机器学习工程师纳米学位毕业项目
基于语音识别的性别识别

杨昱

2017 年 2 月 18 日

1.定义

1.1 项目概述

问题来自于 Kaggle。[1]项目训练一个模型，使之能够根据声音识别说话人的性别。

语音识别是一个热门领域，人们一直希望有朝一日计算机能够直接识别语音，从而能够使用语言和机器进行交流。计算机的输入方式不断演变，从最初的带孔的卡片，到现在键鼠的输入方式，人机交流方式越来越友好。现在，人们正在尝试教机器识别自然语言，这样大众用户在操作机器时就可以摆脱遥控器等设备，进一步解放人类的双手。

在语音识别和说话人识别实验中发现，事先知道说话人性别时所得到的正确识别率要比不知道说话人性别时高（邓英（2003）[2]）。所以根据语音识别性别对语音识别技术具有一定意义。

项目使用支持向量机，决策树，随机森林算法为基础训练预测模型，并注重调试优化支持向量机，使模型的准确率接近人类。项目使用的数据集来自真实的男女人声，并经过声学分析处理，保存了声音频率在 0-280hz。数据集一共包括关于声音频率的 20 个特征和一个性别标签。原始声音文件来自于哈弗大学等机构的数据库。

1.2 问题陈述

根据声音识别一个人的性别对人来说是一个很容易的事情，人可以通过仅仅数个词语，或者语音片段就可以判断出说话的人是男是女。然而，设计实现一个计算机程序来完成相同的事情却有些棘手。

语音性别识别的目标是根据一段语音的特征识别出说话人的性别。

数据集包含 20 个特征作为备选特征和一个二值标签，这意味着可以使用监督学习算法来解决这个问题。

1.3 评价指标

评价指标选用的是准确率（accuracy），即识别正确的样本数量在总体样本中的比例，用公式表达：

$$准确率=\frac{预测正确的样本数}{样本总数}$$

2 分析

2.1 数据研究

数据集一共包括 3168 条记录，以 csv 格式保存。其中部分数据截图如图 1 所示

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
meanfreq	sd	median	Q25	Q75	IQR	skew	kurt	sp.ent	sfm	mode	centroid	meanfun	minfun	maxfun	meandom	mindom	maxdom	dfrange	modindx	label
0.059781	0.064241	0.032027	0.015071	0.090193	0.075122	12.86346	274.4029	0.893369	0.491918	0	0.059781	0.084279	0.015702	0.275862	0.007813	0.007813	0.007813	0	0	male
0.066009	0.06731	0.040229	0.019414	0.092666	0.073252	22.42329	634.6139	0.892193	0.513724	0	0.066009	0.107937	0.015826	0.25	0.009014	0.007813	0.054688	0.046875	0.052632	male
0.077316	0.083829	0.036718	0.008701	0.131908	0.123207	30.75715	1024.928	0.846389	0.478905	0	0.077316	0.098706	0.015656	0.271186	0.00799	0.007813	0.015625	0.007813	0.046512	male
0.151228	0.072111	0.158011	0.096582	0.207955	0.111374	1.232831	4.177296	0.963322	0.727232	0.083878	0.151228	0.088965	0.017798	0.25	0.201497	0.007813	0.5625	0.554688	0.247119	male
0.13512	0.079146	0.124656	0.07872	0.206045	0.127325	1.101174	4.333713	0.971955	0.783568	0.104261	0.13512	0.106398	0.016931	0.266667	0.712813	0.007813	5.484375	5.476563	0.208274	male
0.132786	0.079557	0.11909	0.067958	0.209592	0.141634	1.932562	8.308895	0.963181	0.738307	0.112555	0.132786	0.110132	0.017112	0.253968	0.298222	0.007813	2.726563	2.71875	0.12516	male
0.150762	0.074463	0.160106	0.092899	0.205718	0.112819	1.530643	5.987498	0.967573	0.762638	0.086197	0.150762	0.105945	0.02623	0.266667	0.47962	0.007813	5.3125	5.304688	0.123992	male
0.160514	0.076767	0.144337	0.110532	0.231962	0.12143	1.397156	4.766611	0.959255	0.719858	0.128324	0.160514	0.093052	0.017758	0.144144	0.301339	0.007813	0.539063	0.53125	0.283937	male
0.142239	0.078018	0.138587	0.088206	0.208587	0.120381	1.099746	4.070284	0.970723	0.770992	0.219103	0.142239	0.096729	0.017957	0.25	0.336476	0.007813	2.164063	2.15625	0.148272	male
0.134329	0.08035	0.121451	0.07558	0.201957	0.126377	1.190368	4.78731	0.975246	0.804505	0.011699	0.134329	0.105881	0.0193	0.262295	0.340365	0.015625	4.695313	4.679688	0.08992	male
0.157021	0.071943	0.16816	0.10143	0.21674	0.11531	0.979442	3.974223	0.965249	0.733693	0.096358	0.157021	0.088894	0.022069	0.117647	0.460227	0.007813	2.8125	2.804688	0.2	male
0.138551	0.077054	0.127527	0.087314	0.202739	0.115426	1.62677	6.291365	0.966004	0.752042	0.012101	0.138551	0.104199	0.019139	0.262295	0.246094	0.007813	2.71875	2.710938	0.132351	male
0.137343	0.080877	0.124263	0.083145	0.209227	0.126082	1.378728	5.008952	0.963514	0.73615	0.108434	0.137343	0.092644	0.016789	0.213333	0.481671	0.015625	5.015625	5	0.0885	male

