计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2020,29(4):248−253 [doi: 10.15888/j.cnki.csa.007336]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

课程推荐预测模型优化方案及数据离散化算法①

张　戈

(中国社会科学院大学 计算机教研部, 北京 102488)
通讯作者: 张　戈, E-mail: goaller@sina.com

摘　要: 本研究基于 k-NN 算法建立了课程推荐预测模型. 由于原始样本数据的局部不均衡和数据叠交性, 预测模型在不进行任何参数调整和数据优化的情况下, 模型预测评分并不理想. 针对上述问题, 本研究设计了一套预测模型参数优化方案和样本数据优化方案, 包括最优 k 值选择算法设计、距离公式优化、数据离散化算法设计. 本研究提出的“数据离散化算法”驱使 kd 树的分类空间排序按照我们期望的特征向量的权重排序, 该算法对提升模型预测评分起到了积极作用. 上述优化方案和算法设计使课程推荐预测模型的评分从 0.67 提升到 0.85, 预测结果的准确度提高了 27 个百分点, 学生对课程推荐的满意度得到显著提升.

关键词: k-NN 算法; 最优 k 值选择; 距离公式优化; 数据离散化算法; 预测模型评分

引用格式: 张戈.课程推荐预测模型优化方案及数据离散化算法.计算机系统应用,2020,29(4):248–253. http://www.c-s-a.org.cn/1003-3254/7336.html

Optimization Scheme of Course Recommendation Prediction Model and Data Discretization Algorithm

ZHANG Ge

(Department of Computer Teaching and Research, University of Chinese Academy of Social Sciences, Beijing 102488, China)

Abstract: In this study, the course recommendation prediction model based on k-NN algorithm has been built. Due to the original sample data of the local imbalance and data overlapped, the prediction score of the prediction model is not ideal without any parameter adjustment and data optimization. Aiming at the above problems, this study designed a set of parameter optimization scheme and sample data discretization algorithm of the prediction mode, including the best k value selection algorithm, distance formula optimization, and data discretization algorithm design. In the study, the design of the “data discretization algorithm” drives kd tree classification feature space order sorted by the weight of the characteristic vector that we expect, this algorithm plays a positive role in improving model prediction score. Therefore, all of that increases the grade of the model from 0.67 to 0.85, and the accuracy of prediction results is increased by 27 percentage points, and students' satisfaction with course recommendation is significantly improved.

Key words: k-NN algorithm; selection of optimal k value; distance formula optimization; data discretization algorithm; prediction model score

在大学选课系统的“课程推荐预测”模块中, 系统需要根据学生的选课要求预测出最适合他的课程, 并给出课程推荐建议. 本研究的目标是通过对预测模型的参数优化和算法改进, 尽可能地提高模型评分即预测准确率, 给出课程推荐的最优解.

机器学习的预测算法众多, 根据样本数据特征值

① 基金项目: 2020 年中国社会科学院大学校级科研项目
Foundation item: Year 2020, Scientific Research Program of University of Chinese Academy of Social Sciences
收稿时间: 2019-08-14; 修改时间: 2019-09-06; 采用时间: 2019-09-23; csa 在线出版时间: 2020-04-05

的特性, 本研究选择 k 近邻算法 (k-Nearest Neighbors algorithm, k-NN) 拟合模型进行预测. k-NN 算法主要靠周围有限的邻近样本, 而不是靠判别类域的方法来确定目标点的所属类别. 由于本研究中原始样本数据具有局部不均衡和数据叠交性, 因此对于这种类域的交叉或重叠较多的样本集来说, k-NN 算法较其他算法更为适合. 但是在不进行任何参数调整和算法改进的情况下, 推荐课程的预测结果不能够覆盖学生对所选课程的要求, 模型预测结果不够准确. 为了更好地解决这些问题, 获得推荐和预测结果的最优解, 本研究从 k-NN 算法类的选择入手, 逐步探讨参数的调整方案, 在分析了 kd 树搜索最近邻算法之后, 依据样本数据特点研究和设计了“数据离散化算法”[1–3].

