基于杂交聚类算法的学生状态分析系统研究*

肖立中[a]，胡 婷[b]，刘云翔[a]，林振骏[a]，吴雁林[a]
（上海应用技术学院 a. 计算机科学与信息工程学院；b. 电气电子与工程学院，上海 201418）

摘　要：针对高校学生人数增多、学生信息量增大，原有的学生管理和评估模式已经不能满足需求的问题，研究了学生状态分析系统，实现了学生综合数据的采集，为辅导员与学生提供了交互的平台；针对 K-均值聚类算法在全局优化中的不足，研究了基于粒子群的 K-均值聚类算法对学生综合数据进行分析。通过与 K-均值聚类算法、基于遗传算法的 K-均值聚类算法和人工评估的结果比较可以看出，本系统对学生的状态评估更加全面和客观；同时系统提供了便于查询的视图信息反馈，帮助辅导员及时发现问题、提高工作效率。

关键词：学生管理；学生状态评估；粒子群优化算法；K-均值算法
中图分类号：TP301　　**文献标志码：**A　　**文章编号：**1001-3695(2012)06-2168-04
doi:10.3969/j.issn.1001-3695.2012.06.044

Student status analysis system research based on hybrid clustering algorithm

XIAO Li-zhong[a], HU Ting[b], LIU Yun-xiang[a], LIN Zhen-jun[a], WU Yan-lin[a]
(*a. School of Computer Science & Information Engineering, b. School of Electrical & Electronic Engineering, Shanghai Institute of Technology, Shanghai* 201418, *China*)

Abstract: With the number of students and their information increasing, the original student management and evaluation pattern could not meet demand in college. This paper researched a student status analysis system, which could collect the students' comprehensive data and provide platform to communicate between counsellor and students. In view of the deficiency of global search ability for K-means clustering algorithm, this paper proposed K-means algorithm based on particle swarm optimization to analyze the students' data. Compared with those of K-means algorithm, K-means algorithm based on genetic algorithm and artificial evaluation, the results show the evaluation obtained from the system is more comprehensive and objective. The system could also offer visual information convenient to query, which helped counsellor find sticking point timely and improve efficiency.

Key words: student management; student status evaluation; particle swarm optimization algorithm; K-means algorithm

0　引言

随着高校的扩招，学生人数不断增多，单个辅导员管理的学生越来越多，辅导员很难准确了解每个学生的状态。同时，随着高校图书馆资源的增加、实验室的开放、网络环境的优化、课外活动的丰富等，学生的学习和生活环境发生了较大的变化，有关学生的各种信息量也成倍增长，原有的学生管理模式，即单纯以考试成绩评估学生的方式已经不能适应目前高校学生的管理。

本文设计的学生状态分析系统基于 B/S 架构，能够收集学生的各类信息，并以图形或报表的形式反馈给辅导员，方便辅导员的查询和了解。同时，对收集到的学生信息进行了基于粒子群的 K-均值(PSO-KM)聚类算法的分析，对学生的状态进行了评估。系统还提供了辅导员和学生的交互平台，方便了学生和辅导员的沟通；另外，系统根据辅导员使用系统的情况对辅导员的工作进行了评估。

1　系统介绍

系统包括三个子模块：数据采集模块、留言互动模块和数据分析模块。系统架构如图 1 所示。

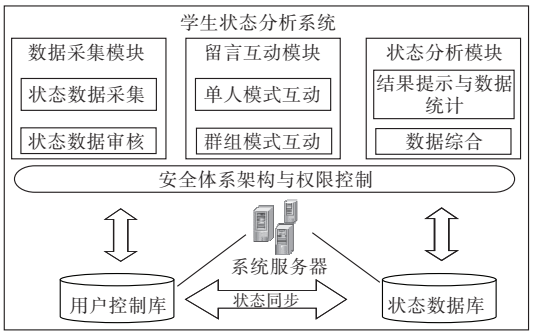

图 1　系统架构

数据采集模块主要包括数据的采集与审核；留言互动模块[1,2]主要包括单人留言和群组留言；状态分析模块主要是对数据的统计呈现以及利用算法对数据进行处理并显示综合结果。这三个子模块都通过安全与权限控制同数据库以及外界进行交互，保证了状态数据的可靠性以及真实性。数据库主要分为两部分，一部分专注于处理用户权限，另一部分专注于用户状态数据，这两部分数据库通过 SQL Server 服务器进行数据同步。