图 1 数据集部分数据展示

所有除标签（label）之外，所有的特征的值的类型都是数字类型。而且其中没有空值，即没有数据丢失。统计男女样本人数均为 1584 人，两种样本人数相等，按性别为两部分。所以，在使用数据前需要将数据顺序打乱，除此之外不需要对数据集进行额外的预处理。

2.2 数据可视化

为了进一步查看特征之间的关系，将两个特征作为一组，遍历所有的特征组合，使用 python 的 matplotlib 库输出性别关于这些特征组合的二维分布图。一共得到 210 副图，通过观察这 210 副图像，可以看出平均基音频率（meanfun）是一个非常有效的特征，任何特征和平均基音频率组合都能够将男女分为两个明显的簇。如图二所示。

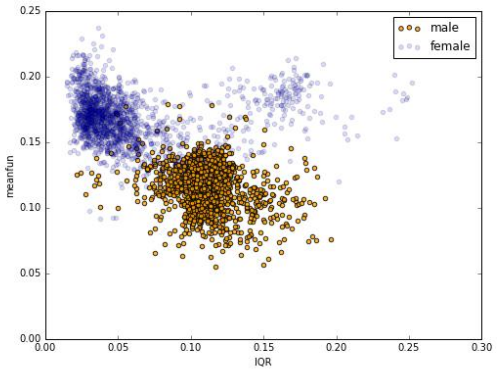


图 2 性别频率四分位间距（IQR）和关于平均基音频率（MeanFun）的分布

其他大多数图像，男女样本分布区分并不明显。以图 3 为例

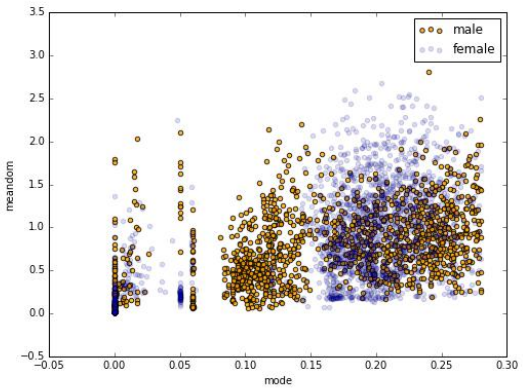


图 3 性别关于频率众数（mode）和平均频率（meandom）

从图 3 中可以看出，虽然有大约一半的男性样本单独集中在图像中部，可是剩余另一半与女性样本交叉严重。

以数字方式显示各特征之间的关系，即相关系数，如图 4.