1 k-NN 算法基本原理和研究采用的机器学习方式

在模式识别中, k-NN 算法是一种用于分类和回归的预测方法. 在这两种情况下, 输入由特征空间中 k 个最邻近的训练实例组成. 输出则取决于 k-NN 是用于分类还是回归: 在 k-NN 分类中, 搜索出和目标点最近邻的 k 个样本点, 按多数投票原则选出最多的分类作为目标点标签. 在 k-NN 做回归时, 一般是用最邻近的 k 个样本的分类标签的平均值作为预测结果.

本研究使用 Python 语言机器学习工具包 scikit-learn 中的 KNeighborsClassifier 类建立课程推荐预测模型, 我们将围绕预测模型参数优化和数据离散化展开研究工作.

2 预测模型优化方案和数据离散化算法设计

表 1 列出了 KNeighborsClassifier 类的主要参数. 其中 n_neighbors 是近邻值, 即 k 值, 默认是 5, 分类器会选取 5 个与新数据点最接近的样本. Weights 是分类器在进行预测时用来计算样本权重的函数. 如果该参数为“uniform”, 则表示每个邻域中的所有样本的权重都是相等的. 如果该参数为“distance”, 则样本的权重值与它距新数据点的距离成倒数关系. algorith 决定了 k-NN 最近邻的核心算法, 该参数可以是“auto”、“brute”、“ball_tree”和“kd_tree”, 分别代表自动选取算法、暴力搜索、球树算法和 kd 树算法. metric 参数表示距离度量公式, 可以是曼哈顿距离或欧氏距离[4–7].

表 1 KNeighborsClassifier 主要参数

参数名称	类型	含义
n_neighbors	int	optional (default = 5) Number of neighbors to use by default for kneighbors queries.
Weights	str or callable	optional (default = ‘uniform’) weight function used in prediction.
Algorithm	——	{‘auto’, ‘ball_tree’, ‘kd_tree’, ‘brute’}, optional.Algorithm used to compute the nearest neighbors.
leaf_size	int	optional (default = 30) Leaf size passed to BallTree or KDTree.
P	integer	optional (default = 2)
Metric	string or callable	default ‘minkowski’
metric_params	dict	optional (default = None). Additional keyword arguments for the metric function.
n_jobs	int or None	optional (default=None)

2.1 最优 k 值选择算法

本研究首先对 k-NN 的重要指标 k 值进行优化. 在优化之前, 我们采用交叉验证方法对拟合模型进行准确性评估. 图 1 是使用测试集对训练样本模型的评分结果, 可以看到, 在没有任何参数调整和算法设计的情况下, 拟合模型评分仅为 0.67, 即预测结果有 67% 的准确率, 可以说模型的测评结果非常不理想.

k 值 (即 n_neighbors) 的选择高度依赖于样本数据. 一般来说如果 k 值较大, 则可以达到抑制噪声的作用. 当然 k 值过大会使分类边界不那么明显, 模型过于简化, 预测标签会产生多个结果均可的情况. 比如我们设定 k 值为 30, 那么 KNeighborsClassifier 分类器会选取 30 个与新数据最接近的训练样本点, 并按照最多投票原则, 选取它们中的最多分类标签作为预测标签. 相反, 如果 k 值过小, 分类线会竭尽全力的包含到每一个该类的点, 即使是噪点, 也会被包含, 预测模型变得复杂, 容易产生过拟合现象. 当 k 为 1 时, 就只有一个最邻近的样本点被选中, 它的标签即为目标新数据的预测标签. 一旦该样本点是个噪点, 那么预测结果就是错误的, 预测模型失去意义.