收稿日期：2011-11-03；**修回日期：**2011-12-19　**基金项目：**上海市教育委员会科研创新资助项目(12YZ164)；上海应用技术学院校内科技发展基金资助项目(KJ2011-03)；上海应用技术学院计算机科学与技术重点学科资助项目(1020Q121002)

作者简介：肖立中(1981-)，男，山东寿光人，副教授，博士，主要研究方向为数据挖掘、软件工程(lyexp@163.com)；胡婷(1981-)，女，讲师，硕士；刘云翔(1964-)，男，教授，博士，主要研究方向为人工智能、计算机软件与理论、信息融合；林振骏(1990-)，男，本科生；吴雁林(1989-)，男，学士.

1.1　权限管理

系统包括四类人员,即学校管理员、学院管理员、辅导员和学生。主要权限设置如下:

a)学校管理员,初始数据的输入及对学院管理员的人员管理。

b)学院管理员,用户的角色控制、对辅导员工作的考评、对系统分析得出的问题学生的处理。

c)辅导员。(a)维护班级和课程信息;(b)导入学生信息数据,如考试成绩、宿舍门禁、图书馆门禁数据等;(c)审核以及修改学生提交的非敏感信息,如第二课堂信息、获奖信息等;(d)与学生进行留言互动;(e)查看状态图表;(f)查看分析报表。

d)学生,输入非敏感信息、查看或发布留言。

1.2　数据采集

对学生的评估不能仅从成绩上下定论,应该进行综合考虑。本系统主要从学生课堂情况和课余生活两个方面出发,选取了包括竞赛获奖、参加讲座、图书馆学习、体育锻炼,课程的旷课、迟到、早退、作业、成绩以及品德评价和娱乐情况,对学生进行综合评估。

以上数据中一部分通过与学校相关管理系统接口同步或数据导入获取,一部分通过学生输入、辅导员审核来获取。

1.3　留言互动

通过权限的管理,管理员、辅导员和学生被分配到社区,在社区平台上进行公开或私密的交流,克服了空间的障碍以及面对面交流的不便。交流内容经过统计处理后用于对辅导员工作的评估,同时交流内容可以作为原始数据资料保存以便分析。

1.4　状态分析

数据采集模块获取了学生课堂和课余的一些原始数据,而状态分析模块则是要对学生的信息进行挖掘,以便对学生状态进行评估。本文系统中的学生信息包括 11 个方面,但是由于每学期有多门课,构成了多维的样本空间,每个学生样本形成一个向量。通过计算向量之间的相似性来度量学生间的相似性,从而对学生进行无监督分类,最后通过对类群的评估完成对学生的评估。

2　状态分析算法

2.1　K-均值算法

K-均值(K-means)算法是一种无监督的聚类算法,它将相似的数据划分到同一个聚类中,而将不相似的数据划分到其他的聚类。

设模式样本集为 $X=\{X_i, i=1,2,\cdots,n\}$,其中 X_i 为 D 维模式向量,聚类问题就是要找到一个划分 $C=\{C_1,C_2,\cdots,C_k\}$,满足

$$X=\bigcup_{i=1}^{k} C_i$$

$$C_i \neq \varnothing \quad i=1,2,\cdots,k$$

$$C_i \cap C_j = \varnothing \quad i=1,2,\cdots,k;\ i\neq j$$

并且使得目标函数

$$J_c=\sum_{j=1}^{k}\sum_{X_i\in C_j} d(X_i,Z_j) \tag{1}$$

达到最小。其中,Z_j 为第 j 个聚类的中心;$d(X_i,Z_j)$ 为样本到对应聚类中心的距离;J_c 为各类样本到对应聚类中心距离的总和。这里 $d(X_i,Z_j)$ 采用欧氏距离。

