Universe Scientific Publishing
Computer System Networking and Telecommunications
计算机系统网络和电信，2019,1(2)
ISSN:2661-3719 (Print); 2661-3727 (Online)

Machine Learning Algorithm-based Dating Predictions

Yuxuan WANG Ke WANG Chenxi LI
School of Computer and Software, Jincheng College, Sichuan University, Chengdu, 611731

Abstract
In this paper, we studied the relationship between frequent-flier miles obtained every year, the percentage of time spent playing video games, the number of ice cream liters consumed per week and the final friendship situation in a group of dating data, learned KNN and decision tree algorithm, and established a two-person model, and finally analyzed and validated the results. Experimental results show that KNN algorithm and decision tree algorithm are both simple and convenient classification algorithms.

Key Words
Artificial Intelligence, K-nearest Neighbor Algorithm, Decision Tree Algorithm, Machine Learning
DOI:10.18686/jsjxt.v1i2.693

基于机器学习算法的交友人选预测

王宇轩 王 科 李晨曦
四川大学锦城学院计算机与软件学院，四川成都，611731

摘 要
本文通过研究一组交友数据中关于每年获得的飞行常客里程数，玩视频游戏所耗时间百分比，每周消费的冰淇淋公升数与最终交友情况之间的关系，学习 KNN 和决策树算法，并建立了 2 者的模型，最终分析验证了结果。实验结果表明 KNN 算法与决策树算法都是行之有效的简单方便的分类算法。

关键词
人工智能；k 近邻算法；决策树算法；机器学习

1.引言

随着科学技术的蓬勃发展，人工智能也不断进行了技术革新，并不断涌现许多优秀的算法，K-近邻算法和决策树算法在机器学习领域中都早已被提出，虽然相比之后的不少算法并不算优秀，甚至拥有许多缺点，然而因为这两者本身简单有效的特性，使我们能够较为简单掌握其思路的同时，理解算法的思想和内涵，从而方便我们进一步深入探究，进而推开人工智能领域研究的大门。

本文采用的试验平台为 anaconda，anaconda 是一个开源的 python 发行版本，兼容了 pandas，numpy，sklearn 等实验中需要使用的 python 库，方便进行代码的编写。

2.KNN 算法简介

KNN，也称为 K 近邻算法或邻近算法，是数据挖掘的基础算法之一。KNN 算法自身简略有效，只存在向前传达过程，不存在学习过程，是机器学习中一种 lazy-learning 的分类算法。

KNN 算法的核心思想是在一个样本空间内，对于希望预测的点 n，选择距离 n 最近的 k 个值[1]，这 k 个值所在类别的出现概率的最大概率所代表的类别，即为未知点 n 的类别，并具有这个类别的所有特性。 KNN 算法属于分类与逻辑回归算法，将候选人的特点信息加入分类集合中，就可以进行分类判断。

KNN 算法的几个主要优点为：1)KNN 由于提出时间较早，经过了多年的发展，理论非常纯熟，还有着简单有效的算法思路，在回归问题和分类问题中都可以使用。2)可以用于非线性分类。3)训练时间复杂度较低。4)对数据没有假设，导致 KNN 算法相比其它机器学习算法准确度更高，对异常点相当不敏感。5)KNN 算法

自身只涉及计算测试样本到训练样本的距离，计算本身较为简单。

KNN 算法作为早期的人工智能领域算法，有一些缺点，例如：1)计算时间也就是时间复杂度受样本量大小的影响，在特征数十分多的时候，计算量会很大。2)样本不平衡的时候，对稀有类别的预测准确率很低。3)KNN 是一种 lazy-learning 算法，基本上不学习，导致预测速度相对其它算法如逻辑学习等方法慢。4)比较决策树模型，KNN 模型的可解释性不强。

在实现 KNN 算法的过程中，需要注意几个核心要点，比如 k 的取值，如果 k 值较小，就相当于只计算少数一个近邻的样本类型，设想一下，在最极端的情况下，k=1 时，最终结果将只受最近的一个样本类型影响，这会产生相当大的误差。这说明了假如 k 值取值过低，会导致过拟合的问题。相反，如果 k 值较大，例如极端情况下 k=n(n 为样本数量)，那么结果将是 n 中最多的样本类型，即产生了欠拟合的问题。为了避免产生过拟合或欠拟合，k 一般取值为 3~10。一种常见的做法是设 k 为训练集中案例数量的平方根[2]。

3.基于 KNN 模型预测交友人选

3.1 数据的处理

本文使用的是一组交友数据，数据样本共有 1000 行，样本量较大，方便进行训练和测试，候选人主要包含以下三种特征：每年获得的飞行常客里程数，玩视频游戏所耗时间百分比，每周消费的冰淇淋公升数[3]。