	meanfreq	sd	median	Q25	Q75	IQR	skew	kurt	sp.ent	sfm	mode	centroid	meanfun	minfun	maxfun	meandom	mindom	maxdom	dfrange	modindx
meanfreq	1	-0.73904	0.925445	0.911416	0.740997	-0.62761	-0.32233	-0.31604	-0.6012	-0.78433	0.687715	1	0.460844	0.383937	0.274004	0.536666	0.229261	0.519528	0.51557	-0.21698
sd	-0.73904	1	-0.5626	-0.84693	-0.16108	0.87466	0.314597	0.346241	0.71662	0.838086	-0.52915	-0.73904	-0.46628	-0.34561	-0.12966	-0.48273	-0.35767	-0.48228	-0.476	0.12266
median	0.925445	-0.5626	1	0.774922	0.731849	-0.47735	-0.25741	-0.24338	-0.502	-0.66169	0.677433	0.925445	0.414909	0.337602	0.251328	0.455943	0.191169	0.438919	0.435621	-0.2133
Q25	0.911416	-0.84693	0.774922	1	0.47714	-0.87419	-0.31948	-0.35018	-0.64813	-0.76687	0.591277	0.911416	0.545035	0.320994	0.199841	0.467403	0.302255	0.459683	0.454394	-0.14138
Q75	0.740997	-0.16108	0.731849	0.47714	1	0.009636	-0.20634	-0.14888	-0.17491	-0.3782	0.486857	0.740997	0.155091	0.258002	0.285584	0.359181	-0.02375	0.335114	0.335648	-0.21647
IQR	-0.62761	0.87466	-0.47735	-0.87419	0.009636	1	0.249497	0.316185	0.640813	0.663601	-0.40376	-0.62761	-0.53446	-0.22268	-0.06959	-0.33336	-0.35704	-0.33788	-0.33156	0.041252
skew	-0.32233	0.314597	-0.25741	-0.31948	-0.20634	0.249497	1	0.97702	-0.19546	0.079694	-0.43486	-0.32233	-0.16767	-0.21695	-0.08086	-0.33685	-0.06161	-0.30565	-0.30464	-0.16932
kurt	-0.31604	0.346241	-0.24338	-0.35018	-0.14888	0.316185	0.97702	1	-0.12764	0.109884	-0.40672	-0.31604	-0.19456	-0.2032	-0.04567	-0.30323	-0.10331	-0.2745	-0.27273	-0.20554
sp.ent	-0.6012	0.71662	-0.502	-0.64813	-0.17491	0.640813	-0.19546	-0.12764	1	0.866411	-0.3253	-0.6012	-0.51319	-0.30583	-0.12074	-0.29356	-0.29487	-0.32425	-0.31905	0.198074
sfm	-0.78433	0.838086	-0.66169	-0.76687	-0.3782	0.663601	0.079694	0.109884	0.866411	1	-0.48591	-0.78433	-0.42107	-0.3621	-0.19237	-0.42844	-0.28959	-0.43665	-0.43158	0.211477
mode	0.687715	-0.52915	0.677433	0.591277	0.486857	-0.40376	-0.43486	-0.40672	-0.3253	-0.48591	1	0.687715	0.324771	0.385467	0.172329	0.491479	0.19815	0.477187	0.473775	-0.18234
centroid	1	-0.73904	0.925445	0.911416	0.740997	-0.62761	-0.32233	-0.31604	-0.6012	-0.78433	0.687715	1	0.460844	0.383937	0.274004	0.536666	0.229261	0.519528	0.51557	-0.21698
meanfun	0.460844	-0.46628	0.414909	0.545035	0.155091	-0.53446	-0.16767	-0.19456	-0.51319	-0.42107	0.324771	0.460844	1	0.339387	0.31195	0.27084	0.162163	0.277982	0.275154	-0.05486
minfun	0.383937	-0.34561	0.337602	0.320994	0.258002	-0.22268	-0.21695	-0.2032	-0.30583	-0.3621	0.385467	0.383937	0.339387	1	0.213987	0.375979	0.082015	0.31786	0.316486	0.002042
maxfun	0.274004	-0.12966	0.251328	0.199841	0.285584	-0.06959	-0.08086	-0.04567	-0.12074	-0.19237	0.172329	0.274004	0.31195	0.213987	1	0.337553	-0.24343	0.35539	0.35988	-0.36303
meandom	0.536666	-0.48273	0.455943	0.467403	0.359181	-0.33336	-0.33685	-0.30323	-0.29356	-0.42844	0.491479	0.536666	0.27084	0.375979	0.337553	1	0.099656	0.812838	0.811304	-0.18095
mindom	0.229261	-0.35767	0.191169	0.302255	-0.02375	-0.35704	-0.06161	-0.10331	-0.29487	-0.28959	0.19815	0.229261	0.162163	0.082015	-0.24343	0.099656	1	0.02664	0.008666	0.200212
maxdom	0.519528	-0.48228	0.438919	0.459683	0.335114	-0.33788	-0.30565	-0.2745	-0.32425	-0.43665	0.477187	0.519528	0.277982	0.31786	0.35539	0.812838	0.02664	1	0.999838	-0.42553
dfrange	0.51557	-0.476	0.435621	0.454394	0.335648	-0.33156	-0.30464	-0.27273	-0.31905	-0.43158	0.473775	0.51557	0.275154	0.316486	0.35988	0.811304	0.008666	0.999838	1	-0.42927
modindx	-0.21698	0.12266	-0.2133	-0.14138	-0.21647	0.041252	-0.16932	-0.20554	0.198074	0.211477	-0.18234	-0.21698	-0.05486	0.002042	-0.36303	-0.18095	0.200212	-0.42553	-0.42927	1