K-均值算法作为一种有效的聚类算法已经获得了广泛的应用,但是它通常从数据对象中随机选择作为初始聚类中心,并且产生的是局部最优解,而不是全局最优解,这使得产生的聚类结果具有很大的不确定性[3,4]。针对这一情况,本文采用了 PSO-KM 算法。

2.2　粒子群优化算法

粒子群优化(PSO)算法同遗传算法类似[5,6],是一种基于群体的优化工具。系统初始化一组随机解,通过迭代搜寻最优值,但是它没有遗传算法用的交叉以及变异操作,而是通过粒子(潜在的解)在解空间追随最优的粒子进行搜索。与遗传算法比较,它的优势在于容易实现同时又有深刻的智能背景,既适合科学研究,又特别适合工程应用。

PSO 算法是由 Kennedy 等人[7]于 1995 年提出的一种基于迭代的优化方法。在 PSO 中,每个粒子都是 D 维解空间的一点,并且都具有一个速度,不同粒子具有对应目标函数相关的个体适应度。每个粒子根据自身的经验和群体的经验来调整自己的轨迹,向最优点靠拢。对于某个粒子 i,它的位置表示为 $X_i=(x_{i1},x_{i2},\cdots,x_{iD})$,速度 $V_i=(v_{i1},v_{i2},\cdots,v_{iD})$,它所经历过的最好位置记为 $\mathrm{Pbest}_i=(\mathrm{pbest}_{i1},\mathrm{pbest}_{i2},\cdots,\mathrm{pbest}_{iD})$,它所能获取的群体的最好位置记为 $\mathrm{Gbest}=(\mathrm{gbest}_1,\mathrm{gbest}_2,\cdots,\mathrm{gbest}_D)$。在每一步中,粒子根据以下公式更新自己的速度和位置:

$$V_i(t)=wV_i(t-1)+c_1r_1(\mathrm{Pbest}_i-X_i(t-1))+c_2r_2(\mathrm{Gbest}-X_i(t-1)) \tag{2}$$

$$X_i(t)=X_i(t-1)+V_i(t-1) \tag{3}$$

其中:w 称为惯性权值;c_1、c_2 是两个正常数,称为加速因子;r_1 和 r_2 是两个[0,1]间的随机数,通常使用一个常量 $v_{d\max}$ 来限制粒子的速度,改善搜索结果。w 起着平衡全局搜索能力与局部搜索能力的作用,对于算法的成功至关重要。Shi 等人[8]发现较大的 w 值有利于跳出局部极小点,而较小的值有利于算法收敛,从而提出了一种根据算法叠代次数使惯性权值线性递减的方法。算法在初期使用较大惯性权值,具有较强全局搜索能力;后期则使用较小惯性权值,提高局部搜索能力。w 的计算公式如下:

$$w=(w_1-w_2)\times\frac{\max_\mathrm{iter}-\mathrm{iter}}{\max_\mathrm{iter}}+w_2$$

其中:w_1 和 w_2 是惯性权值的初始值和最终值;max_iter 和 iter 分别为算法最大叠代次数和当前次数。

PSO 的基本算法步骤描述如下:

初始化粒子群。在 D 维搜索空间中随机设定各粒子的初始位置 X_i 和初始速度 V_i;计算每个粒子的适应度值;把每个粒子的当前位置设为 Pbest_i,把具有最佳适应度值的位置设为 Gbest。

```
for t = 1 to max_iter
w_t = (w_1 - w_2) * ((max_iter - t)/max_iter) + w_2
for i = 1 to N            // N 为粒子数目
  for d = 1 to D
```

$v_{id}(t) = w_t v_{id}(t-1) + c_1 r_1 (pbest_{id} - x_{id}(t-1)) +$
$c_2 r_2 (gbest_d - x_{id}(t-1))$
$v_{id}(t) = \min(v_{dmax}, \max(-v_{dmax}, v_{id}(t)))$
$x_{id}(t) = x_{id}(t-1) + v_{id}(t-1)$
end for d
if $X_i \in [X_{min}, X_{max}]$
计算 X_i 的适应度值,必要时更新 $Pbest_i$ 和 Gbest
end if
end for i
达到结束指标就停止
end for t

