Hans Journal of Data Mining 数据挖掘, 2019, 9(3), 96-103
Published Online July 2019 in Hans. http://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2019.93012

Hans汉斯

Analysis of Alcohol Preference Characteristics and Prediction of Alcohol Abuse Tendency Based on Machine Learning

Xiaona Zhao

University of International Business and Economics, Beijing
Email: 2264018660@qq.com

Received: Jun. 26th, 2019; accepted: Jul. 9th, 2019; published: Jul. 16th, 2019

Abstract

Heavy drinking refers to drinking exceeds the standard of moderate drinking or general social drinking. Heavy drinking has serious negative effects on personal development and family, so the analysis of the causes of alcohol abuse and the behavior prediction of heavy drinking are important. Features include basic information (educational level, age, gender, country of residence, ethnicity) and five-factor personality measures (Neuroticism, Extraversion, Openness to experience, Agreeableness, Conscientiousness) as well as Barrett impulsiveness and impulsive sensations seeking. We used machine learning-based decision trees, Naive Bayes, K-nearest neighbors, support vector machines, logistic regression and other classification methods to predict and analyze. Then we can predict whether there is a tendency to alcohol abuse according to a person's basic information and personality characteristics. And the characteristics of alcoholics were analyzed: extroversion had a great influence on drinking behavior; people with high openness were less inclined to drinking alcohol, as for other personality traits, the higher Neuroticism, Agreeableness, Conscientiousness, Barrett impulsiveness and feeling seeking, the greater the possibility of drinking.

Keywords

Machine Learning, Heavy Drinking, Prediction, Feature Analysis

基于机器学习的酒精偏好者特征分析及酗酒倾向预测

赵萧娜

文章引用: 赵萧娜. 基于机器学习的酒精偏好者特征分析及酗酒倾向预测[J]. 数据挖掘, 2019, 9(3): 96-103.
DOI: 10.12677/hjdm.2019.93012

对外经济贸易大学，北京
Email: 2264018660@qq.com

收稿日期：2019年6月26日；录用日期：2019年7月9日；发布日期：2019年7月16日

摘　要

饮酒超出适量饮酒或一般社交性饮酒的标准为重度饮酒，重度饮酒无论对个人发展还是对家庭、社会都会产生严重的负面影响，因此，通过特征进行酗酒原因分析和行为预测具有重要意义。特征包括基本信息(教育水平、年龄、性别、居住国、种族)以及五因素人格测量(神经质、外向性、开放性、友善性、严谨性)，还有巴瑞特冲动性和冲动感觉寻求，基于机器学习的决策树、朴素贝叶斯、K近邻、支持向量机、逻辑回归多种分类方法进行预测，通过对一个人的基本信息和性格特征的分析来预测是否具有酗酒倾向。对酗酒者的特征进行了分析，外向对饮酒行为有较大影响，开放性高的人更不倾向于饮酒，在其他性格特征中，神经质，友善性，严谨性，巴瑞特冲动性和感觉寻求越高，饮酒可能性越大。

关键词

机器学习，重度饮酒，预测，特征分析

1. 引言

1.1. 重要性

当今社会，人们面临学习、就业、经济、交际、恋爱等某方面或者多方面压力，饮酒作为一种暂时性的消除烦恼的方式很容易被人们采用。然而，如若不能合理控制自己饮酒消遣的频率，则会对酒精产生依赖，对自己的身体和心灵以及他人均产生不好的影响。一个人是否更容易借助饮酒逃避和屈服于遇到的挫折与个人的成长经历和性格特征相关。通过机器学习的多种分类和预测方法(决策树，朴素贝叶斯，K近邻和支持向量机)对这些特征进行分析，找出影响较大的因素，可以对重度饮酒者进行针对性的帮助，也可以对一个人是否具有饮酒倾向做出预测，及时采取措施，防患于未然。

