Transactions on Computer Science and Technology
March 2015, Volume 4, Issue 1, PP.12-19

Factor Analysis Based on RoboCup Midfielders' Positioning

Zekai Cheng, Liang Nie, Feng Qin

School of Computer Science and Technology, Anhui University of Technology, Maanshan 243032, China

#Email: nlh200717@sina.com, chengzk@ahut.edu.cn

Abstract

In order to study the influence of the midfielders' positioning for the matches' result, a method based on factor analysis was put forward to evaluate the midfielders' positioning. This method constructs mathematics model through parsing the log files of the games ,and calculate comprehensive evaluation scores of the midfielders' positioning for each team. The experimental result indicates that the team's rankings which were inferred using this method are coincide with the result rankings of the RoboCup simulation 2D games in recent years, and through drawing the scatter diagram of the relevance between comprehensive evaluation scores and result of the match, it could find that the team which has a higher comprehensive evaluation score would have a triumph with a great probability. After integrating theory with practice, it indicates that the key of the match is midfield flanks' positioning of midfielders in their own half for midfielder's active area.

Keywords: Midfielders' Positioning; Factor Analysis; Log Files; Comprehensive Evaluation Score; Scatter Diagram; Variance Contribution

基于因子分析的 RoboCup 中场球员跑位研究[1]

程泽凯，聂亮，秦锋

安徽工业大学 计算机科学与技术学院，安徽 马鞍山 243032

摘 要：为了深入研究中场球员的跑位对比赛胜负的影响，提出了基于因子分析的评价方法。该方法通过解析比赛日志文件，建立因子分析模型，并计算双方中场球员跑位的综合评价得分。实验结果表明，该方法所推测出的球队排名与近几年 RoboCup 仿真比赛的排名大致吻合，并且通过对综合评价得分与比赛胜负得分差绘制散点图，可以发现，综合评价得分较大的球队获胜的概率也较大。理论结合实际后得出，对中场球员活动的区域来说，本方中场侧翼位置的跑位对比赛胜负的影响最大。

关键词：中场球员跑位；因子分析；日志文件；综合评价得分；散点图；方差贡献率

引言

RoboCup 仿真 2D 是一个多智能体协作与对抗的平台，它提供了一个完全分布式控制、实时异步的多智能体环境[1]。比赛采用 Client/Server 方式，在一个标准的计算机环境内进行。比赛中，Client 与 Server 之间通过 UDP/IP 协议进行信息通信[2]。参赛队伍编制程序，着重解决实时、动态、复杂的环境下多智能的协作问题，为人类研究多智能体协作提供了一个不错的平台[3]。仿真组比赛中，中场球员一般由 1 名防守中场、1 名左进攻中场和 1 名右进攻中场组成，主要任务是衔接球队的进攻和防守。中场球员一般处在球场的中间、前锋队员的身后，但在必要情况下，它需要跑到球场上的任何一个角落[4]。一支球队的中场在引领球队进攻和协同队友防守上起着至关重要的作用。如何使中场球员跑位更合理，这个问题在仿真组比赛中已经越来

[1] *基金项目：受国家自然科学基金项目 (61300059) 和安徽高校省级自然科学研究基金项目 (KJ2014ZD05) 支助。

越重要了。

因子分析 (Factor Analysis) 是一种从变量群中提取共性因子的统计技术。在众多变量中挖掘出隐藏的具有代表性的因子，从而将变量降至一个更易掌握的水平，这种方法已被越来越多的研究人员所关注。 因子分析在实际生活中用途十分广泛，诸如在金融、旅游等领域上已有较为成功的应用[5-6]。

1 因子分析的概念

因子分析 (Factor Analysis) 最早由英国心理学家 Charles Spearman 于 1904 年提出，最终成为现代统计学的一个重要分支，基本思想是把一些信息重叠、具有错综复杂关系的变量归纳为少数几个不相关的综合因子[7]。文献[8]提出用因子分析法评价了 2002 年江苏省众多城市的经济发展水平，并提出了几点推动江苏省经济发展的建议；文献[9]运用因子分析法评价了我国中小企业板块中 22 家上市公司的综合业绩，在其 2003 年十三个主要财务指标中抽取出四个公因子，然后根据因子得分情况对其进行综合业绩排序；文献[10]针对传统科技创新效率评价的不足之处，提出一种基于因子分析和数据包络分析的两步骤高校科技创新效率评价模型；文献[11]运用因子分析法对福建南平市区域经济发展不平衡进行研究，得出几条有利于南平市区域经济协调发展的建议；文献[12]对沪市 604 家上市公司 2001 年度财务报表中十个主要的财务指标进行了因子分析，然后根据每个股票的因子得分情况做出了综合评价；文献[13]运用因子分析的方法分析了中超联赛 2008 年 210 场比赛的统计数据与比赛结果之间的关系，找出了硬性比赛成绩的五个公因子，为中超联赛提供了一定的科学评价依据。