2.3 PSO-KM 算法的编码与适应度选择

粒子群算法采用实数编码,一个编码对应于一个可行解。在本文的算法中,采用的是基于聚类中心的编码方式,也就是每个粒子的位置是由 k 个聚类中心组成,粒子除了位置之外,还有速度和适应值。样本向量维数为 D,因此粒子的位置是 $k \times D$ 维变量,速度也应当是 $k \times D$ 维变量。这样,粒子就可以采用以下编码结构[9]:

$$\begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1D} \\ z_{21} & z_{22} & \cdots & z_{2D} \\ \vdots & \vdots & & \vdots \\ z_{k1} & z_{k2} & \cdots & z_{kD} \end{bmatrix}, \begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1D} \\ v_{21} & v_{22} & \cdots & v_{2D} \\ \vdots & \vdots & & \vdots \\ v_{k1} & v_{k2} & \cdots & v_{kD} \end{bmatrix}, f(S)$$

其中:z_{ij}表示第 i 个聚类中心的第 j 维变量;$f(S)$ 为适应度函数,本文构造了一个平衡类中距与类间距的函数:

$$f(S) = J_c + \frac{1}{k} d(Z)$$

其中:J_c 为 K-均值算法的目标函数,其计算方法如式(1)所示;$d(Z) = \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} d(Z_i, Z_j)$ 为各聚类中心距离之和。$f(S)$ 越小,说明聚类质量越好。

在判断聚类结果是否为最优时,仅使用类中距 J_c 是不够的[10]。因为分类的目标是使类中距尽可能小,同时使类间距尽可能大,然而当聚类个数 k 增加时,类中距呈递减趋势而类间距总体呈增加趋势。为了限制类中距单调,本文构造的适应度函数加上了惩罚因子 $d(Z)$,使得适应度函数能平衡类中距和类间距。

2.4 PSO-KM 算法描述

a)将每个样本随机指派为某一类,作为最初的聚类划分,并计算各类的聚类中心,作为初始粒子的位置编码,计算粒子的适应度,并初始化粒子的速度,反复进行 N 次,共生成 N 个初始粒子群。

b)对每个粒子,比较它的适应度值和它经历过的最好位置 $Pbest_i$ 的适应度值及群体所经历的最好位置 Gbest 的适应度值,如果更好,更新 $Pbest_i$ 和 Gbest。

c)根据式(2)(3)调整粒子的速度和位置。

d)进行 K-均值处理。

e)按照处理结果,计算新的聚类中心,更新粒子的适应度值,取代原来的编码值。

f)如果达到结束条件(足够好的位置或最大迭代次数),则结束,否则转步骤 b)。

PSO-KM 算法在产生下一代解群时有较大的随机性,所以不容易陷入局部极小值,而且每代种群中的解具有自我学习提高和向他人学习的双重优点,所以具有较快的收敛速度。由于不存在随机寻优的退化现象,后期收敛比较平稳,很少有波动现象。

3 实验与实验结果

3.1 实验数据描述

通过数据采集模块,获取了计算机学院 07 级学生八个学期的数据。由于相同专业的学生更具有可比性,本文将软件工程专业 80 个学生作为例子进行实验。

3.2 数据预处理

由于不同的学生状态数据属性有不同的度量标准,如果不对原始数据进行预处理,就有可能产生大数“吃”小数的问题;另外,收集到的多维数据对于评估学生的影响力是不同的。为了解决以上问题,本文进行如下变换:

$$y_{ij} = \frac{a_j x_{ij}}{\max_j}$$

其中:x_{ij}为收集到的第 i 个学生的第 j 个属性;$\max_j$ 为属性的最大值,这样使得各属性的值都映射到[0,1]间;a_j 为属性的权值,该属性由学校的办学定位确定,不同学校可以根据自身对学生的培养要求进行调整。