1.2. 相关研究

导致药物滥用的因素包扩心理、社会、个人、环境以及经济方面[1] [2] [3]，这些方面也与很多性格特质相关[4] [5]。而糖、酒精与烟草这些合法药物比非法药物更可能导致早逝[6]。使用消遣药物对社会和个人的影响应该加以重视[7]。心理学家普遍认同五因素人格测量模型可以很好地反应个人差异[8]，五因素人格测量模型包括神经质(N)，外向性(E)，开放性(O)，友善性(A)和严谨性(C)。很多研究表明人格特质与药物的滥用相关，Flory 等[9]发现滥用酒精者的 A 和 C 较低，E 较高，他们也发现较低的 A 和 C，较高的 O 与大麻使用相关。

2. 研究方法

2.1. 分类方法介绍

机器学习属于人工智能研究较早的分支，分为有监督学习，无监督学习和半监督学习。其中监督学习是通过已知的数据训练出一个模型，从而对新的数据通过此模型可以得出其分类。本文用到的的监督学习分类器有：

1) 朴素贝叶斯

基于贝叶斯定理与特征条件独立假设的分类方法。采用计算每一个样本属于每一类的概率，然后将样本划分为具有最大概率的那一类中。

朴素贝叶斯能处理多分类任务，尤其对小规模的数据表现很好，因此可以在训练时把数据集分成批次，一批一批的增量式训练。

2) CART 决策树

对所有属性分别计算 Gini 系数增益，取 Gini 系数增益值最大的属性作为决策树的根节点属性，对划分后的子数据集采用同样的方法决定子树的根节点，递归进行。决策树有非常良好的优点：决策树的构造不需要任何领域知识，就是简单的 IF...THEN...思想；决策树能够很好的处理高维数据，并且能够筛选出重要的变量；由决策树产生的结果是易于理解和掌握的；决策树在运算过程中也是非常迅速的；一般而言，决策树还具有比较理想的预测准确率。

3) 支持向量机

支持向量机是解决分类问题的一种算法，SVM 就是无数条可以分类的直线中最完美的一条，其恰好在两个类的中间，距离两个类的点一样远。

4) K 近邻

给定一个训练集，对新的样本，在训练集中找到与该样本最邻近的 K 样本，这 K 个样本多数属于某个类，则新样本属于该类。

由于 KNN 方法主要靠周围有限的邻近的样本，而不是靠判别类域的方法来确定所属类别的，因此对于类域的交叉或重叠较多的待分样本集来说，KNN 方法较其他方法更为适合。该算法在分类时有个主要的不足是，当样本不平衡时，如一个类的样本容量很大，而其他类样本容量很小时，有可能导致当输入一个新样本时，该样本的 K 个邻居中大容量类的样本占多数。因此可以采用权值的方法(和该样本距离小的邻居权值大)来改进。该方法的另一个不足之处是计算量较大，因为对每一个待分类的文本都要计算它到全体已知样本的距离，才能求得它的 K 个最近邻点。目前常用的解决方法是事先对已知样本点进行剪辑，事先去除对分类作用不大的样本。

5) 逻辑回归

逻辑回归思想基于线性回归，线性回归的主要思想是通过历史数据拟合一条直线，用这条直线对新的数据进行预测。而逻辑回归是寻找合适的分类函数，对该函数的数据预测并得到最终的分类结果，然后构造代价函数即损失函数，用以表示预测的输出结果与训练数据的实际类别之间的偏差，最小化代价函数从而获得最优的模型参数。

2.2. 评测标准

本研究采用准确率、召回率、F 值(综合衡量指标)作为评测标准，准确率 = 被正确分类的样本数/所有被分为该类别的样本总数，召回率 = 实际被分类为该类别的样本数/应该被分类为该类别的样本数，准确率与召回率越高代表分类效果越好，但两者存在矛盾，F 值综合考虑了准确率和召回率，寻找一个