图 4 特征之间的相关系数

2.2 算法

根据前面的分析可知，语音性别分类本质上是一个分类问题。用于分类的机器学习算法有很多，比如逻辑回归、支持向量机、决策树、Boosting 等。这些算法统称为机器学习算法。算法的输入是选取的特征组合构成了特征向量；输出的是性别；整个数据集是特征空间；特征数量时空间的维度。

2.2.1 决策树

决策树是一种常见的机器学习算法。它基于树结构来进行决策，与人类在面临决策问题时的思考方式类似。举例来解释，要对“这个西瓜好吗？”这样的问题进行决策时，通常会进行一系列的判或“子决策”：先看瓜的颜色，如果是“青绿色”，那么再看根蒂的形态，如果是“蜷缩”，那么可以得出好瓜的结论。（周志华 2016）[3]

优点：

- 使用决策树不需要对数据进行太多的预处理，例如标准化，能够减少计算量。

- 决策树的可解释性前，模型的结果也可以进行可视化。
- 能够处理多输出问题。

缺点：

- 决策树最主要的问题是容易过拟合，这意味着泛化能力差。
- 决策树有一定的不稳定性。对同一个数据集，每次运行决策树算法产生的结果可能不相同。[4]

2.2.2 支持向量机

支持向量机，SVM，也是通过求解超平面，从而划分样本类别。不同的是，SVM 要求样本点距离超平面距离尽量远，使得超平面具有一定稳健性。SVM 经过理论论证能够应对能多情况，并且取的比较理想的效果。支持向量机支持核技巧。核技巧隐式地把数据映射到高维，然后在高维求解分类超平面，这等价于在原有空间求解一个分类的曲面。支持向量机加上核技巧能够解决许多非线性问题。

优点：

- 在高维空间有效。
- 在尺寸数量大于样本数量的情况下仍然有效。

缺点：

- 如果特征的数目远大于样本的数目，则该方法可能给出差的性能。[4]

2.2.3 随机森林

随机森林顾名思义，是用随机的方式建立一个森林，森林里面有很多的决策树组成，随机森林的每一棵决策树之间是没有关联的。在得到森林之后，当有一个新的输入样本进入的时候，就让森林中的每一棵决策树分别进行一下判断，看看这个样本应该属于哪一类（对于分类算法），然后看看哪一类被选择最多，就预测这个样本为那一类。