实验中使用了 pandas 库函数进行数据读取，Pandas 是基于 Numpy 的一种工具，该工具提供了大批快速处理数据的函数和方法，使我们可以更好地处理数据分析工作。该实验中我们可以采用 pandas.read_csv()函数读取后缀为 csv 的样本数据。

因为本文中数据的几个特征的取值范围差异很大，而 KNN 算法以距离为度量，这会导致更新的结果也产生较大的差异，这是我们所不希望的。我们会希望认为特征的重要水平是一样的，并不想偏袒某个特征，所以在使用这些特征之前，对数据进行预处理是必要的，预处理数据的 2 种方式分别为标准化和归一化。

标准化要求均值为 0，标准差为 1，转换公式如下：

$$Z=\frac{\mathrm{x}-\mu}{\sigma} \quad (1)$$

其中，

$$\mu=\frac{\left(\mathrm{x}_1+x_2+\cdots+x_n\right)}{n} \quad (2)$$

$$\sigma=\sqrt{\frac{\left(\mathrm{x}_1-\mu\right)^2+\left(\mathrm{x}_2-\mu\right)^2+\cdots+\left(\mathrm{x}_i-\mu\right)^2}{n}} \quad (3)$$

归一化能将所有特征的值在处理后压缩到 0 到 1 的区间上，这样做可以抑制离群值对结果的影响，归一化公式如下：

$$X_{norm}=\frac{X-X_{\min}}{X_{\max}-X_{\min}} \quad (4)$$

3.2 使用 KNN 算法建立模型并评估

在建立模型之前，需要选择向量距离度量规则，在上述数据处理过程中，数据已经标准化/归一化，特征重要程度相同，因而应用欧式距离作为距离度量规则，欧式距离越小，说明相似度越大。

在二维空间中，设 A 点的坐标为(q1,p1)，B 点的坐标为(q2,p2)，以此类推，第 N 个点的坐标为(qn,pn)，那么这一组向量的欧式距离 d 的公式为：

$$\mathrm{d}=\sqrt{(q_1-p_1)^2+(q_2-p_2)^2+\cdots+(q_n-p_n)^2} \quad (5)$$

在该实验中，由于只有一个样本数据集，可以将这个样本数据集依据一定条件划分为训练集和测试集，其中训练集是用来训练模型的数据集，测试集的作用是测试训练集训练的模型的效果，并且测试集的特征和训练集的特征不能不同，样本规模要足够，从而保证可产生具备统计意义的结果，因此在实验的具体实施中，将 1000 条数据以 2:8 比例分开，前一部分数据作为测试集，其它数据作为训练集进行计算，从而得到实验结果。

实验通常使用 RMSE（均方根误差）评估模型，RMSE 的公式为：

$$RMSE=\sqrt{\frac{(actual_1-predicted_1)^2+(actual_2-predicted_2)^2+\cdots+(actual_n-predicted_n)^2}{n}} \quad (6)$$

均方根误差又称标准误差，反映了一组数据集的离散程度，标准差越小，说明数据更加精确，模型效果越好。

3.3 实验结果分析

运行结束后加权平均值 RMSE 为 0.29232644，由此可得，该模型表现效果良好。由于样本数越大，标准差越靠近总体标准差，标准误差也会随测量次数的增加而减小，继续提高样本数将有利于提高识别的准确度，也就是说，如果能够进一步提升样本数量，模型将有进一步找到合适拟合的可能。

本次测试只涉及到了简单的分类应用，KNN 算法在分类与回归中都有各自的应用空间，若未来需要对更加复杂的问题进行分析，还需进一步在分类效率和分类效果两方面优化算法。

4.决策树算法简介

决策树算法是一种基本的分类与回归算法，是能够将“不确定转化为确定”的分析措施，是直观利用概率分析的一种图解法，由于决策分支画出的图形很像一棵树，故称为决策树。决策树算法从根节点一步步走到叶子节点（决策），并且所有的数据都能落到叶子节点，既可以做分类也可以做回归[4]。

决策树有以下几个优点：1)易于理解和实现，使人们在学习过程中不需要使用者了解很多的背景知识，这同时是它的能够直接体现数据的特点，只要通过解释后都有能力去理解决策树所表达的意义[5]。2)几乎不需要数据预处理，相比前文的 KNN 算法，决策树的数据并不需要进行标准化/归一化或是创建虚拟变量等预处理。3)既可以处理分类问题，对回归问题的效果也十分良好，如 ID3 和 C4.5 是分类算法，CART 是回归算法。4)决策树使用白盒模型，相比黑盒模型算法（例如人工神经网络），更容易运用逻辑判别体现这种规则。

