## 基于数据挖掘的学生画像与学业预警方法研究

**摘要**

(介绍背景)随着物联网、大数据和人工智能技术的飞速发展，各行业信息化水平不断提高。信息化的发展为教育的变革提供了重要条件，而探索智慧课堂、构建智慧校园已成为学校变革的必然之路。《中国教育现代化 2035》中指出，结合大数据、机器学习等技术，构建一个数字化、智能化的智慧校园平台，实现智慧教育、智慧管理和智慧服务，是当前教育信息化发展的必然趋势。(提出问题)在智慧校园建设过程中，许多高校仍然沿用传统的学生管理方式，简单地按专业、成绩等条件划分学生，缺乏针对学生行为特点的个性化管理及服务。(本文提出的解决方法)本文将机器学习技术应用到学生画像和学业预警的研究中，构建学生数据画像和学业预警模型，帮助高校教育管理人员全面认识学生、精准预测学生和个性化服务学生。(研究成果)针对河南某高校的校园数据集，设计了基于数据挖掘的学生行为分析系统。系统共分为四个模块，分别是登录模块、数据处理模块、学生画像模块和学业预警模块。对系统进行实验与测试，结果表明该系统可以直观地展现在校学生的行为特点，能够帮助学校管理人员个性化服务与管理学生，具有一定的实用价值。

**1.Introduction**

(研究背景)智慧校园代表着在学校范围内，借助学生行为数据结合数据挖掘相关技术合成“智慧”，来提高学校教育管理人员对学生的服务与管理效率。《教育信息化 2.0 行动计划》[2]和《中国教育现代化2035》共同指出，结合大数据、机器学习等新一代信息技术，融合高校师生普遍需求，构建一个数字化、智能化的智慧校园平台，实现智能教育、智能管理和智能服务，是当前教育信息化发展的必然趋势。目前，高校在建设智慧校园的过程中依然面临着许多问题：

（1）在实现教育教学和信息管理的信息化过程中，高校内的各类信息平台收集了大量的学生数据。由于这些数据信息种类繁多、结构复杂，因此没有得到充分地利用。许多高校只开发了简单的查询功能，没有去挖掘这些学生数据背后隐藏的价值。

（2）许多高校仍然沿用传统的学生管理方式，简单地按专业、成绩等条件对生进行划分，缺乏依据学生行为特点的个性化管理及服务，远远不能满足当前智慧校园的发展需求。

**(本文研究内容、创新点)**针对这个问题，将大数据和机器学习技术应用到学生画像和学业预警的研究中，构建学生数据画像和学业预警模型，为高校教育管理人员提供相应的决策支持。本文主要工作如下：

（1）学生行为数据集的建立。本文通过对河南某高校各类信息管理系统进行调研分析，从校园一卡通系统、教务系统、图书馆门禁系统和图书馆借阅系统中采集学生的行为数据。对采集到的数据进行归纳总结和标准化处理，建立可以直接用于挖掘分析的学生行为数据集，为接下来学生画像和学业预警模型建立奠定数据基础。

（2）基于改进 K-means 的学生画像方法。使用传统 K-means 算法构建学生画像模型时，聚类簇数和初始聚类中心需要人工选择，因此聚类结果容易陷入局部最优。针对这个问题，对 K-means 算法进行了两点改进：引入 Canopy 算法，来确定 K-means 算法聚类簇数；引入最大最小距离算法，选择距离相对较远的 K 个样本作为 K-means 初始聚类中心，以此提高学生行为数据的聚类效果。

（3）基于改进 SVM 算法的学业预警方法。使用 SVM 算法构建学业预警模型时，惩罚参数γ和核参数 gamma 使用默认值会导致预测模型无法取得较高准确率。针对这个问题，采用改进的 FOA 算法对 SVM 模型的惩罚参数γ和核参数 gamma 进行寻优，来提高学业预警模型的准确率。

（4）基于数据挖掘的学生行为分析系统。研究了机器学习聚类和分类算法，利用本文构建的学生行为数据集，设计并实现了基于数据挖掘的学生行为分析原型系统。通过对学生画像和学业预警功能的测试，验证系统的有效性。

(论文结构)第二章介绍了本文研究的一些相关理论和方法，第三章详细介绍了改进学生画像的算法及初步实验分析，第四章介绍了改进学业预警的详细算法，进行实验并分析结果，第五章建立学生行为分析系统，进行可视化系统展示，第六章对本文进行总结。

**2.Related Work**

数据挖掘是一项比较复杂的系统工程，可以将其分为五个阶段：明确问题阶段、数据采集阶段、数据预处理阶段、模型建立阶段、模型解释与评价阶段[37-39]。数据挖掘流程图如下图2-1所示。