优点：

- 能够处理多维数据，并且不用做特征选择。
- 和决策树类似，随机森林的最终模型也是可以解释的。

2.3 技术

项目中会使用 scikits.learn（sklearn）函数库。Sklearn 是一个基于 python 的免费机器学习算法库。它包含了多种机器学习算法，例如，支持向量机，随机森林，k-means，可用于分类、回归、聚类问题。[4]

scikit-learn 项目开始为 scikits.learn，David Cournapeau 的 Google Summer of Code 项目。 它的名字来源于它是一个“SciKit”（SciPy 工具包），一个单独开发和分布式第三方扩展 SciPy 的概念。 原来的代码库后来被其他开发人员重写。

Sklearn 与 numpy 和 pandas 库组合可以非常快速建立一个机器学习模型。并且它还提供了多种函数辅助为机器学习算法调试参数。

2.4 基准模型

在 kaggle 的问题描述页上，给出了到目前为止不同模型最好的成绩。

监督学习算法	训练集准确率	测试集准确率
BaseLine(总预测男性)	50%	50%
Logistic Regression	97%	98%
CART	96%	97%
Random Forest	100%	98%
SVM	100%	99%
XGBoost	100%	99%

表格 2

所有的机器学习算法训练出的最终模型都得到了近乎完美的准确率。Kaggle 上提供的一个快速入门，使用决策树以 meanfun 和 IQR 为区分特征，实现了 96.0%的准确率。因为 96.0%已经是一个非常高的准确率，已经非常接近人根据声音识别性别的准确率，所以本项目尝试再将准确率提高 0.5%，也就是将准确率提高到 96.5%。

3 具体方法

3.1 数据预处理

数据集中，label 栏下的数据是字符串形式的，为了便于处理，需要将这些字符串映射成数字，’ male’映射到 1，‘female’映射到 2。

在划分训练集和测试集之前，需要将数据进行随机打乱，因为原始数据文件中，数据是按照性别排序的，不具有随机性。在将数据随机打乱后，将其中 20%的数据作为测试集，80%的数据作为训练集。为了评估模型的泛化能力，使用交叉验证。即将数据分为 5 组互斥子集，每组占总数据的 20%，每次使用一组数据作为测试集，剩下 4 组做训练集，重复 5 次，使得每一组数据都充当过测试集的角色。取 5 个测试集最终准确率的平均数作为最终准确率。

对于本实验，数据集一共包括 3168 条数据，数据集的 20%也就是 634 条数据，也就是测试集包括 634 条数据，训练集包括 2534 条数据。

3.2 实现

3.2.1 决策树与随机森林

实现一个简单的决策树，查看预测准确率。从 sklearn 库中引入决策树模块，设置 random_state 为 1，训练决策树。查看训练结果，测试集准确率为 95.8%，训练集的准确率为 1.0。从测试集的准确率上看，着已经非常接近实验的目标 96.5%了。

同理，训练一个随机森林模型，参数设置为默认参数。查看模型在测试集上的表现，准确率达到 98%，这一个非常高的分数，和 kaggle 上最佳决策树模型的准确率几乎相同。

3.2.2 支持向量机

和训练决策树模型相似，首先实现一个简单的支持向量机模型，使用 sklearn 库中的 SVC 模块。在不改变其默认参数情况下，使用训练集进行训练一个模型。这个模型在测试集中达到了 74%的准确率，相对决策树和随机森林模型来说低很多。

3.3 改进

因为决策树和随机森林的表现都很好，对这两种算法进行改进难以看出改进效果，所以本文主要研究支持向量机的改进方法。

可以考虑从不同角度优化支持向量机，本项目主要从两方面优化支持向量机，数据和参数方面。

首先，从数据角度。这里使用的方式是标准化。标准化是指将每个属性的分布偏移为具有零的平均值和标准偏差为 1（单位方差）。它对标准化模型的属性很有用。 数据集的标准化是在 scikit-learn 中实现的许多机器学习估计器的常见要求;，如果单个功能没有或多或少看起来像标准的正态分布数据，它们可能表现不好。