本文将因子分析应用于 RoboCup 仿真 2D 中场球员跑位上，期望得到众多球队各场次比赛的中场球员的综合评价得分，然后找出综合评价得分与比赛胜负之间的关系，并进一步发现对比赛胜负影响最大的跑位区域。

2 因子分析的原理

2.1 因子分析模型

第一步，对实验样本的变量定义矩阵如式 (1) 所示：

$$X=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \tag{1}$$

其中，n 为实验样本的数目，p 为待分析的变量个数，矩阵的一行表示一组实验变量，n 行即表示 n 组实验数据。

第二步，对上述数据进行标准化处理，使标准化后的变量均值为 0，方差为 1。为了表述简单，不妨设标准化后的矩阵仍记作 X。

所以，相关系数矩阵如式 (2) 所示：

$$A=X'X \tag{2}$$

假设相关系数矩阵 A 的 p 个非负特征值分别为 λ_1，λ_2，…，λ_p，而对应于这些特征值的正交特征向量矩阵如式 (3) 所示：

$$V=\begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1p} \\ v_{21} & v_{22} & \cdots & v_{2p} \\ \vdots & \vdots & & \vdots \\ v_{p1} & v_{p2} & \cdots & v_{pp} \end{bmatrix} \tag{3}$$

第三步，先找出其因子载荷矩阵 $B=(b_{lk})_{p*m}$，其中 b_{lk} 为因子载荷，求出此矩阵的方法有很多，本文采用了主成分法，如式 (4) 所示：

$$b_{lk} = v_{lk} \cdot \sqrt{\lambda_k} \tag{4}$$

b_{lk} 表示第 l 个变量在第 k 个公因子上的载荷，反映了第 l 个变量在第 k 个公因子上的相对重要程度。

然后，令公因子矩阵为 F，则有

$$F = VX' \tag{5}$$

上式中 X' 表示 X 的转置矩阵。

不难得出

$$FF' = diag(\lambda_1, \lambda_2, \cdots, \lambda_p) \tag{6}$$

式(6)中 diag (λ1, λ2,…, λp) 表示主对角元素分别是λ1，λ2，…，λp 的对角阵，并且 Fi=Vi*X’，i=1,2,3,...p，即 Fi 为第 i 组实验样品的公因子得分。现选择 m (m<p) 个公因子，由变量之间的相关矩阵可抽取出第一公因子 F1，让其在所有变量的公因子方差中所占的方差比重最大，此后就不再考虑 F1 的影响；再从余下的相关矩阵中选出与 F1 不相关的公因子 F_2，按此类推，直到所有变量的公因子方差被分解完为止。

最后，建立模型如式 (7) 所示：

$$\begin{cases} X_1 = b_{11}F_1 + b_{12}F_2 + \cdots + b_{1m}F_m + \varepsilon_1 \\ X_2 = b_{21}F_1 + b_{22}F_2 + \cdots + b_{2m}F_m + \varepsilon_2 \\ \vdots \\ X_{\mathrm{p}} = b_{p1}F_1 + b_{p2}F_2 + \cdots + b_{pm}F_m + \varepsilon_p \end{cases} \tag{7}$$

残差 ε_1，ε_2，…，ε_p 为特殊因子，彼此相互独立，且服从正态分布 $N(0,\sigma*\sigma)$，代表着公因子以外的影响因素，实际分析时忽略不计。

2.2 KMO-Bartlett 与变量共同度

KMO (Kaiser-Meyer-Olkin) 检验统计量是用于比较变量间简单相关系数与偏相关系数的指标。KMO 统计量取值在 0 和 1 之间。当所有变量间的简单相关系数平方和远远大于偏相关系数平方和时，KMO 值越接近于 1，越适合做因子分析；反之，不适合。

Bartlett 球度检验用于检验相关阵是否为单位阵，显著性概率小于 0.05 时，才适合做因子分析。

变量共同度是指对某一个原始变量来说，其在所有因子上的载荷的平方和。它反映了所有公因子对原始变量方差的解释程度。如果大部分变量的变量共同度都高于 0.7，则因子分析效果较好。

2.3 方差贡献率与方差累积贡献率

公因子 F_i 的方差贡献率计算公式为 $\lambda_i/(\lambda_1+\lambda_2+\ldots+\lambda_m)$，即公因子 F_i 的方差在全部方差所占的比重。此值越大，则 F_i 对原始变量 X_1，X_2，…，X_p 的解释能力越强。