KNeighborsClassifier
(algorithm=' auto', leaf_size=30, metric=' minkowski',
metric_params=None, n_jobs=None, n_neighbors=5, *p*=2,
weights=' uniform')
拟合模型评分: 0.671 717 171 717 1717
推荐课程编号为: [3]
推荐课程编号名称为: Python 程序设计

图 1 默认参数下的拟合模型测评结果

本研究对 k 值的选择不采取固定取值的方式, 而是通过一个自定义函数完成 k 值的自动选取, 该函数的功能是在 k 值的一定选取范围内对预测模型进行交叉验证, 根据测评结果选出模型评分最高的 k 值. 图 2 为选取最优 k 值的活动图. 算法首先给出一个 k 的取值范围, 根据原始数据量设置为 1 到 50. 使用交叉验证方法建立训练集和测试集, 依次使用每个 k 值建立拟合模型, 比较它们的模型评分, 将模型评分最高的 k 值记录下来, 用该 k 值拟合的模型获取分类标签结果. 实验数据可以证明 k 值的选择尤为重要.

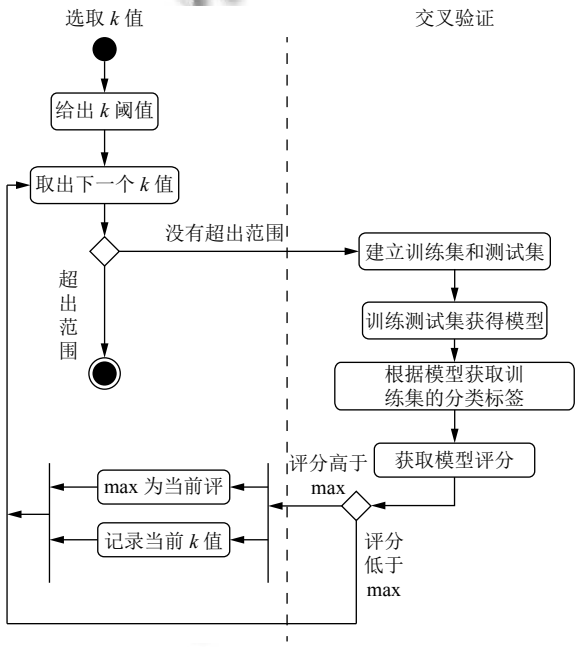

图 2 选取最优 k 值算法活动图

2.2 距离公式优化

预测模型的参数 Weights 的默认值是“uniform”, 表示邻域中的具有投票权利的各样本点的权重都是相等的. 这显然是不合理的. 目标新数据的标签应尽量依据距离它最近邻的样本点的标签给出, 并且这些投票样本点的标签对最终结果的贡献应该和它们与目标新数据的距离有关. 本研究将该参数调整为“distance”, 表示投票样本点的权重和其距新数据点的距离相关 (倒数关系), 即距离越近的投票样本点影响力越大.

本研究中 k-NN 最近邻使用的算法是 kd_tree, kd_tree 中距离公式的选择至关重要. 预测模型默认采用“闵可夫斯基”距离公式:

$$dist(A,B)=\left(\sum_{i=1}^{n}|a_i-b_i|^p\right)^{\frac{1}{p}} \tag{1}$$

当式 (1) 中的 p 为 2 时, 即为欧几里得距离, 当 p 为 1 时, 即为曼哈顿距离.

如果距离度量采用欧式距离 (euclidean), 分类器会计算样本点和新数据点之间的绝对距离. 本研究中样本数据每个特征取值范围较小, 均是 0 到 5 区域内的整数, 那么新数据点距离每个特征的欧式距离值就非常相近, 分类界线不明显, 欧式距离无法完成更好地分类作用, 分类边界模糊的情况仍旧没有得到改善. 式 (2) 是曼哈顿距离公式:

$$dist(A,B)=\sum_{i=1}^{n}|a_i-b_i| \tag{2}$$

曼哈顿距离计算的是目标数据点和各个对应特征之间距离的总和.