对数据标准化之后，再一次训练模型，查看训练结果。出乎意料，这一次支持向量机预测准确率达到了 98%。使用交叉验证查看预测准确率为 96%。

方案	训练集准确率	测试集准确率	交叉验证准确率
标准化	98.3%	98.4%	96.7%

表格 3

从表格中的数据可以看出，交叉验证的准确率要明显低于使用一个测试集和训练集时的模型准确率。这是因为交叉验证将数据集分成了多个不同的训练集和测试集，并将模型在所有测试集上的准确率的平均值作为最终结果。因此交叉验证的准确率会低一些，不过这也使得交叉验证的准确率更接近模型的真实准确率。

接下来调试支持向量机的参数，查看当核函数（kernel）不同时，交叉验证的准确率。使用的数据还是经过标准化处理的数据。

方案		交叉验证
kernel='linear'		96.71%
kernel='rbf'		96.79%
kernel='poly'		93.39%

表格 4 采用不同核函数支持向量机的表现

从表 4 中可以看出，采用’linear’和’rbf’和函数的模型表现比较好，而使用’poly’核函数的模型相对差一些。所以接下来在调试其他参数时，核函数（kernel）均使用 rbf。

除了核函数之外，支持向量机还有两个重要参数，C、gamma。首先调试参数 C。

参数 C 相当于松弛变量，当 C 的值越大，对错误分类的惩罚越严重，此时模型趋向于训练集全分类对的情况。这时，训练集的准确率很高，不过模型的泛化能力会下降，也就是倾向于过拟合。C 的值越小，对错误分类的惩罚越小，允许出错，将一些分类出错的点当成噪声忽略。

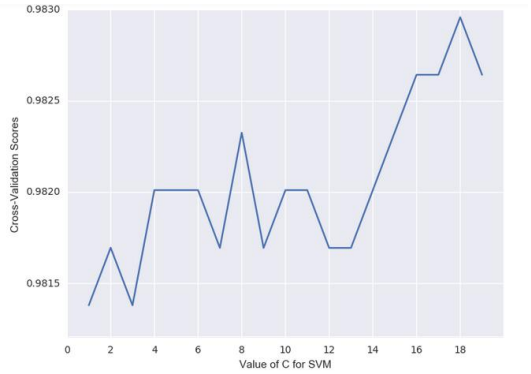


图 5 参数 C 取值 1-20 时的准确率

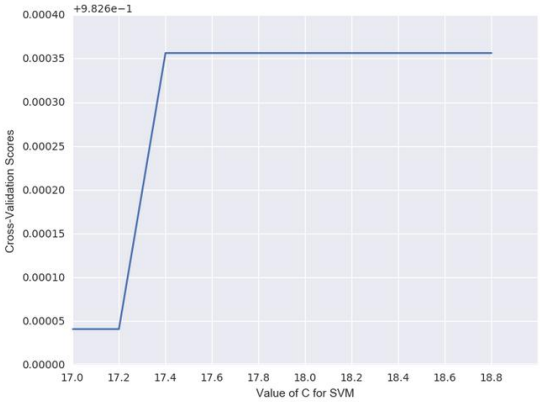
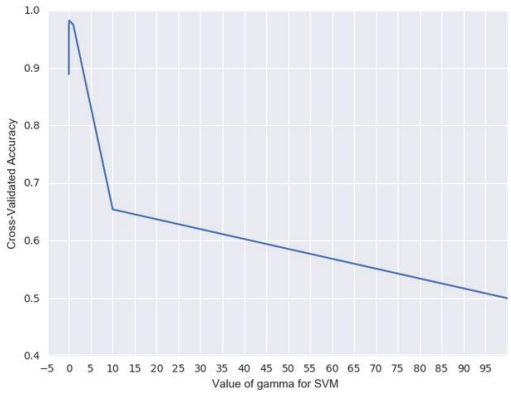


图 6 参数 C 取值 17.0-19.0 时的准确率

在通过图 5，可以观察到模型准确率在 C=18 时达到最高，接下来将尝试使用更小的步长，查看 C 取值在 17.0 和 19.0 之间时，什么时候能够取得模型准确率最大值。如图 6 所示，当 C 取值 17.4 时，模型准确率最高。

gamma 参数是支持向量机众多参数中出参数 C 之外另一个比较重要的参数。Gamma 参数是 rbf 核函数（高斯函数）的标准差的倒数。gamma 的值用于评估两个点的相似性。小的 gamma 值意味着模型有一个大的方差。在这种情况下，即使两个点相距较远也会被认为是相似的。另一方面，大的 gamma 值以为一个小方差的模型，这种情况下，当两个点距离比较近时才会被认为是相似的。