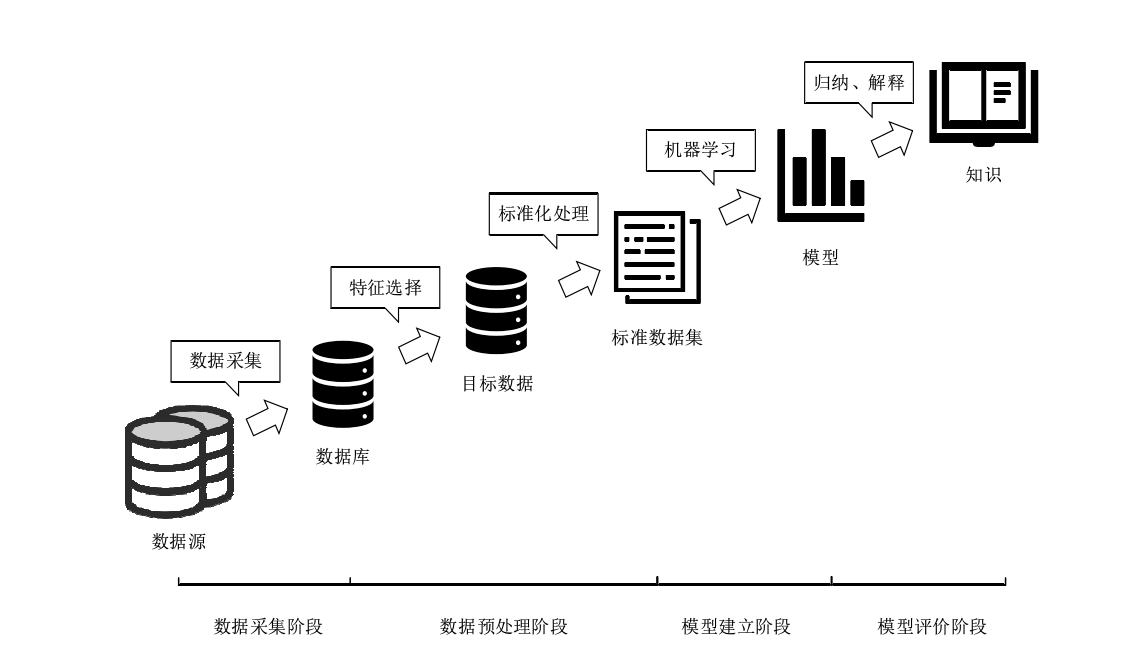


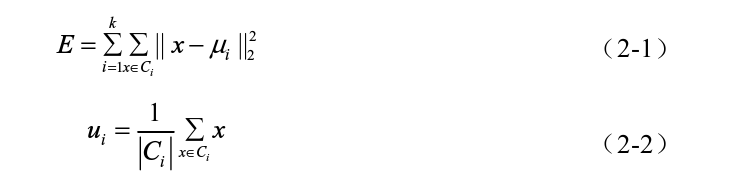
图2-1 数据挖掘流程图

**K-means算法**

K-means 是由 Macqueen 提出的一种迭代求解算法[53-55]，其中 K 指的是聚类簇数，而 Means 则被理解为每个聚类族数中的数据均值，因此 K-means 算法也常被称为是 K均值算法。虽然 K-means 已经被提出六十多年，但是由于其具备简单、高效等优点，至今仍然被广泛应用于各种领域，如文档聚类[56]、图像分割[57-59]、特征学习[60]等。

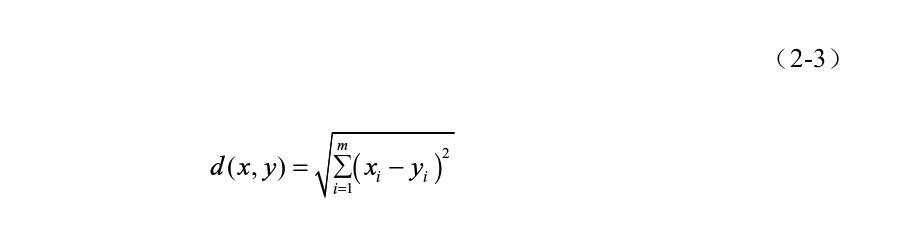
K-means存储用于定义聚类的k个质心，如果某一个点距离某个质心最近，则将该点归道该聚类中。

对于数据集D= { x1,x2,....,xm}K-means 算法可以根据数据集中样本与簇中心之间的距离远近，将数据集中所有样本划分为K个簇。聚类结果中簇中心集合C={C1,C2,....,Ck }的最小平方误差 E 计算方法如下所示。

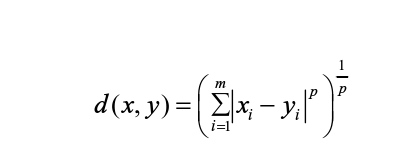


公式 2-1 中 E 值的大小代表着每个簇内样本点与聚类中心之间的紧密程度。而公式2-2 中ui 是Ci 的均值向量。K-means 算法计算数据集中样本之间距离的方法有许多，其中欧氏距离和闵可夫斯基距离这两种方法比较常用。

欧式距离的计算方法如公式2-3所示。



闵可夫斯基距离的计算方法如公式2-4所示。

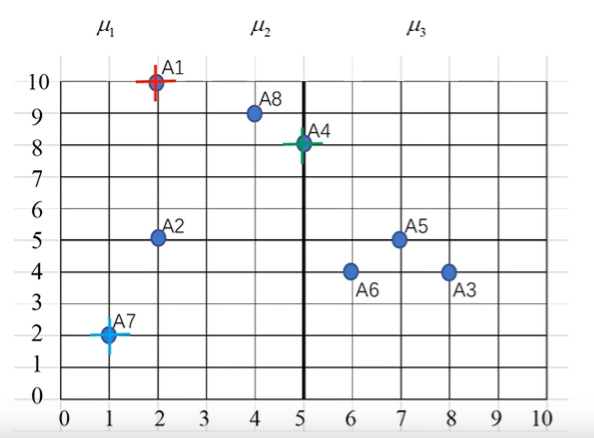
 （2-4）

举例：

利用k-means算法和欧氏距离，将下面8个点(使用（x，y）代表位置)聚类为三个簇。

A1(2,10), A2(2,5), A3(8,4),A4(5,8),A5(7,5),A6(6,4),A7(1,2),A8(4,9)

假设初始我们选择A1，A4，A7分别为每个簇的中心C1，C2，C3。则所有点的分布：



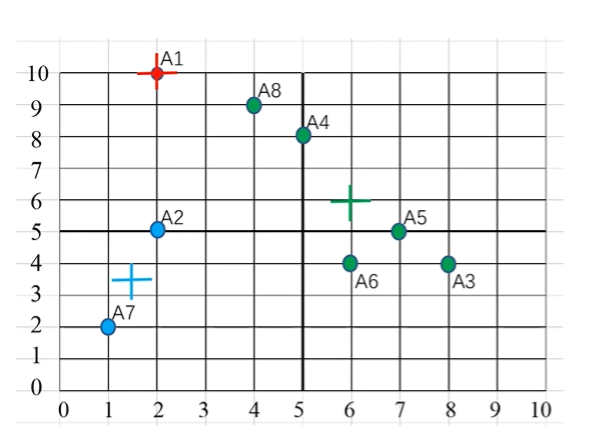
第一轮：（分别计算每个点到簇中心的距离）

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 簇  点 | A1 | A2 | A3 | A4 | A5 | A6 | A7 | A8 |
| C1 | 0 | 5 | 6 |  |  |  |  |  |
| C2 |  |  | 5 | 0 |  |  |  |  |
| C3 |  |  |  |  |  |  | 0 |  |