3.3 学生状态评估结果

为了对本文算法的聚类能力进行评估,针对数据集实施了 K-均值算法、基于遗传算法的 K-均值(GA-KM)算法和 PSO-KM 算法,比较结果如图 2 所示。

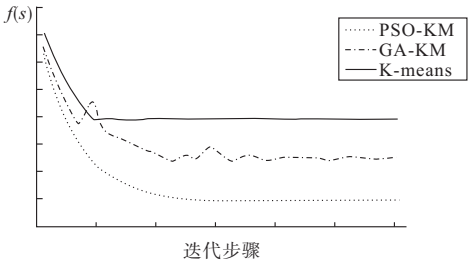

图 2 三种算法的聚类结果

从图中可以看出,在同等条件下,K-means 收敛速度很快,但很容易陷入局部最小值,而 PSO-KM 和 GA-KM 算法寻优能力优于 K-means。由于粒子群算法不存在随机寻优的退化现象,因此与 GA-KM 相比,PSO-KM 后期收敛比较平稳,很少有波动现象。同时可以看出,GA-KM 算法产生了早熟现象,而 PSO-KM 算法通过调整加速因子和惯性权值减轻了早熟的影响。

经过 PSO-KM 算法处理后,学生样本被分为四类。为了评估四类学生的状态,本文进行了类别的横向评估,对于四类的聚类中心作如下综合评估:

$$\tau_i = \sum^{j} z_{ij}$$

其中:z_{ij}为第 i 类聚类中心的第 j 个属性。对 τ_i 作降序排列,即对应 A、B、C、D 四个等级。

为了验证算法的正确性,本文对 80 名学生收集了八个学期来自教师、辅导员和同学的评估,并与基于 K-均值算法、基于 GA-KM 算法和本系统的结果进行对比,对比情况如图 3 ~ 5 所示。其中匹配度用如下公式计算:

$$匹配度=\frac{算法评估与人工评估相同的学生数}{算法评估类别的学生数}$$

由于 K-均值算法产生的是局部最优解,另外,它受初始聚类中心的影响比较大,产生的聚类结果具有很大的不确定性。本文使用 K-均值算法对数据集进行了 10 次实验,每次得到的聚类结果都不一样,图 3 是对匹配度计算了均值所得的结果。而用 GA-KM 和 PSO-KM 算法对数据集进行了 10 次实验,得到的聚类划分都一致,稳定性很好。通过图 3 ~ 5 的对比可以看出,本文算法的匹配度高。

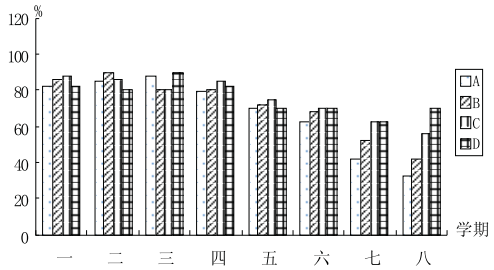

图3　基于K-均值算法评估匹配度

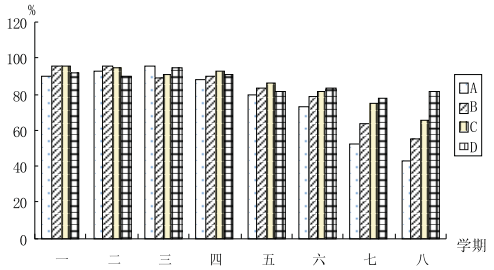

图4　基于GA-KM算法评估匹配度

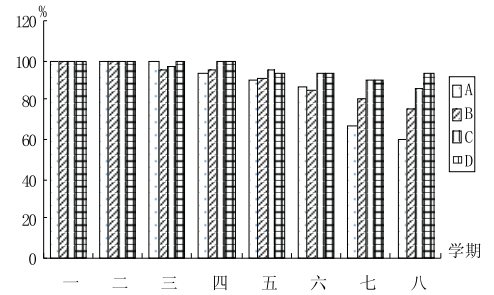

图5　系统评估匹配度

由图 5 可以看出,一至四学期匹配程度比较高,这是因为这四个学期中学生的活动主要以课堂学习、娱乐活动、体育锻炼等为主,教师、辅导员和同学的评估基本上能够反映学生的状态,也说明系统的准确率比较高。从第五学期匹配度开始下降,综合分析发现,第五学期开始,学生参加竞赛、讲座、图书馆学习等的内容比较多,由于参加评估的教师、辅导员和同学了解的内容有一定局限性,造成了评估的差异。经过对评估差异学生数据的人工分析,发现系统评估更具有说服力。第八学期匹配度最低,这是因为该学期学生主要以实习和毕业设计为主,系统收集的数据较少。A 类学生的匹配度不高,这是因为该类学生数量较少,少量评估差异的学生就能引起较大的不匹配度。