方差累积贡献率为前 K 个公因子的方差贡献率的总和，即 $(\lambda_1+\lambda_2+\ldots+\lambda_k)\ /(\lambda_1+\lambda_2+\ldots+\lambda_m)$。

2.4 因子旋转及命名

基于所得到的因子载荷矩阵难于给出合理的解释，为更好地阐明公因子的意义，减少解释的主观性，采用方差最大正交旋转的方法，即对因子载荷矩阵加以正交变换，变换的目的就是使因子载荷矩阵的所有元素的平方值向 0 和 1 两极分化，从而有利于对公因子的解释。观察所提取的公因子在因子旋转后与原始变量的相关性，可合理地对公因子进行命名。

2.5 因子得分及综合评价得分

因子分析的数学模型就是将原始变量表示为公因子的线性组合。因为提取的公因子能够反映原始变量之间的相关性，用公因子代表原始变量时，往往更有利于描述研究对象的特征，所以需要将公因子表示为原始变量的线性组合，如式(8)所示：

$$\begin{cases} F_1 = \rho_{11}X_1 + \rho_{12}X_2 + \cdots + \rho_{1p}X_p \\ F_2 = \rho_{21}X_1 + \rho_{22}X_2 + \cdots + \rho_{2p}X_p \\ \vdots \\ F_m = \rho_{m1}X_1 + \rho_{m2}X_2 + \cdots + \rho_{mp}X_p \end{cases} \tag{8}$$

式(8)即为因子得分函数，利用此函数可算出每个样本的因子的得分情况。估算因子得分的方法很多，本文采用回归法。

各个公因子分别从不同的角度上反映了整体水平，但单独某一公因子并不能全面地体现综合能力，故建立综合评价得分函数。本文采用了因子旋转法，设因子旋转后的方差贡献率为权值，与提取的各公因子建立线性方程，则最终的综合评价得分函数如式 (9) 所示：

$$F = \frac{\alpha_1 F_1 + \alpha_2 F_2 + \cdots + \alpha_m F_m}{\alpha_1 + \alpha_2 + \cdots + \alpha_m} \tag{9}$$

（α_i为旋转后第 i 个公因子的方差贡献率）。

不难发现，由式 (8) 和式 (9) 可推出综合评价得分 F 与原始变量 X_1，X_2，…，X_p之间线性关系，如式(10)所示：

$$F = t_1X_1 + t_2X_2 + \cdots + t_pX_p \tag{10}$$

由式 (10) 可知相关系数 t_1，t_2，…，t_p的大小，进而推断出最能影响综合评价得分 F 的原始变量。

3 实验

3.1 数据建模

3.1.1 日志文件获取

在 RoboCup 仿真 2D 比赛过程中，服务器会产生日志文件，分别为 RCG 文件和 RCL 文件。前者记录了任一周期内的球场状态，包括球的位置、球员的位置以及球员体能等相关信息；后者则主要记录了一些通信信息，如球员之间喊话、球员与教练之间通信等。本文的实验数据就是通过建立程序解析日志文件来获取的。

3.1.2 变量的选择

Agent 底层中，球场分为 8 个区域，分别为 A、B、C、D、E、F、G、H 等 8 个区域，通过建立程序输出双方中场球员在上述 8 个区域分布的数目，建立模型时将上述 8 个区域分布数目分别用原始变量 X_1，X_2，…，X_8表示。如图 1 所示。

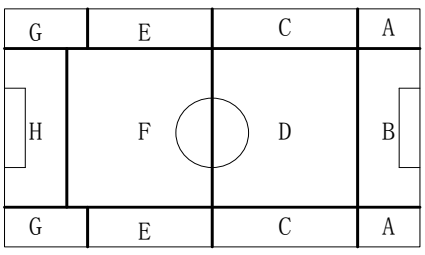


图 1 球场区域的划分

表 1 中场球员数据 (单位：周期数)

变量	1 队			2 队		
	DF	OFL	OFR	DF	OFL	OFR
X_1	0	127	64	0	36	0
X_2	0	163	123	0	6	42
X_3	197	377	1400	3	1028	1006
X_4	2497	3111	2077	2648	1712	1474
X_5	406	117	1079	0	533	529
X_6	2263	2108	1131	2848	2147	2382
X_7	328	0	118	0	372	112
X_8	308	0	6	499	175	453

现任意解析一场比赛中场球员的数据，本文所述的中场球员包括 1 名防守中场(DF)、1 名左进攻中场(OFL) 和 1 名右进攻中场 (OFR) ，如表 1 所示。