将每个点到各簇中心距离比较，取最小的属于该簇，属于C1用红色表示，C2用绿色，C3用蓝色。

确定新簇 μ1：{A1} μ2：{A3，A4，A5，A6，A8} μ3：{A2，A7}

确定新簇中心 C1=（2，10）；C2=（6，6）；C3=（1.5，3.5）

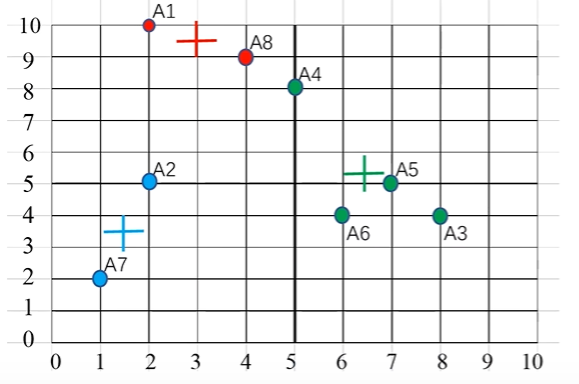


第二轮：

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 簇  点 | A1 | A2 | A3 | A4 | A5 | A6 | A7 | A8 |
| C1 | 0 | 5 | 6 |  |  |  |  |  |
| C2 |  |  |  |  |  | 2 |  |  |
| C3 |  | /2 |  |  |  |  | 2 |  |

新簇 μ1：{A1，A8}；μ2：{A3.A4，A5，A6}；μ3：{A2，A7}

确定新簇中心C1=（3，9.5）C2=（6.5，5.25）C3=（1.5，3.5）

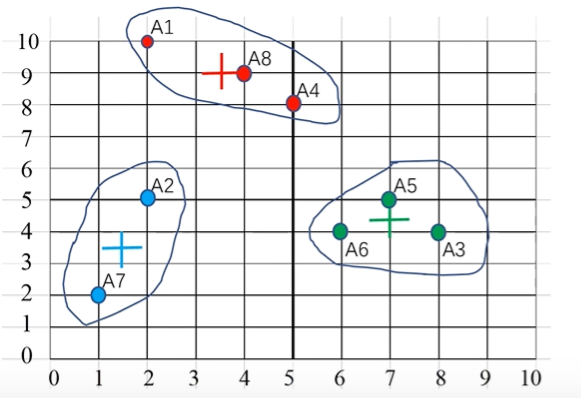


第三轮：

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 簇  点 | A1 | A2 | A3 | A4 | A5 | A6 | A7 | A8 |
| C1 |  |  |  |  |  |  |  |  |
| C2 |  |  |  |  |  |  |  |  |
| C3 |  |  |  |  |  |  |  |  |

新簇：μ1：{A1，A4，A8}；μ2：{A3，A5，A6}；μ3：{A2，A7}

确定新簇中心C1=（3.67，9）C2=（7，4.33）C3=（1.5，3.5）



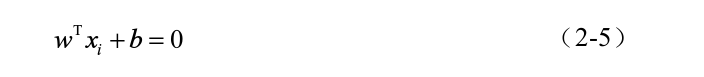
迭代收敛。算法结束。

(缺陷)K-means 算法对大型数据进行聚类分析时，往往能够取得不错的效果，但是 K-means算法也存在着一些问题，比如聚类簇数和初始聚类中心需要人为确认，如果这两个参数选择不合理，容易使算法陷入局部最优。

**SVM算法**

支持向量机 (Support Vector Machine, SVM) 是由 Vapnik 提出的一种二分类模型[60-62]，它的基本原理是通过核函数、稀疏损失等方法将低维空间中的原始数据映射到近似线性可分的一个高维空间中，然后在高维空间中基于最大间隔原则进行线性分类[63-64]。SVM 算法适用于小样本、多维度、非线性的情况，使其在文本分类[65]以及模式识别[66]等领域都有广泛的应用。

对于数据集 D={(x1,y1 ) ,( x2,y2)...., (xn,yn)}，SVM 算法的目标是找到一个超平面(w,b) ，将属于不同类型的样本分开。超平面计算方法如下所示。

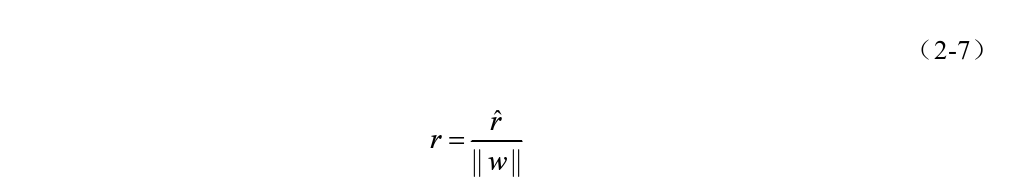


字母w代表着超平面的法向量，而字母 b 代表着数据集中样本点到超平面之间的垂直距离。

数据集中任意的样本点(xi,yi)到超平面 (w,b) 的函数间隔为表达式如下所示。



几何间隔为 r 表达式如下所示。



支持向量指的是数据集中距离超平面最近的点，两个支持向量之间的几何间隔如图2-2所示。

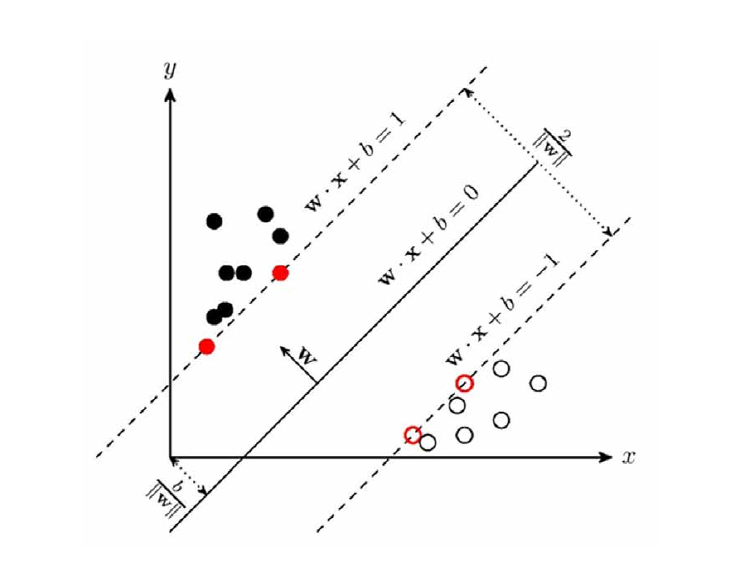
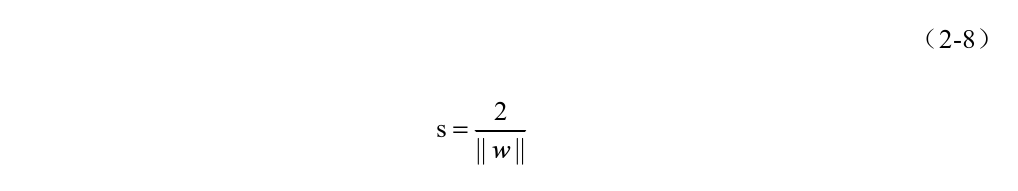