准确率和召回率都比较高的平衡点，measure 通过调节 γ 来调节准确率和召回率两者之间的具体联系，当 γ 大于 1 时，F 值受到准确率的影响较大，当 γ 小于 1 时，受到召回率影响更大。当 γ 为 1 时，为 F1 值，召回率与准确率对该分类系统的影响同样重要。

$$\text{F-measure} = \frac{\left(\gamma^2+1\right)*\text{precision}*\text{recall}}{\gamma^2\left(\text{precision}+\text{recall}\right)}$$

$$\text{F1} = \frac{2*\text{precision}*\text{recall}}{\text{precision}+\text{recall}}$$

下面通过一个例子展示各个指标的详细计算方法，A1 为被正确判别为 A 类的样本数，B1 为被正确判别为类别 B 的样本数；A2 为所有被判别为类别 A 的样本数，B2 为所有被判别为类别 B 的样本数；A3 为类别 A 实际的样本数，B3 为类别 B 实际的样本数，具体公式如下表 1。

Table 1. Description of evaluation indicators
表 1. 评价指标说明

评价指标	类别 A	类别 B
准确率	$\frac{A_1}{A_2}*100\%$	$\frac{B_1}{B_2}*100\%$
召回率	$\frac{A_1}{A_3}*100\%$	$\frac{B_1}{B_3}*100\%$
F1	$2*\frac{\frac{A_1}{A_2}*\frac{A_1}{A_3}}{\frac{A_1}{A_2}+\frac{A_1}{A_3}}*100\%$	$2*\frac{\frac{B_1}{B_2}*\frac{B_1}{B_3}}{\frac{B_1}{B_2}+\frac{B_1}{B_3}}*100\%$

3. 数据来源与预处理

3.1. 数据来源

本研究数据来源与 UCI 网站(http://archive.ics.uci.edu/ml/index.php)的 drug consumption 数据集，数据集包含 1885 名受访者记录。每个受访者具有 12 个属性，包含：

1) 基本信息：教育水平、年龄、性别、居住国、种族，属性类型及取值见表 2。

2) 性格因素：五因素人格测量(神经质、外向性、开放性、友善性、严谨性)、巴瑞特冲动性、冲动感觉寻求，性格因素变量解释：

神经质(Neuroticism)：指紧张、焦虑、抑郁等负面情绪的长期倾向；

外向性(Extraversion)：指外向开朗，健谈，积极，开心的程度；

开放性(Openness to experience)：想象力，创造力，标新立异的想法，广泛的兴趣爱好；

友好性(Agreeableness)：人际关系方面的利他性，信任，谦虚善良富有同情心，协同性；

严谨性(Conscientiousness)：可靠性，意志坚定性以及快速高效井井有条。

五因素人格测量的结果是通过受访者回答与这些特质相关的 60 道问题得出对于其每个特质的结果。

巴瑞特冲动性(impulsiveness)由三方面组成：第一个现代是不经思考就行动，第二是注意力不集中思想容易被打乱，第三方面是做事情不考虑后果，综合此三方面的程度来度量一个人的冲动性。

冲动感觉寻求(sensation seeking)由冲动性和感情诉求组成，此因素与高风险行为息息相关。

采访结果经过量化处理，使得各个性格特质的取值具有相差不大的尺度，表 3 为一些样本取值。

Table 2. Basic information values
表 2. 基本信息取值

特征类型	特征划分
Age	1~24 25~34 35~44 45~54 55~64 65+
Gender	Female male
Education	Left school before 16 years Left school at 16 years Left school at 17 years Left school at 18 years Some college or university, no certificate or degree Professional certificate/ diploma University degree Masters degree Doctorate degree
Country	Australia Canada New Zealand Other Republic of Ireland UK USA
Ethnicity	Asian Black Mixed-Black/Asian Mixed-White/Asian Mixed-White/Black Other White

Table 3. Examples of values of personality factor characteristics
表 3. 性格因素特征的取值示例