我们从原始数据中随机抽取 300 条数据对上述两种距离公式进行比较. 图 3(只显示部分样本数据) 中分别计算了目标数据点和每个样本数据的欧式距离和曼哈顿距离, 欧式距离的样本方差为 1.4, 曼哈顿距离的样本方差为 3.4. 可以看出使用欧式距离度量的训练集样本点分布较密集, 样本点之间的差距不大, 不利于分类. 曼哈顿距离会分散样本点分布, 分类时的界线识别会略好于欧式距离. 实验数据验证了采用曼哈顿距离的预测模型评分略高于采用欧式距离的评分[8–10].

2.3 数据离散化算法设计

k-NN 邻近算法的核心规则在 brute、kd_tree 和 ball_tree 三种算法中进行选择. 本研究中, 特征的维度不会超过 20 个, 因此我们采用更高效、速度更快的 kd_tree 搜索最邻近值.

图 3 样本点两种距离度量比较

kd_tree 搜索最邻近算法首先会找出方差最大的特征向量, 然后将其作为当前分割维度, 按中位数分割该维度空间, 在当前维度上小于中位数的数据集作为左子树的数据集, 大于等于中位数的数据集作为右子树的数据集, 依次重复递归直到建立一棵 kd 树, 从而可以搜索最近邻的点[11,12].

可以看出, 特征向量的权重依据它们各自的数据方差. 本研究中的 6 个特征向量取值范围均为 0~5. 从收集的样本数据来看, “对课程通过率重视程度”和“对课程趣味性的重视程度”两个特征相比其他特征向量, 其数据更集中在 3~5 之间, 数据更密集, 方差更小. 如果按照 kd 树算法的空间分割依据, 这两个特征向量会最后被分割, 也就是它们的权重排序是最后两位. 但实际上, 我们希望上述两个特征向量的权重排序分别为第 4 位和第 5 位. 因此我们设计了“数据离散化算法”, 以期达到“人为”修改 6 个特征向量方差的目标, 从而让模型按我们希望的特征权重排序进行分类. 如果采用传统的标准化数据的方法, 可以将 6 个特征向量数据统一映射到[0,1]区间上, 但是它不能改变特征向量的权重, 其特征向量方差的排序仍旧没有改变.

本研究建立了“数据离散化算法”的核心公式:

$$X^* = \sqrt{X\left(X_{\max} - \lambda\left(\sum_{i=1}^{n} X_i\right)/n\right)} \tag{3}$$

每个特征向量的所有样本数据均通过该公式进行预处理. 在式 (3) 中, X 是原始数据, X^*是离散化后的数据. 原始数据 X 乘以一个倍数后发生离散, 该倍数等于样本数据最大值减去样本数据均值乘以系数 λ. λ 系数为人工给出的期望权重值, 取值范围在[0,1]之间, 其作用是降低数据分布密集在[3,5]之间的两个特征的均值. 经过该离散化公式的处理, 原样本数据各个特征向量的方差排序变为了我们希望的排序顺序. 但是如果考虑到分布区间内数据点的纯度对改变特征权重排序的影响, 我们需要进一步引入信息熵并研究其对数据离散化的作用[13–15].

3 实验分析

本研究使用 Python 语言工具包 scikit-learn v0.21.3 对采集的 Excel 格式的近千条样本数据拟合预测模型, 按 3:1 的比例抽取训练集 train 和测试集 test 数据, 通过交叉验证方法评测预测模型, 给出模型准确率评分. 代码编辑和输出结果在 jupyter notebook 环境下完成.

3.1 最优 k 值选择实验

本研究首先设计了最优化 k 值的算法, 功能通过自定义函数 selectK() 实现. 图 4 是函数的代码及模型测评结果, 可以看到当 k=7 时, 模型评分最高, 未进行任何优化前的模型评分为 0.67, 现在为 0.76, 提高了约 13 个百分点.