图2-2 支持向量机间隔

两个支持向量到超平面的垂直距离之和s如下所示。



线性不可分的SVM分类模型核函数使用高斯核，高斯核 SVM 分类模型有两个非常重要的参数γ与 gamma。其中γ为惩罚系数，也就是对误分类点的惩罚度，γ的值如果设置的太高，对误分类点的惩罚度就会很高，容易造成过拟合，反之γ如果设置的太低，对误分类点的惩罚度就会很低，容易造成欠拟合。而gamma 则是高斯核自带的一个参数，它决定了数据集中数据从低维空间映射到高维空间后的分布情况。如果 gamma 的值设置的过大，会导致训练误差和测试误差之间的差距太大，也就是过拟合现象；如果 gamma 设置的过小，则会造成平滑效应过大，使算法无法在训练集上取得较高的准确率。

**3.Methods**

K-means 算法在学生数据画像的应用中，存在 K 值和初始中心需要人为选择的缺点。针对这个问题，对 K-means 进行了两方面的改进。一方面，使用 Canopy 算法进行初步聚类，将结果提供给 K-means 作为 K 值的参考，消除了 K 值选取的不确定性。另一方面，利用最大最小距离算法选择距离较远的 K 个样本作为 K-means 的初始聚类中心，避免了随机选择初始聚类中心容易陷入局部最优的问题。

**聚类簇数的确定**

Canopy 算法是 McCallum 提出的一种快速聚类技术，虽然精度较低，无法给出精确的簇结果，但能给出最优的簇数量。Canopy 算法需要设置两个变量 T1 和 T2，随机选择初始聚类中心，计算样本与初始中心的欧氏距离，根据阈值 T1 和 T2 将初始数据集划分为多个不完全重叠的子集。算法的基本原理如图 3-1所示：

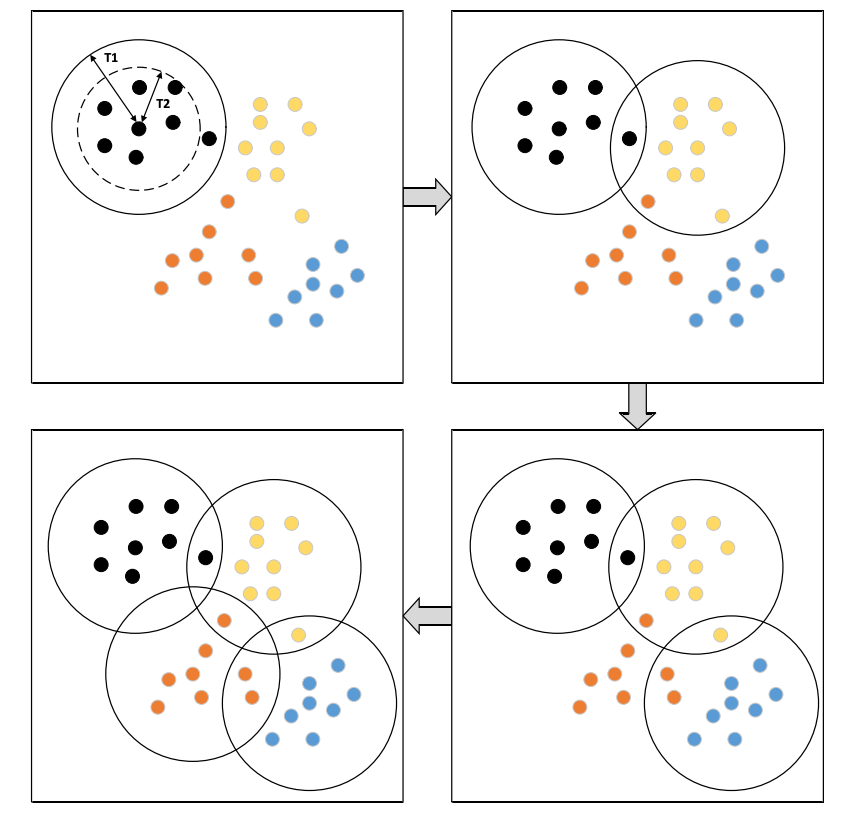


图3-1 Canopy算法原理

具体算法如下：

输入：数据集X={ x1,x2,...xn}

步骤 1：通过交叉验证调参或先验知识确定两个距离阈值T1和T2，其中T1大于T2。

步骤 2：从数据集 X 中随机选择一个样本点 P 作为第一个 Canopy 聚类中心，并将P 加入集合 C 中。从 X 中将 P 删除。

步骤 3：随机从集合 X 中选取一个样本 Q，并计算 Q 到集合 C 中每一个 Canopy 聚

类中心的欧几里得距离，选择这些距离中的最小值。

步骤 4：将T1与距离 d 作比较。如果 d 大于T1，则将样本点 Q 添加到与其距离最小的 Canopy 中，并标注一个弱标记。如果 d 小于T1，将 Q 设为新的 Canopy 聚类中心，将样本点 Q 添加到集合 C 中，并从数据集 X 中将 Q 删除。