Neuroticism	Extraversion	Openness to experience	Agreeableness	Conscientiousness	Impulsiveness	Sensation seeking
0.31287	−0.57545	−0.58331	−0.91699	−0.00665	−0.21712	−1.18084
1.23461	1.93886	1.65653	−0.15487	−0.52745	1.29221	1.2247
−0.67825	1.93886	1.43533	0.76096	−0.14277	−0.71126	−0.21575
−0.46725	0.80523	−0.84732	−1.6209	−1.0145	−1.37983	0.40148
−0.14882	−0.80615	−0.01928	0.59042	0.58489	−1.37983	−1.18084
0.73545	−1.6334	−0.45174	−0.30172	1.30612	−0.21712	−0.21575
−0.67825	−0.30033	−1.55521	2.03972	1.63088	−1.37983	−1.54858
−0.46725	−1.09207	−0.45174	−0.30172	0.93949	−0.21712	0.07987
−1.32828	1.93886	−0.84732	−0.30172	1.63088	0.19268	−0.52593
0.62967	2.57309	−0.97631	0.76096	1.13407	−1.37983	−1.54858
−0.24649	0.00332	−1.42424	0.59042	0.12331	−1.37983	−0.84637
−1.05308	0.80523	−1.11902	−0.76096	1.81175	0.19268	0.07987
−1.32828	0.00332	0.14143	−1.92595	−0.52745	0.52975	1.2247
2.28554	0.16767	0.44585	−1.6209	−0.78155	1.29221	0.07987

3.2. 数剧处理

对于每个受访的酒精使用情况，分为：“从未使用”，“十年前使用”，“过去十年使用”，“去年使用”，“上个月使用”，“上周使用”，和“昨天使用”，本研究为了方便，将“从未使用”，“十年前使用”，“过去十年使用”，“去年使用”，“上个月使用”样本分类的取值归类为“month−”，频率为每月及更少，视为正常饮酒者；将“上周使用”，和“昨天使用”的样本归类为“week+”，频率为每周及以上，视为重度饮酒者。

4. 结果分析

本研究过程分为两部分，第一部分对于基本信息与是否重度饮酒之间的关系进行分析，第二部分对于性格因素对重度饮酒的影响进行分析。

4.1. 利用基本信息分类与预测

将 1885 条数据的 2/3 作为训练集，1/3 作为预测集，进行以下分析：

1) 运用 CART 决策树通过基本信息进行分类。决策树的优点在于算法不受数据缩放影响，特征不需要归一化，标准化等预处理，因此对于五类离散型变量的分析适合选取决策树模型。

决策树 graphviz 可视化图中的 samples 给出该节点中的样本数 values 给出每个类别的样本数。

调用模型时需选择参数：树的深度，若为 None 表示不限，直到所有的叶子节点中所有的样本都属于同一类别，或是每个叶子节点包含的样本数小于分类内部节点的最少样本数；分裂节点策略选择，最优或是随机切分两种选择；分类内部节点的最少样本数；每个叶子节点的最少样本数；叶子节点中样本的最小权重；寻找最佳划分时考虑的特征数量。

