隔在五天到十五天包括十五天内，则G=0.5；若时间间隔在15天至30天之间包括30天，则G=0.25；若时间间隔大于30天，则G=0，此时时间间隔对预测播放量的预测关系已经趋近于无关。

2.连续五天内歌手歌曲总播放量，用S1-S5表示。在测试过连续三天，五天，一周的歌曲播放量进行预测后，根据训练效果和速度发现连续五天的播放量是最优的预测特征。

3.歌手总歌曲的试听量的最大值，最小值，均值。使用Smin，Smax，Smid来表示。

4.预测的日期是否是周一，周二…周日，这是时间属性的因素，可以使用week1-week7使用one-hot编码方式。

5.预测日期是否在节假日内，该因素同是时间属性。

6.歌手近期时间内是否发布了新的歌曲，新的歌曲通常会引起歌手歌曲播放量的增加。

上面共有18个影响播放量的因素，如下表所示：

|  |  |
| --- | --- |
| 字段 | 说明 |
| G | 时间间隔 |
| S1-S5 | 所预测日期T前G-G+4的播放量 |
| Smin,Smax,Smid | 预测日期T前试听总量的最小值最大值平均值 |
| Week1-week7 | 预测日期T是否为周一-周日 |
| F | 预测日期T是否为节假日 |
| W | 预测日趋T前七天是否有新歌发布 |

后面三个字段week1-week7，F,W都使用1或者0来表示，1代表“是”，0代表“否”。

所以建模的输入神经元应该有18个，而由于对结果的要求和算法复杂度的控制，输出神经元为1个,用X来表示预测日的播放量。

对训练集提取特征后的部分数据表如下：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 字段\序号 | 1 | 2 | … | 11126 | 11127 |
| G | 0 | 0 | … | 0 | 0 |
| S1 | 6848 | 6023 | … | 5098 | 5153 |
| S2 | 6443 | 6544 | … | 5354 | 5712 |
| S3 | 6894 | 6534 | … | 5578 | 5884 |
| S4 | 6786 | 8766 | … | 6098 | 6140 |
| S5 | 7980 | 6885 | … | 5560 | 6093 |
| Smin | 6443 | 6023 | … | 5098 | 5153 |
| Smax | 7980 | 8766 | … | 6098 | 6140 |
| Smid | 6990 | 6950 | … | 5537 | 5796 |
| Week1 | 0 | 0 | … | 0 | 0 |
| Week2 | 0 | 0 | … | 0 | 0 |
| Week3 | 0 | 0 | … | 1 | 0 |
| Week4 | 0 | 0 | … | 0 | 1 |
| Week5 | 1 | 0 | … | 0 | 0 |
| Week6 | 0 | 1 | … | 0 | 0 |
| Week7 | 0 | 0 | … | 0 | 0 |
| F | 0 | 0 | … | 0 | 0 |
| W | 0 | 0 | … | 0 | 0 |
| Y | 7023 | 6598 | … | 5137 | 5256 |

表中的输入层数据在实际使用时会通过前面提的min-max公式进行归一化处理，使其在区间（0，1）之间。

## 4.4 模型训练及隐层神经元确定

目前已经有理论证明只含有一个隐藏层的BP神经网络具有逼近任意非线性连续函数的能力，所以为了减少训练复杂度和训练时间，本文只使用了一层隐层神经元，隐层神经元计算公式如下：