图 4 选取最优 k 值实验及测评结果

3.2 距离公式优化实验

经过分析后, 距离度量公式采用曼哈顿距离, 即 p=1, weights 改为“distance”, 从图 5 可以看出调整距离公式参数后, 模型评分从 0.76 变为 0.79, 提高了约 4 个百分点. 使用曼哈顿距离度量提高了 kd 树搜索算法在建树过程中分割特征空间时的辨识度, 一定程度地提高了预测分类的准确度. 但是它对模型测评分提升效果并不明显. 从实验结果可以看出, 样本数据的特征维度在 20 以内、样本数据量较大时, 采用欧式距离还是曼哈顿距离对最终模型预测结果的影响力没有参数 k 值最优化的影响力大.

3.3 数据离散化算法实验

到目前为止模型的评分并没有超过 0.8, 在模型优化工作之后, 我们的重点转到对样本数据的离散化上. 表 2 是数据离散化前后的各个特征向量方差及其排序对比. 可以看到我们最为关心的两个分布较密集的特征, 其方差排序由之前的第 6 位和第 5 位, 变为现在的第 4 位和第 5 位, 特征向量的权重排序也因此而变为了我们希望的权重排序.

KNeighborsClassifier
(algorithm=’ auto’, leaf_size=30, metric=’ minkowski’,
metric_params=None, n_jobs=None, n_neighbors=3, *p*=1,
weights=’ distance’)
拟合模型评分为： 0.794 117 647 058 8235

图 5 采用曼哈顿距离公式实验测评结果

图 6 是数据离散化后的模型测评得分, 以及对新数据点[2,4,5,2,0,2]进行预测的结果. 可以看到此时的预测模型得分为 0.85, 即预测准确率为 85%. 而且模型给出了正确的课程预测结果, 推荐课程准确.

表 2 数据离散化实验前后方差和特征权重对比

	喜欢艺术类课程程度	喜欢编程类课程程度	具有的计算机基础程度	对课程通过率的重视程度	对作业量的重视程度	对课程趣味性重视程度
原始数据各特征向量方差	1.707 0272	1.715 4936	1.702 08	1.682 632	1.686 309	1.685 806 46
原始数据各特征向量权重排序	2	1	3	6	4	5
数据离散化后各特征向量方差	1.707 0272	1.715 4936	1.702 08	1.701 395	1.683 37	1.700 050 73
数据离散化后各特征向量权重排序	2	1	3	4	6	5

图 6 数据离散化实验结果

表 3 是本研究在模型未优化时和经过最优化 *k* 值、距离公式优化、数据离散化几个过程的预测模型测评分对比. 可以看到 *k* 值的自动选取函数对预测准确率的提升贡献最大. 其次是数据的离散化处理使模型测评准确率提高了 7.4%. 那么可以看到距离公式的优化对提高模型准确率只贡献了不到 4 个百分点, 这个和样本数据的特征维度个数有关. 在经过了预测模型参数优化和数据离散化过程后, 模型预测准确率由 0.67 提高到了最后的 0.85, 效果非常显著.

表 3 各个优化过程模型测评分对比

	模型未优化	最优化 *k* 值	距离公式优化	数据离散化
模型评分	0.6717	0.7647	0.7941	0.8529
提高百分比 (%)	——	13.8	3.8	7.4

4 结论与展望

为了在学生选课时给他们推荐更适合他们的课程, 本研究建立了课程推荐预测模型. 在对样本数据的特点进行详细分析之后, 本研究设计了一套预测模型优化方案和数据离散化算法, 使预测模型的准确率评分提高了约 26.8%.