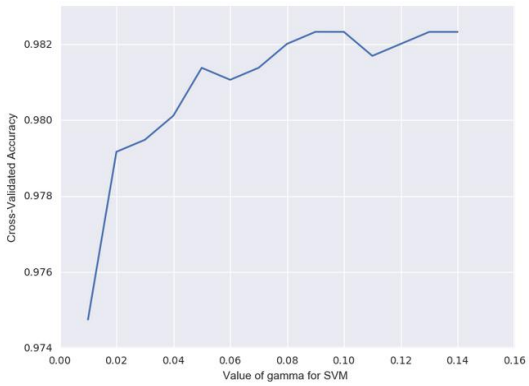


图 7　gamma 取值变换大时的准确率　　图 8 gamma 取 0.10 左右时的准确率

通过对图 7 的观察，可以发现，当 gamma 取 0.001 到 100 之间，模型在 gamma 等于 0.1 左右时，准确率达到最高。图 8 显示，当 gamma 等于 0.090 时，模型准确率达到最高值，98.20%。

最后，进行整体调试支持向量机的参数，通过上面的探索，可以确定，最优参数 C 的取值在 16.0 到 19.0 之间，而参数 gamma 的最优取值在 0.010 到 0.20 之间，核函数可以选用的有”linear”和”rbf”。所以，可以使用网格搜索，选取一个最优的参数组合。最终，选用”rbf”核函数，参数 C 取 1.4，参数 gamma 取 0.30 时，模型准确率最高，交叉验证准确率为 98.32%

4 结果

4.1 模型评估

使用上述讨论得出的最优参数组合{kernel=’rbf’, C=’16.5’, gamma=0.06}训练一个支持向量机的模型，并输出学习曲线，即准确率与训练集规模的关系。

训练集中样本至少包含 250 个样本，最多不超过 2500 个样本。从图 9 中可以看出，测试集的准确率随样本数目的增长一直增长，同时增长速率不断减小，最终趋于平缓。训练集的准确率在前半段随样本数目的增长而上升，后半段有些波动。整体来看，训练集的准确率和测试集的准确率在不断接近。图像中没有显示训练集的准确率最终远超过测试集准确率，所以不认为有过渡拟合和趋势。

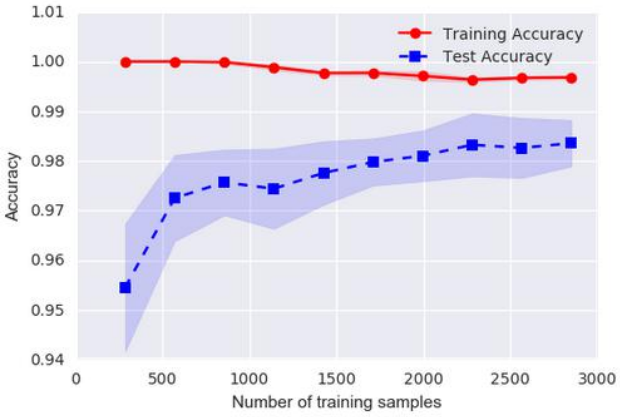


图 9 学习曲线

4.2 结果分析

在数据集中随机选取 30 个样本，将数据标准化后，使用训练好的支持向量机模型进行预测。部分数据如图 10 所示，预测结果如图 11 所示。通过结果图，可以看出随机选的小规模样本预测结果与真实性别一样。