步骤 5：将与距离 d 作比较。如果 d 小于T2，则为其附加一个强标记，将所有强标

记样本的中心位置更新为该 Canopy 的聚类中心，并从数据集 X 中删除样本点 Q。如果d 大于T2，将 Q 添加到当前 Canopy 集合当中。

步骤 6：重复步骤 3 至步骤 5，直到数据集 X 为空。

输出：簇中心集合 C 中样本个数 K

**初始聚类中心的选取**

在 K-means 算法中，初始聚类中心的选取是随机的。随机选取初始聚类中心可能会使算法陷入局部最优，并且如果初始聚类中心聚集在一起，会增加算法的迭代次数，耗费更多时间，因此初始聚类中心的选择是 K-means 算法改进的重要一环。 最大最小距离算法是一种基于试探的聚类算法，该算法可以使用欧氏距离选取尽可能远的 K 个聚类中心。采用最大最小距离算法为 K-means 算法选取尽可能远的样本点作为初始聚类中心，可以避免了 K-means 算法以避免陷入局部最优并且能够减少算法的迭代次数。

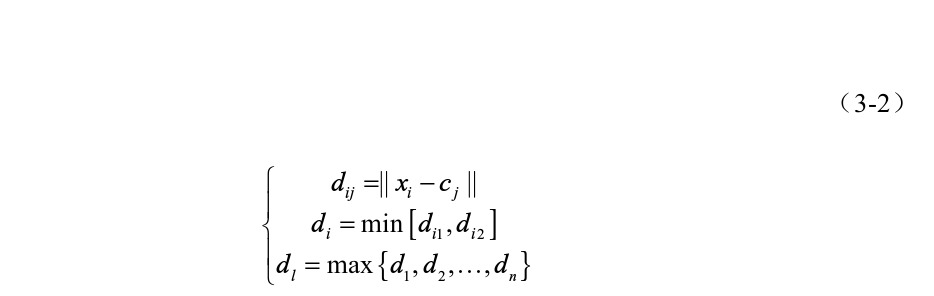
具体算法如下：

输入：数据集X={x1,x2,...,xn}，簇数 K（K≥3）

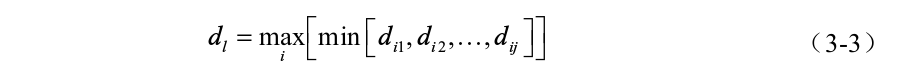
步骤 1：从数据集 X 中随机抽取一个样本，作为第一个聚类中心c1 。

步骤 2：使用欧式距离，选取距离c1 最远的样本作为第二个聚类中心c2 。

步骤 3：计算每个聚类中心与数据集 X 中的样本之间的欧式距离dij（其中i =1,2,…,n; j =1,2），求其中的最小值。选取{d1,d2,...,dn }中最大值dl对应的样本点，设置为第三个聚类中心c3。设 m =4 。



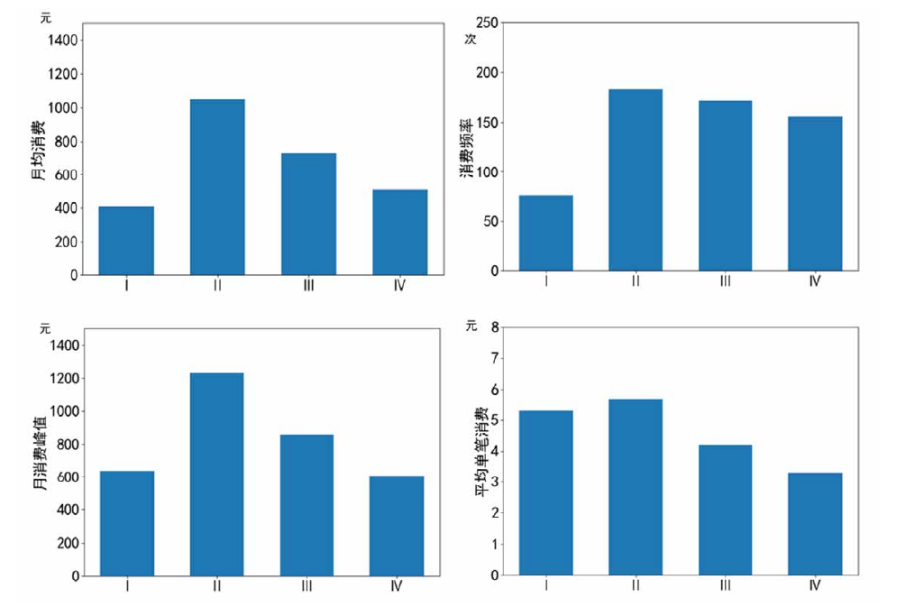
步骤 4：选择第 m 个聚类中心。对于 j =1,2,…,m −1，分别计算数据集 X 中的样本到集合{c1,c2,...,cm-1}中所有聚类中心的距离dij，并通过公式 3-3 计算dl。选择dl对应的样本作为第m个聚类中心cm 。



步骤 5：如果 m 等于 K，则输出聚类中心集 C；否则 m =m +1，重复步骤 4。

输出：初始聚类中心集合C={c1,c2,...,ck}

**3.3 实验结果分析**

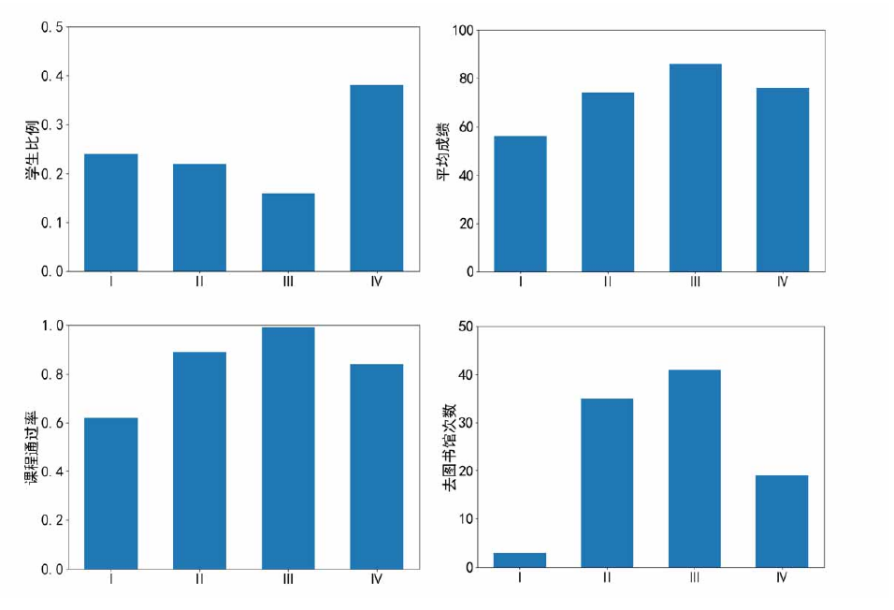


从图中可以看出来第一类学生月平均消费和消费频次在所有组中都是最低的，而单笔消费金额偏高，说明这类学生不经常在学校内消费。老师应核实这类学生在校情况，对这些经常外出的学生安全问题予以重视。