决策树结果如图 1 所示。

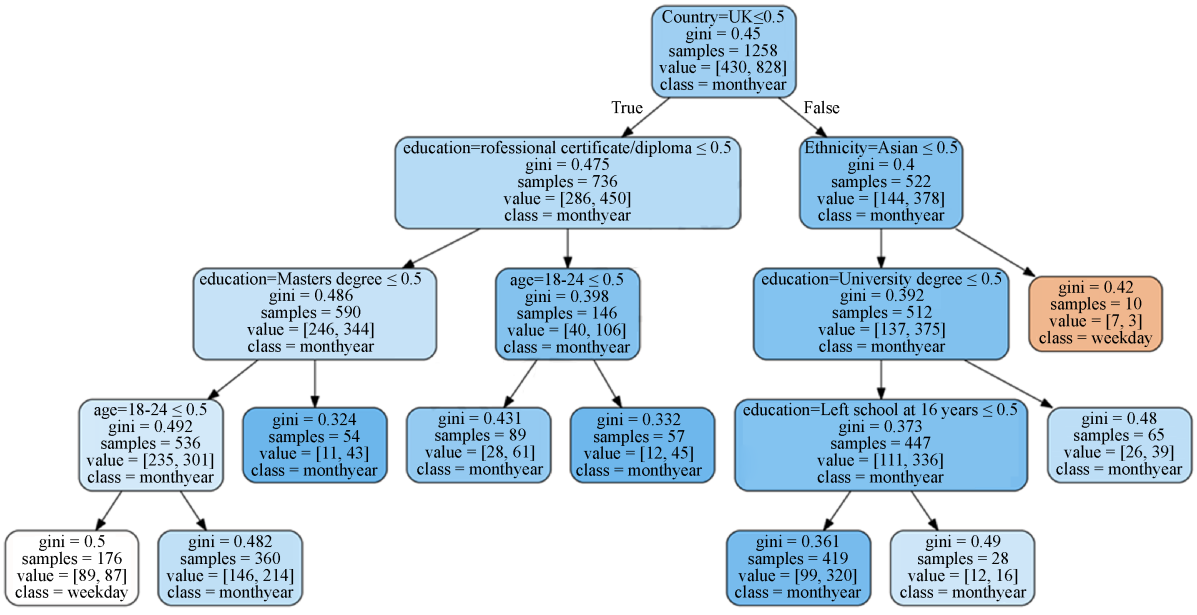

Figure 1. Decision tree model
图 1. 决策树模型

2) 多项式贝叶斯分类，以基本信息为特征的结果如下表 4。

Table 4. Polynomial Bayesian classification results
表 4. 多项式贝叶斯分类结果

	Precision	Recall	f1-score	Support
Weekday	0.64	0.11	0.19	191
Monthyear	0.71	0.97	0.82	437

3) 通过支持向量机法，以基本信息为特征的结果见表 5。

Table 5. Support vector machine classification results
表 5. 支持向量机分类结果

	Precision	Recall	f1-score	Support
Weekday	0.6	0.03	0.06	191
Monthyear	0.7	0.99	0.82	437

4) 通过 K 近邻法，以基本信息为特征的结果见表 6。

Table 6. K nearest neighbor classification results
表 6. K 近邻法分类结果

	Precision	Recall	f1-score	Support
Weekday	0.37	0.22	0.27	191
Monthyear	0.71	0.83	0.77	437

通过对比以上三种方法的分类结果发现，贝叶斯具有综合的最好的分类效果，三种方法对不酗酒的预测效果均较好。

4.2. 利用性格特征分类与预测

1) 高斯贝叶斯模型分类效果评价，对性格特征的分析，结果见表 7。

Table 7. Gauss Bayes classification results
表 7. 高斯贝叶斯分类结果

	Precision	Recall	f1-score	Support
Weekday	0.37	0.15	0.21	191
Monthyear	0.70	0.89	0.78	437

2) 通过逻辑回归方法对性格特征的分析

逻辑回归参数如下：0.05528946、0.18008346、−0.06434509、0.03357274、0.01287108、0.01264426、0.08029382 分别对应神经质、外向性、开放性、友善性、严谨性、巴瑞特冲动性、冲动感觉寻求的系数，从中可以看出外向性对饮酒产生较大的影响，开放性，即想象力，创造力，标新立异的想法高的人更不倾向与饮酒，其他性格特征，神经质，友善性，严谨性，巴瑞特冲动性和感觉寻求越高，饮酒可能性越大。逻辑回归模型结果见表 8。

以上两种方法的分类效果相比，高斯贝叶斯具有更好的综合效果。

Table 8. Logistic regression classification results
表 8. 逻辑回归分类结果

	Precision	Recall	f1-score	Support
Weekday	0.42	0.04	0.08	191
Monthyear	0.70	0.97	0.81	437