本研究在进行过程中发现了一些问题和需要进一步探讨的内容. 首先, 距离公式参数调整对提高模型准确率效果不显著, 随着数据量和特征的增加, 距离公式的影响权重需做进一步的研究. 第二, 实验证明了数据离散化对模型优化的显著效果, 但是还有一些问题需要做进一步的思考. 例如两个分布较密集的特征向量其区间明显被分割为[0,2]和[3,5]两个分布. 相对于区间[3,5]的数据来说, 区间[0,2]的数据点是否可以看做异常点被“抛弃”? 区间数据的纯度对数据有什么影响? 如何改进“数据离散化”公式, 用一个新的算法自动给出合理的 λ 值, 而不是人工给出 λ 值. 第三, 样本数据本身是否合理是本研究进行过程中最困扰研究者的一个问题. 从实验结果能够看到, 在做出了预测模型优化和离散化数据处理之后, 模型的评分仍没有达到 0.9 以上, 更不要说接近 1 的高模型评分. 究其原因, 原始样本数据在特征向量设计上存在优先级不明确、课程的特征属性相互叠交的情况, 一条含糊不清的特征数据, 可能对应 1 到 2 个标签结果, 并且这两个结果均合理. 如何让样本数据更可用是下一步要进行的研究.

参考文献

1 严晓明. 基于类别平均距离的加权 KNN 分类算法. 计算机系统应用, 2014, 23(2): 128–132. [doi: 10.3969/j.issn.1003-3254.2014.02.022]

2 应毅, 任凯, 刘亚军. 基于 GIS 技术和加权 kNN 算法的实时揽件调度方法. 计算机工程与应用: 1–6. http://kns.cnki.net/KCMS/detail/11.2127.tp.20190911.1128.004.html. (2019-09-28)[2019-12-18].

3 Shi KS, Li LM, Liu HT, *et al*. An improved KNN text

classification algorithm based on density. Proceedings of 2011 IEEE International Conference on Cloud Computing and Intelligence Systems. Beijing, China. 2011. 113–117.

4 张万桢, 刘同来, 邬满, 等. 使用环形过滤器的 K 值自适应 KNN 算法. 计算机工程与应用, 2019, 55(23): 45–52, 85. [doi: 10.3778/j.issn.1002-8331.1905-0388]

5 张清清, 李长云, 李旭, 等. 基于不规则区域划分方法的 k-Nearest Neighbor 查询算法. 计算机系统应用, 2015, 24(9): 186–190. [doi: 10.3969/j.issn.1003-3254.2015.09.033]

6 刘星毅, 韦小铃. 基于欧式距离的最近邻改进算法. 广西科学院学报, 2010, 26(4): 409–411. [doi: 10.3969/j.issn.1002-7378.2010.04.006]

7 桑应宾, 刘琼荪. 一种基于特征加权的 K Nearest Neighbor 算法. 海南大学学报 (自然科学版), 2008, 26(4): 352–355.

8 文武, 李培强. 基于 K 中心点和粗糙集的 KNN 分类算法. 计算机工程与设计, 2018, 39(11): 3389–3394.

9 陆凯, 徐华. 基于最近邻距离权重的 ML-KNN 算法. 计算机应用研究: 1–5. http://kns.cnki.net/KCMS/detail/51.1196.TP.20190122.1326.007.html. (2019-09-28)[2019-12-18].

10 路敦利, 宁芊, 臧军. 基于 BP 神经网络决策的 KNN 改进算法. 计算机应用, 2017, 37(S2): 65–67, 88.

11 高亮, 谢健, 曹天泽. 基于 Kd 树改进的高效 K-means 聚类算法. 计算技术与自动化, 2015, 34(4): 69–74. [doi: 10.3969/j.issn.1003-6199.2015.04.015]

12 万家山, 陈蕾, 吴锦华, 等. 基于 KD-Tree 聚类的社交用户画像建模. 计算机科学, 2019, 46(S1): 442–445, 467.

13 刘云, 袁浩恒. 数据挖掘中并行离散化数据准备优化. 四川大学学报 (自然科学版), 2018, 55(5): 993–999.

14 董跃华, 刘力. 基于自适应改进粒子群优化的数据离散化算法. 计算机应用, 2016, 36(1): 188–193. [doi: 10.11772/j.issn.1001-9081.2016.01.0188]

15 姜楠, 周晓沧. 基于非线性规划的数据离散化方法及其应用. 清华大学学报 (哲学社会科学版), 2006, 21(S1): 54–59, 70.