第二类学生生活费比较充足，生活水平高。

第三类学生相对比较稳定。

第四类学生月均消费和单笔消费都较低，消费频次高，说明这类学生在校消费活跃，但消费水平有限，比较节俭，在评定助学金等困难奖金时可以有限考虑。

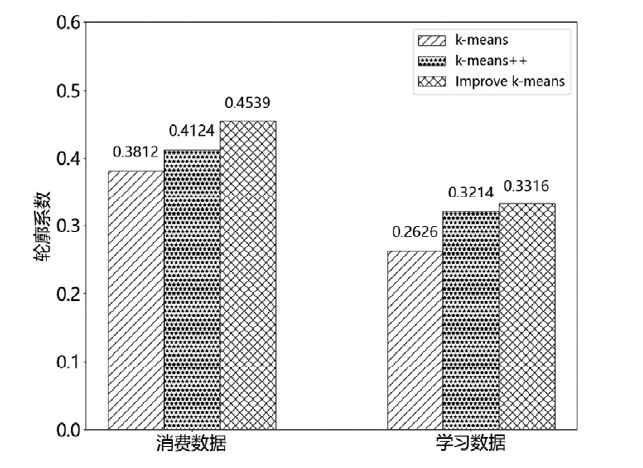


第一类学生平均成绩和课程通过率都较低，去图书馆次数也低，老师们可以纠正这类学生的学习态度等。

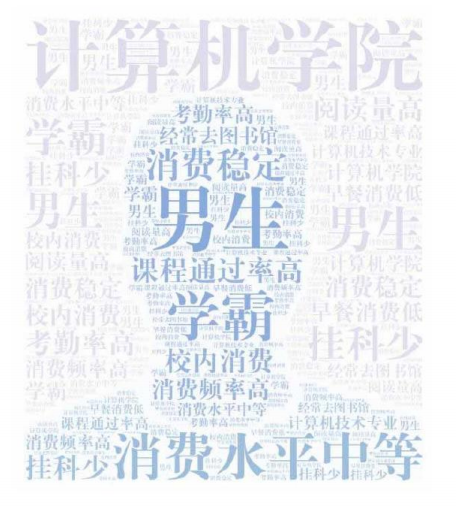
第二类学生处于中等偏上水平。

第三类学生从表中发现属于学霸类型。

第四类学生与其他几组相比属于中等水平。



通过对比可以发现改进后的K-means算法相较于原本算法轮廓系数较高，在聚类效果上有明显提高。



**基于改进FOA的SVM算法**

学业预警模型的数据集包含学生成绩数据、图书馆门禁数据、图书馆借阅数据和校园一卡通消费数据，为线性不可分数据，因此使用线性不可分的 SVM 分类模型，核函数使用高斯核。高斯核 SVM 分类模型有两个非常重要的参数γ与 gamma。其中γ为惩罚系数，也就是对误分类点的惩罚度，γ的值如果设置的太高，对误分类点的惩罚度就会很高，容易造成过拟合，反之γ如果设置的太低，对误分类点的惩罚度就会很低，容易造成欠拟合。而 gamma 则是高斯核自带的一个参数，它决定了数据集中数据从低维空间映射到高维空间后的分布情况。如果 gamma 的值设置的过大，会导致训练误差和测试误差之间的差距太大，也就是过拟合现象；如果 gamma 设置的过小，则会造成平滑效应过大，使算法无法在训练集上取得较高的准确率。 通过以上分析，可以看出惩罚系数γ和 gamma 参数的选择对 SVM 分类模型十分重要，在一定程度上能够影响 SVM 分类模型的准确率。本文计划采用 FOA 算法来对 SVM分类模型进行超参数寻优，选择最佳的惩罚系数γ和 gamma 参数，并针对 FOA 算法每次迭代搜索步长范围固定的缺点，提出了一种自适应步长的方法来提高 FOA 算法的效率。

**FOA算法（果蝇优化算法）**

果蝇在寻找食物时，会利用敏锐的嗅觉寻找空气中食物的气味，将发现的食物位置分享给其他果蝇并接受其他果蝇分享来的信息，对收集到的浓度信息进行比较后找出味道浓度最高的位置，再利用发达的视觉发现食物和其他果蝇聚集的位置，向着这个位置飞过去，不断地重复这个流程来寻找食物味道浓度最高的位置。FOA 算法是一种简单可行、参数较少、易于实现的优化算法，在一般情况下可以获得全局最优解，算法的迭代演化路径如图 3-3所示。