得到分类结果后,系统可以查询每个学生状态的变化曲线,从而可以在学生状态变化的关键时期对其进行引导。图 6 给出了一名学生状态变化的曲线图。

由图 6 可知,该学生从第二学期开始状态变差,通过系统可以查询该生的各项数据统计明细,第一学期和第二学期的数据如表 1 所示(隐去学号和姓名)。

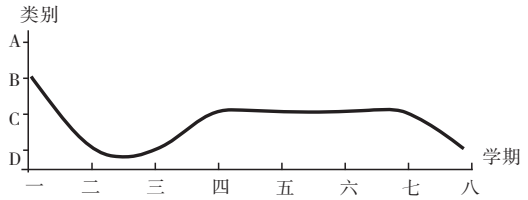

图6　学生状态变化曲线图

表 1　学生综合数据统计明细

学期	课堂扣分	作业情况	成绩绩点	课余生活	竞赛情况	品德情况
一	-43.13	52.5	2.68	43.79	64.42	48.4
二	-20.35	40	2.1	-3.8	64	48

由表 1 可以看出,该生状态变差的原因包括课余生活没有合理安排(娱乐时间过长、自习时间较少),虽然课堂扣分情况有所好转,但是作业情况有所下降,成绩也有所下降。因此,辅导员在工作中就需要对该生的课余生活方面给予更多的关心,并督促其按时完成课后作业,从而引导学生向更好的方向发展。

4　结束语

针对学生状态的综合评估问题,本文研究并实现了学生状态分析系统,实现了学生状态数据的采集、辅导员与学生的交互,并研究了基于粒子群的 K-均值聚类算法用于学生状态的无监督分析。通过实验表明,该系统能够对学生进行较全面的综合评估,方便对学生进行引导和管理,提高了学生管理的效率。另外,除系统收集的学生状态数据外,学生的实习、校外实践等也是重要的数据,更全面的数据采集是将来的研究内容。

参考文献:

[1]　郭靖. ASP. NET 开发技术大全[M]. 北京:清华大学出版社,2009.

[2]　求是科技. ASP 信息管理系统开发实例导航[M]. 北京:人民邮电出版社,2005.

[3]　毛韶阳,李肯立. 优化 K-means 初始聚类中心研究[J]. 计算机工程与应用,2007,43(22):179-181,219.

[4]　胡艳维,秦拯,张忠志. 基于模拟退火与 K-均值聚类的入侵检测算法[J]. 计算机科学,2010,37(6):122-124.

[5]　GOLDBERG D E. Genetic algorithms in search,optimization and machine learning[M]. Boston:Addison-Wesley Longman Publishing,1989:160-173.

[6]　THIERENS D. Scalability problems of simple genetic algorithms[J]. Evolutionary Computation,1999,7(4):331-352.

[7]　KENNEDY J,EBERHART R. Particle swarm optimitzaion[C]//Proc of IEEE International Conference on Neural Networks. Washington DC:IEEE Computer Society,1995:1942-1948.

[8]　SHI Yu-hui,EBERHART R C. Parameter selection in particle swarm optimization[C]//Proc of the 7th International Conference on Evolutionary Programming. London:Springer-Verlag,1998:591-600.

[9]　XIAO Li-zhou,SHAO Zhi-qing,LIU Gang. K-means algorithm based on particle swarm optimization algorithm for anomaly intrusion detection[C]//Proc of the 6th World Congress on Intelligent Control and Automation. 2006:5854-5858.

[10]　肖立中,邵志清,马汉华,等. 网络入侵检测中的自动决定聚类数算法[J]. 软件学报,2008,19(8):2140-2148.