M	N	O	P	Q	R	S	T	U
meanfun	minfun	maxfun	meandom	mindom	maxdom	dfrange	modindx	label
0.190716	0.049231	0.277457	1.272321	0.023438	11.17969	11.15625	0.058357	female
0.172607	0.084656	0.246154	0.859375	0.171875	6.0625	5.890625	0.221003	female
0.185592	0.019656	0.242424	0.30816	0.007813	0.90625	0.898438	0.280978	female
0.095423	0.056497	0.212766	0.519064	0.087891	0.776367	0.688477	0.65087	male
0.119749	0.024814	0.277778	0.384803	0.004883	0.805664	0.800781	0.551575	male
0.149622	0.039526	0.188679	0.354448	0.004883	1.450195	1.445313	0.239489	female
0.118711	0.047105	0.277457	0.89446	0.023438	7.21875	7.195313	0.061335	male
0.123986	0.047059	0.275862	1.40625	0.023438	7.875	7.851563	0.111685	male
0.109296	0.030769	0.217391	0.862109	0.117188	3.950195	3.833008	0.264522	male
0.13042	0.048534	0.27907	0.714844	0.023438	5.34375	5.320313	0.122991	male

图 10 随机选的 30 个测试样本

	1	2	3	4	5	6	7	8	9	10
真实性别	female	female	female	male	male	female	male	male	male	male
预测性别	female	female	female	male	male	female	male	male	male	male
	11	12	13	14	15	16	17	18	19	20
真实性别	male	female	female	female	male	male	female	male	female	male
预测性别	male	female	female	female	male	male	female	male	female	male
	21	22	23	24	25	26	27	28	29	30
真实性别	male	male	female	female	female	male	female	male	male	female
预测性别	male	male	female	female	female	male	female	male	male	female

图 11 测试的结果图。数字是样本编号，编号下第一行是样本的真实性别，第二行是模型预测的性别。

将错误分类的样本保存成一个单独的文件，然后查看这些错误分类的样本在数据集中是怎样分布的。图 12 是样本点关于平均频率（meanfreq）、频率四分位间距（IQR）和平均基音频率（meanfun）的分布图。

图 12 红色的点代表正确分类的男性，绿色点代表正确分类的女性。青色方块代表被模型错分为女性的样本，蓝色菱形代表被模型错分为男性的样本。

通过图 12 可以观察到，被错误分类的样本分布集中，主要位于两个簇的交界地区，也就是声音特征比较倾向于中性。因此导致模型难以区分这些人的应该归到哪一类。还有少量样本落到异性簇的中间，使得模型进行了错误的分类。实际上，在现实生活中，有一些人的声音确实很特殊，使得我们仅仅通过声音无法判断说话人的性别。所以模型对极少量的样本分类错误也是正常的。

通过以上讨论，最终训练出来的支持向量机在交叉验证中达到了 98.38%的准确率，已经符合了项目最初要求。能够保证在绝大多数情况下正确识别出说话人的性别，虽然没有达到最优的识别率 99%，但是已经非常接近了。

5 结论

5.1 总结

本文实现了一个根据语音数据识别性别，准确率为98.35%的支持向量机模型。在数据预处理的步骤中，通过将数据标准化，极大地提高了模型的识别率，验证了标准化数据对支持向量机来说非常重要。然后为了进一步提高模型预测的精准度，找到参数C和参数gamma的最佳范围，而后使用网格搜索算法找到了参数C和参数gamma的最佳组合。

通过对学习曲线的分析，证明模型没有出现过拟合现象。此外，还观察了被错误分类的样本被错误分类的原因，因为这些人的声音特征比较中性或者和异性的声音特征比较相似，所以被错误地分类。

从整体来看，准确率98.35%已经可以和人类的识别能力相媲美，而且项目已经完成了既定目标。

5.2 后续改进

项目使用的数据集包括大约3000个样本，从数量方面来说，3000个样本难以代表整个人类。所以，本项目的一个改进方向是扩大项目的数据集，采集更多的样本用于训练。在这种情况下，训练出来的更具有代表性。

从应用角度来讲，可以尝试开发一个网页，允许用户上传一段声音，网页可以根据这段声音判断说话人的性别。或者开发一个app，用户对着手机说话，app识别说话人的性别。

附录

[1]https://www.kaggle.com/primaryobjects/voicegender

[2]邓英、欧文贵《基于 HMM 的性别识别》.全国现代语音学学术会议, 2003:74-75. 链接

[3]周志华《机器学习》.清华大学出版社 .2016

[4]http://scikit-learn.org/