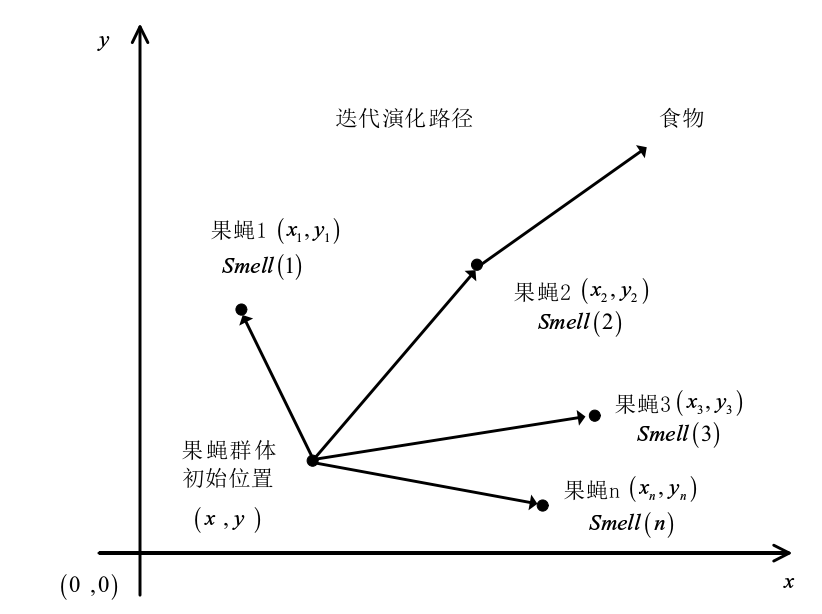


图3-3 FOA算法迭代演化路径

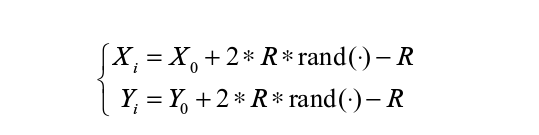
FOA 算法的基本流程如下。

输入：果蝇的最大迭代次数 K，果蝇的种群规模 N，搜索步长 R

步骤 1：随机选取果蝇群体的初始位置(x0,y0)。

步骤 2：果蝇群体以初始位置(x0,y0)作为起点，使用随机方向和距离来搜寻食物的

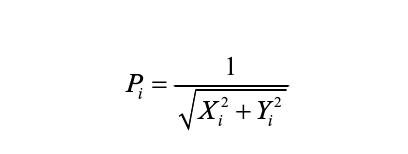
位置，设果蝇群体中第 i 只果蝇找到新的位置为(xi,yi)，计算公式如下所示。



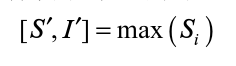
Rand(·)指随机长生一个[0，1]之间的数。

步骤 3：计算出果蝇群体中每个果蝇位置到原点之间的距离，将距离的倒数作为该

果蝇的味道浓度判定值Pi。具体计算公式如下所示。

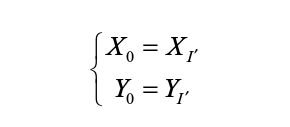


步骤 4：将果蝇群体中每个果蝇的味道浓度判定值带入到浓度判定函数中去，计算出每个果蝇的味道浓度Si。这里设味道浓度越大越好，比较果蝇群体中每个果蝇的味道浓度，找出味道浓度最高的果蝇，保留该果蝇的味道浓度值和位置。



步骤 5：保留最佳味道浓度值，果蝇群体其他果蝇都朝着味道浓度值最高的置飞去。

使用公式更新初始位置(x0,y0)。



步骤 6：重复步骤 2 至步骤 4，如果搜索到的最高味道浓度优于前一代且算法迭代

次数小于最大迭代次数 K，则执行步骤 5；如果不是，则输出最佳味道浓度值，结束算

法。

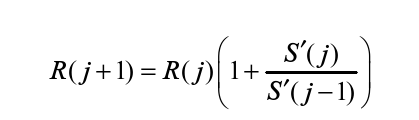
**基于改进 FOA 的 SVM 参数优化**

使用 FOA 算法对 SVM 分类模型进行参数寻优时，由于每次迭代的步长搜索范围是固定的，导致果蝇群体无法快速找到参数最优值，并且结果容易陷入局部最优。针对这个问题，对 FOA 算法进行了改进，引入自适应步长的概念，使 FOA 算法的搜索步长根据当代最优味道浓度值和上一代最有味道浓度值的比值来变化。

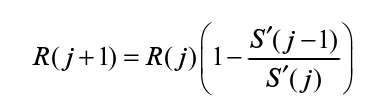
这里设味道浓度越大越好，j 为算法的迭代次数，而 R( j)则是算法第 j 次迭代的搜索步长，S′( j)是第 j 次迭代的最高味道浓度。

当迭代次数 j＝１时，设 R(1) =R(2) =2，而 S′(1) 为算法第一次迭代找到的最高味道浓度值。

当迭代次数 j≥２时，如果本次迭代的最高味道浓度值小于上一次迭代的最佳味道浓度值，则表示上一次迭代的最高味道浓度优于本次迭代的最高味道浓度值，这时可以增大步长，提高算法的全局搜索能力。下一次迭代的步长 R( j +1) 计算方法如下所示。

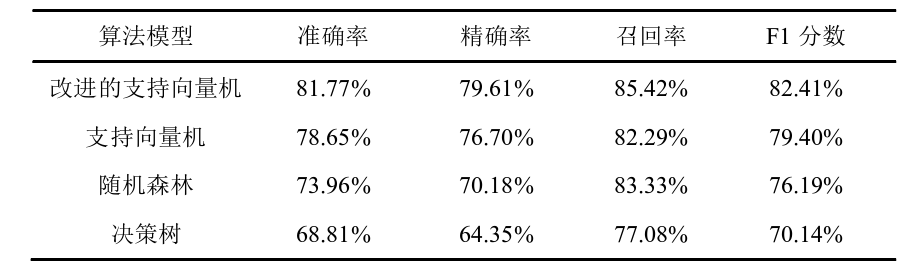


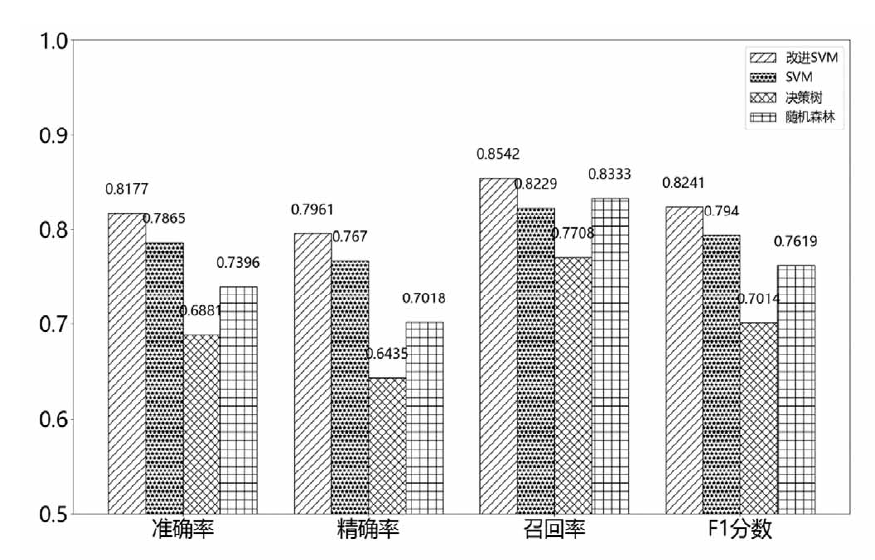
如果本次迭代的最高味道浓度值大于上一次迭代的最佳味道浓度值，则表示本次迭代的最佳味道浓度值优于上一次迭代的最高味道浓度值，这时可以考虑减小步长，提高算法的精度。下一次迭代的步长 R( j +1) 计算方法如下所示。