5.基于决策树模型预测交友人选

5.1 构建决策树模型

为了从根节点开始选择特征，构造出一颗决策树，需要一种衡量规范，从而计算通过不同特征进行分支抉择后的分类状况，决策树通过贪婪思想进行分裂，通常选择最优状况作为根节点，其余节点的选择也以此类推。

本文使用决策树的 ID3 算法，该算法以信息论为基础，以熵值的信息增益作为判断标准，熵是表现随机变量不确定性的度量，不确定性越大，得到的熵值也越大，公式为：

$$H(X)=-\sum p_i * \log(p_i), i=1,2,\cdots,n \qquad (7)$$

信息增益是分裂前后的根的数据复杂度和分裂节点数据复杂度的变化值.

决策树不可能无限制的生长，总有停止分裂的时候，设想在极端条件下，决策树只剩一个数据点的时候，分裂才最终停止，这种情况下无疑会导致过拟合。所以决策树需要剪枝，一般而言有 2 种剪枝方法：预剪枝和后剪枝

预剪枝可以限制深度，叶子节点个数，叶子节点样本数，信息增益量等，边建立决策树边进行剪枝的操作，后剪枝通过一定的衡量标准进行剪枝,当决策树建立结束后，才会进行后剪枝操作。

本文使用使用 sklearn 库中 tree.DecisionTreeClassifier()函数建立决策树，采用限定深度的方法进行剪枝，限定决策树的深度为 4，代码为：

tr = tree.DecisionTreeClassifier(max_depth = 4,criterion = 'gini')

5.2 实验结果分析

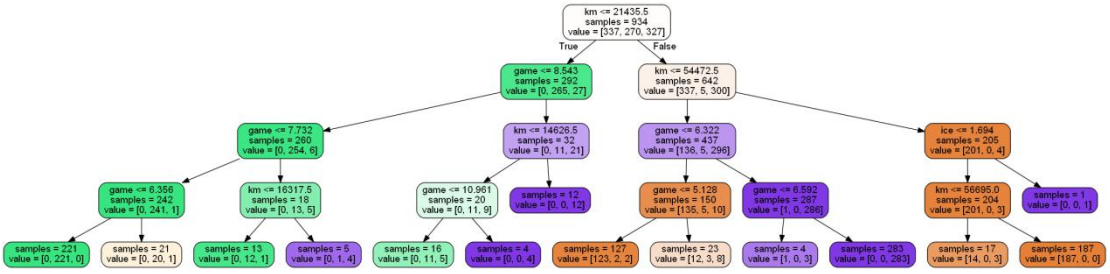

图 1

图 1 为最终建立的决策树。

为展示训练结果，将原数据再次使用 score 函数输

入，得到正确率为 0.93617021276595747，由此可得，该模型表现良好。由于该实验采用 ID3 决策树，存在信息增益偏向选择取值较多特征的问题[6]，如果采用 C4.5 决策树，使用信息增益率作为评判标准，模型正确率还有上涨的空间。

6.结论

综上所述，本文基于 anaconda 和 python 语言简要实现了 k-近邻算法和决策树算法的理论学习和实践应用。K-近邻算法和决策树算法作为人工智能领域的基础算法，是研究学习更高深算法的起点。然而 KNN 算法和决策树算法毕竟是机器学习的早期算法，随着其它算法的提出，本身具有的许多缺点也更为明显。本文虽然只通过简单的例子粗略的使用了 2 种算法，但在例子中算法依然有改进的空间，如果能够不断改进和完善 2 种算法，就能够更深入地理解机器学习的思想，也有利于学习其它算法思路，从而在人工智能领域的研究中更进一步。

参考文献

[1] 赵凯迪 基于 SVM 的局部加权 KNN 分类算法的研究.浙江工商大学硕士学位论文,2018.
[2] 王永波 基于机器学习的花卉分类算法研究.现代计算机：上下旬,2015.
[3] Peter Harrington 机器学习实战.北京：人民邮电出版社,2013.
[4] 王秀岩 决策树算法及其应用.电子技术与软件工程,2014.
[5] 毛聪莉 基于粗糙集的决策树学习算法研究.湖南大学硕士论文,2008.
[6] 叶萌 决策树学习研究综述.黑龙江科技信息,2011.

作者简介

第一作者：王宇轩（1998-），男，汉，四川省成都市，本科，四川大学锦城学院，研究方向：人工智能。
第二作者（通讯作者）：王科（1985—），男，讲师，硕士研究生，四川大学锦城学院，研究方向 ：云计算
第三作者：李晨曦（1998-7），男，汉，贵州省贵阳市，本科，四川大学锦城学院，研究方向：大数据技术开发