5. 总结

随着社会越来越开放和包容，饮酒行为十分司空见惯。大家开心了去喝酒庆祝，难过了去喝酒消愁，然而，如果不进行节制，很容易染上酒瘾，如此循环往复，对身体和心理造成巨大的伤害，饮酒因素包含基本特征和性格特征，通过了解某人的这些特征来分析其是否具有酗酒行为具有重要意义，这可以使相关机构着重进行干预，减少因酗酒造成的事故。也可以从相关的分析中看出，哪些因素与酗酒行为密切相关。本文运用了机器学习的 5 种分类方法，分别将基本信息和性格特征分析代入训练模型，通过测试集对模型的效果进行评价。最后几个模型均得出了不同的效果，在具体情况下，可根据需求调用适合的模型。在 5 种分类方法中，逻辑回归的方法可以进行特征的影响大小分析，表明外向性对饮酒行为有较大影响，开放性高的人更不倾向于饮酒，在其他性格特征中，神经质，友善性，严谨性，巴瑞特冲动性和感觉寻求越高，饮酒可能性越大。

参考文献

[1] Cleveland, M.J., Feinberg, M.E., Bontempo, D.E. and Greenberg, M.T. (2008) The Role of Risk and Protective Factors in Substance Use across Adolescence. *Journal of Adolescent Health*, **43**, 157-164. https://doi.org/10.1016/j.jadohealth.2008.01.015

[2] Ventura, C.A., de Souza, J., Hayashida, M. and Ferreira, P.S. (2014) Risk Factors for Involvement with Illegal Drugs: Opinion of Family Members or Significant Others. *Journal of Substance Use*, **20**, 136-142. https://doi.org/10.3109/14659891.2013.875077

[3] WHO (2004) Prevention of Mental Disorders: Effective Interventions and Policy Options: Summary Report. World Health Organization, Geneva.

[4] Dubey, C., Arora, M., Gupta, S. and Kumar, B. (2010) Five Factor Correlates: A Comparison of Substance Abusers and Non-Substance Abusers. *Journal of the Indian Academy of Applied Psychology*, **36**, 107-114.

[5] Bogg, T. and Roberts, B.W. (2004) Conscientiousness and Health-Related Behaviors: A Meta-Analysis of the Leading Behavioral Contributors to Mortality. *Psychological Bulletin*, **130**, 887. https://doi.org/10.1037/0033-2909.130.6.887

[6] Beaglehole, R., Bonita, R., Horton, R., Adams, C., Alleyne, G., Asaria, P., Baugh, V., Bekedam, H., Billo, N., Casswell, S., *et al*. (2011) Priority Actions for the Non-Communicable Disease Crisis. *The Lancet*, **377**, 1438-1447. https://doi.org/10.1016/S0140-6736(11)60393-0

[7] Bickel, W.K., Johnson, M.W., Koffarnus, M.N., MacKillop, J. and Murphy, G.J. (2014) The Behavioral Economics of Substance Use Disorders: Reinforcement Pathologies and Their Repair. *Annual Review of Clinical Psychology*, **10**, 641-677. https://doi.org/10.1146/annurev-clinpsy-032813-153724

[8] Costa, P.T. and MacCrae, R.R. (1992) Revised NEO-Personality Inventory (NEO PI-R) and and NEP Five-Factor Inventory (NEO-FFI): Professional Manual. Psychological Assessment Center, Odessa.

[9] Flory, K., Lynam, D., Milich, R., Leukefeld, C. and Clayton, R. (2002) The Relations among Personality, Symptoms of Alcohol and Marijuana Abuse, and Symptoms of Comorbid Psychopathology: Results from a Community Sample. *Experimental and Clinical Psychopharmacology*, **10**, 425-434. https://doi.org/10.1037//1064-1297.10.4.425

知网检索的两种方式：

1. 打开知网首页：http://cnki.net/，点击页面中“外文资源总库 CNKI SCHOLAR”，跳转至：http://scholar.cnki.net/new，搜索框内直接输入文章标题，即可查询；
   或点击“高级检索”，下拉列表框选择：[ISSN]，输入期刊 ISSN：2163-145X，即可查询。
2. 通过知网首页 http://cnki.net/顶部“旧版入口”进入知网旧版：http://www.cnki.net/old/，左侧选择“国际文献总库”进入，搜索框直接输入文章标题，即可查询。

投稿请点击：http://www.hanspub.org/Submission.aspx
期刊邮箱：hjdm@hanspub.org