通过改进 FOA 算法对高斯核 SVM 分类模型中惩罚系数γ和参数 gamma 进行寻优。

**实验结果分析**





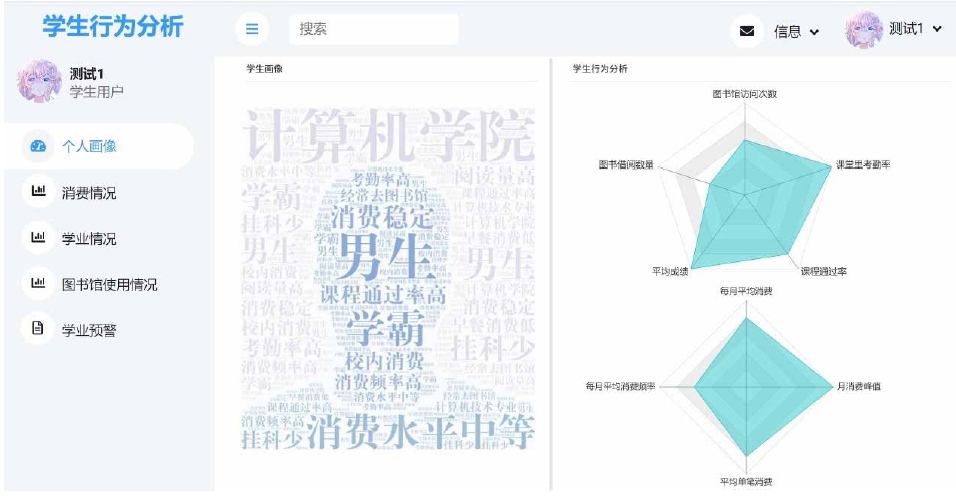
从表和图中可以清楚的看出，改进后的SVM分类模型在准确率、精确率、召回率和 F1 分数四个指标上均优于决策树、随机森林和传统 SVM 算法，可以有效的对无法顺利毕业的学生进行预警，对高校学生的教育管理具有重要意义。

**Results(可视化展示)**

将处理后的数据进行可视化展示。

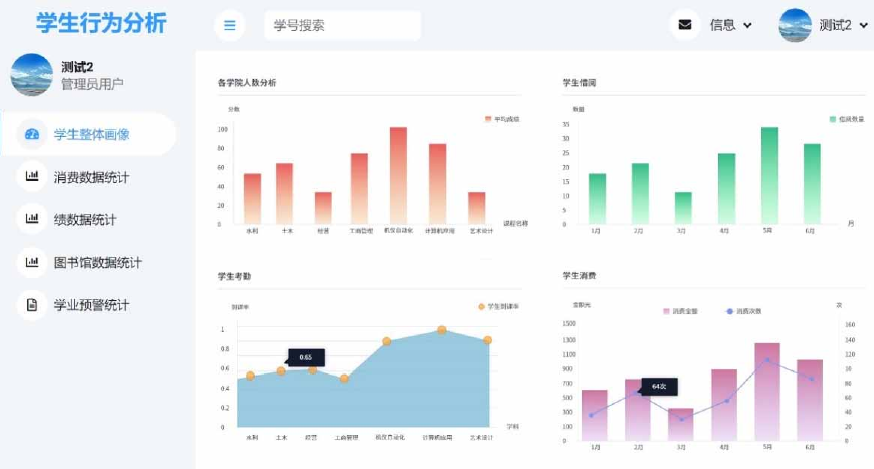
学生个人用户







管理员用户











**Conclusion**

随着数字化校园的不断发展，高校内各类信息管理系统中积累的学生数据越来越多，使用传统的数据统计方法无法有效地分析这些学生数据，高校管理人员必须寻找新的方法来挖掘学生的行为特点，实现学生的个性化管理与服务，推动智慧校园的发展。本文以机器学习算法在校园数据挖掘中的应用为中心，对学生数据画像和学业预警两个问题进行研究，主要工作总结如下：

（1）本文介绍了学生数据挖掘的背景及意义，分析了学生画像和学业预警问题的研究现状及目前所存在的不足之处，解释了数据挖掘技术、机器学习相关算法和数据预处理的概念及流程。

（2）本文从校园一卡通系统、教务系统、图书馆门禁系统和图书馆借阅系统中采集学生的行为数据。对采集到的数据进行归纳总结和标准化处理，建立可以直接用于挖掘分析的学生行为数据集。

（3）针对 K-means 算法在学生画像应用中存在的不足，提出了一个基于 Canopy 和最大最小距离原则的 K-means 算法，将算法应用于对学生行为数据的聚类分析。实验结果表明，基于 Canopy 和最大最小距离原则的 K-means 算法可以有效区分不同行为特征的学生，相对于其它聚类算法具有更高的轮廓系数。将算法聚类结果使用词云的方法进行可视化展示，帮助学校更加全面地认识学生。

（4）针对 SVM 算法在构建学业预警模型时存在的不足，提出了一种基于改进 FOA的 SVM 算法。使用基于改进 FOA 的 SVM 算法预测学生未来能否顺利毕业，对未来可能无法顺利毕业的学生发出学业预警。实验表明基于改进 FOA 的 SVM 预警模型在准确率、召回率等方面优于传统 SVM、决策树和随机森林三个分类模型。

（5）设计并实现了基于数据挖掘的学生行为分析系统，介绍了系统的工作流程和系统框架，对系统的登录模块、学生数据画像模块和学业预警模块进行可视化展示。