D:\softwarefile\1826239165\Image\C2C\{V)(0AZ~_{%5IR_H0OI34_N.png

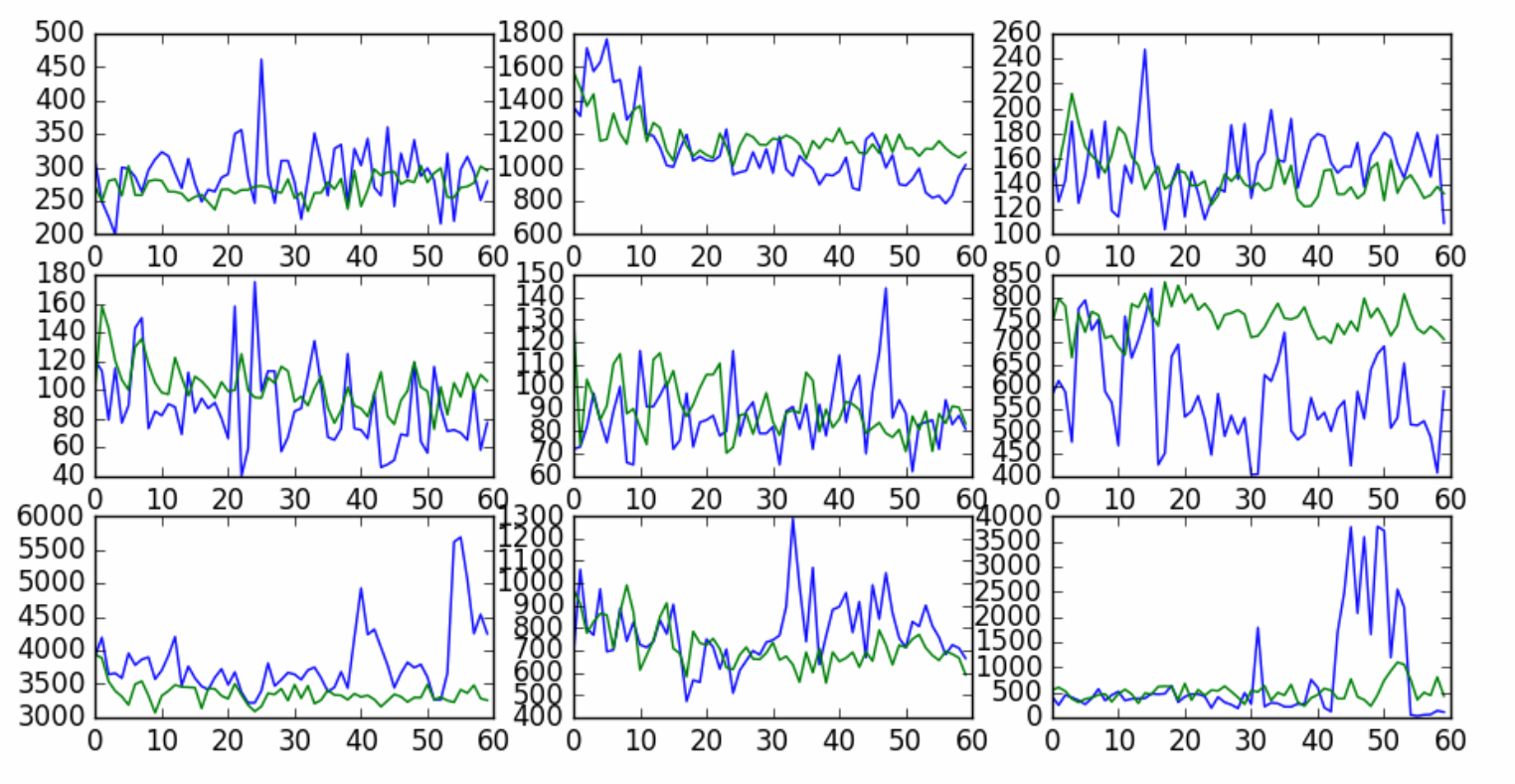
上面公式中d代表输入神经元个数，m为输出神经元个数。根据上面的公式我们的隐层神经元应该在[5,15]之间，我们采用试探法对隐层神经元分别为5，7，9，11，13时各进行10次训练，得到分别的手链时间与误差平方和如下表：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 隐层神经元个数 | 最快收敛速度 | 最慢收敛速度 | 平均收敛速度 | 误差平方和 |
| 5 | 115 | 235 | 177 | 11.89 |
| 7 | 123 | 224 | 180 | 11.32 |
| 9 | 112 | 235 | 197 | 11.12 |
| 11 | 133 | 244 | 204 | 11.01 |
| 13 | 136 | 246 | 212 | 10.96 |

根据表中数据，在收敛速度和误差率的平衡上考虑，决定隐层神经元个数为9个。

## 4.5 实验预测结果

通过前面的数据整理及分析，对该歌手的音乐播放量建立了一个输入元为18个，输出元为一个，只具有一层隐层神经元且隐层神经元9个的BP神经网络模型，使用此模型套入九名歌手的数据特征得到2015年8月1日至2015年8月30日的歌曲试听量图如下（绿色折线表示预测，蓝线表示实际）：

 可见预测数据大值还原了真实数据所表达的大体趋势和播放数量，甚至在某些小的波动上依然能够预测出来，而误差的关键因素还是神经网络模型的训练效果，隐层神经元数量，播放量的特征选取上。本文从实际出发，针对数据情况提取了特征，并对隐层节点进行试探法找到了最优的节点数进行了预测。

## 4.6 本章小结

本章严格的通过BP神经网络算法的步骤序列，对前文清理过后的数据进行了归一化处理，数据集的划分，特征提取，输入输出元，隐层神经元个数的确定，最后对不同歌手进行了音乐播放量的预测，与播放量的发展趋势基本吻合，实现了利用数据挖掘方法对流行音乐播放量的预测，对歌手歌曲是否会流行有了一定的把控，达到了对音乐流行趋势的预测。

# 5 结论

数据挖掘技术自被提出以来就不断的被补充和完善，而人工神经网络模型则是其皇冠上最亮眼的珍珠。几十上百种的神经网络模型被提出，被应用到各行各业，各个领域，潜移默化的提升着人民日常生活的质量和幸福度。而利用数据挖掘的技术对流行音乐进行预测则对人民的娱乐生活有着不凡的意义。既能推动音乐行业的发展，又对推动着科学技术的进步。

本文利用BP算法和已有数据对音乐的播放量进行了预测，大体上预测了流行趋势，然而还是有待精确，利用其他算法进行多次预测模型融合应该能进一步的提升预测的精确度，然而篇幅能力有限，并未过多赘述。在完成后有机会应当继续深入研究这个问题。

这次毕业设计对于我来说是一次挑战，又是一次学习、提升自我的机会，刚开始的时候对题目一筹莫展的我不断的查询资料，了解知识，学习了许多未曾接触过的概念。而通过这次毕业设计，我学到了许多东西，通过实践增强了自己的编码能力和信息搜集能力，让我感受到了知识和应用的联系，对我解决问题的能力有了很大的提升，以后面对问题任务都有经验信心去解决。

# 致谢

我首先要感谢我的论文指导老师，计算机学院的王然老师。王老师对我论文的研究方向做出了指导性的意见和推荐，在论文撰写过程中及时对我遇到的困难和疑惑给予悉心指点，提出了许多有益的改善性意见，投入了许多的心血和精力。王老师对我的帮忙和关怀表示诚挚的谢意!同时，还要感谢计算机学院的授课老师们和所有同学们，大家在杭电的学习中互相学习，互相帮忙，共同度过了一段完美难忘的时光。

　　此外，还要感谢朋友以及同学们在论文编写中带给的大力支持和帮忙，给我带来极大的启发。也要感谢参考文献中的作者们，透过他们的研究文章，使我对研究课题有了很好的出发点。

　　最后，谢谢论文评阅老师们的辛苦工作。衷心感谢我的家人、朋友，以及同学们，正是在他们的鼓励和支持下我才得以顺利完成此论文。

# 参考文献

[1]刘晓曈. 基于深度学习的分类预测方法研究及应用[D].东南大学,2017.

[2]王惠文,孟洁.多元线性回归的预测建模方法[J].北京航空航天大学学报,2007(04):500-504.

[3]Hongxiao F., Li C., Jiabao H., Yanru X., Han L. (2017) SANet: An Approach for Prediction in Music Trends. In: Bhatia S., Mishra K., Tiwari S., Singh V. (eds) Advances in Computer and Computational Sciences. Advances in Intelligent Systems and Computing, vol 553. Springer, Singapore

[4]Making SVMs Scalable to Large Data Sets using Hierarchical Cluster Indexing[J] . Hwanjo Yu,Jiong Yang,Jiawei Han,Xiaolei Li. Data Mining and Knowledge Discovery . 2005 (3)

[5]Python数据分析与挖掘实战[M]. 机械工业出版社 , 张良均, 2015

[6]基于神经网络算法构建电信用户流失预测模型的研究[D]. 孙碧颖.兰州大学 2016

[7]吕倩倩. 基于机器学习的音乐流行趋势预测[D].兰州大学,2017.

[8]人工神经网络在股票预测中的应用[D]. 胡照跃.中北大学 2016

[9] Neuromorphic architecture for adaptive robort control:A preliminary analysis. Guez A,Elibert J,Kan M. Proc.IJCNN‘ 87 . 1987

[10]Vapnik, V. & Izmailov, R. Ann Math Artif Intell (2017) 81: 3. https://doi.org/10.1007/s10472-017-9538-x

[11]Learning multiple layers of representation[J] . Geoffrey E. Hinton. Trends in Cognitive Sciences . 2007 (10)

[12]高速公路软土地基沉降量的人工神经网络预测[D]. 赵亚哥白.东北林业大学 2006

[13]数据挖掘中的新方法[M]. 科学出版社 , 邓乃扬,田英杰著, 2004

# 附录