3.2 数据分析

现采集 60 场比赛日志文件作为实验数据，首先得到 KMO 与 Bartlett 检验结果如表 2 所示。

表 2 结果显示显著性概率值小于 0.005，则应拒绝各原始变量孤立的假设，即各变量之间具有较强的相关性。且 KMO 度量值为 0.721，大于 0.7，可采用因子分析模型来进行降维。

变量共同度提取情况如表 3 所示，结果显示所有的变量共同度都在 70%以上，则提取出来的公因子对各原始变量的解释情况良好。

各公因子的方差贡献率、累积贡献率以及因子旋转后情况如表 4 所示，根据特征值提取条件大于 1，抽取出三个公因子的特征值分别为 3.226、1.724、1.087，这三个公因子对原始变量解释情况累积贡献率是 75.453%。

表 2 KMO-Bartlett 检验

取样足够度的Kaiser-Meyer-Olkin度量		0.721
Bartlett球形度检验	近似卡方	763.309
	自由度	28
	显著性概率	0.000

表 3 变量共同度

变量	初始	提取
X_1	1.000	0.716
X_2	1.000	0.747
X_3	1.000	0.758
X_4	1.000	0.862
X_5	1.000	0.774
X_6	1.000	0.954
X_7	1.000	0.726
X_8	1.000	0.700

表 4 方差贡献率

成分	初始特征根			抽取平方和载入			旋转平方和载入		
	特征根	贡献率	累积贡献率	特征根	贡献率	累积贡献率	特征根	贡献率	累积贡献率
1	3.226	40.325	40.325	3.226	40.325	40.325	2.348	29.344	29.344
2	1.724	21.546	61.871	1.724	21.546	61.871	2.140	26.746	56.090
3	1.087	13.582	75.453	1.087	13.582	75.453	1.549	19.363	75.453
4	0.852	10.644	86.097						
5	0.423	5.287	91.385						
6	0.388	4.850	96.234						
7	0.301	3.766	100.00						
8	4.18E-6	5.23E-5	100.00						

3.3 因子旋转、命名及因子得分

为了便于解释公因子，采用最大方差法因子旋转后的因子载荷矩阵如表 5 所示。

由表 5 可知，F_1 在 X_3、X_4、X_6 上有较大载荷，此因子与球场中场区域相关性较高，可命名为中场因子；F_2 在 X_1、X_2、X_8 上有较大载荷，体现了与球场禁区相关性较高，可命名为禁区因子；F_3 在 X_5、X_7 上有较大相关性，反映了与球场后场边界区域关联性较大，可命名为边路防守因子，命名情况如表 6 所示：

表 5 旋转后因子载荷矩阵

变量	公因子		
	F_1	F_2	F_3
X_1	0.234	0.813	0.009
X_2	0.070	0.860	-0.048
X_3	0.865	0.100	-0.007
X_4	0.736	0.213	-0.524
X_5	0.087	0.032	0.816
X_6	-0.969	-0.079	-0.098
X_7	-0.228	-0.420	0.631
X_8	0.025	-0.707	0.447

表 6 公因子命名

公因子	所解释变量	公因子命名
F_1	X_3, X_4, X_6	中场因子
F_2	X_1, X_2, X_8	禁区因子
F_3	X_5, X_7	边路防守因子

采用回归法求出因子得分函数，则各公因子与各原始变量之间线性关系如式(11)所示：

$$\begin{cases} F_1 = 0.008X_1 - 0.078X_2 + 0.393X_3 \\ \quad + 0.291X_4 + 0.08X_5 - 0.449X_6 \\ \quad - 0.021X_7 + 0.124X_8 \\ F_2 = 0.445X_1 + 0.479X_2 - 0.052X_3 \\ \quad - 0.09X_4 + 0.196X_5 + 0.053X_6 \\ \quad - 0.07X_7 - 0.311X_8 \\ F_3 = 0.207X_1 + 0.165X_2 + 0.063X_3 \\ \quad - 0.311X_4 + 0.633X_5 - 0.143X_6 \\ \quad + 0.371X_7 + 0.178X_8 \end{cases} \tag{11}$$

3.4 综合评价得分及与比赛胜负关系

为整体地体现综合能力，需求出综合评价得分。现将旋转后的各公因子方差贡献率设为权值，建立综合评价得分 F 与公因子 F_1、F_2、F_3 的线性关系式，如式 (12) 所示：

$$F = 0.2943F_1 + 0.26746F_2 + 0.19363F_3 \tag{12}$$

通过公式 (12)，计算出原始 60 组数据的综合评价得分，现列出分值较大的前五位，如图 2 所示。

现从本实验数据 60 场比赛的日志文件中输出每场比赛后双方得分情况，记录下比赛得分差 Y，即比赛双方最终得分之差，也就是说，当 Y 大于零，比赛获胜；当 Y 等于零，比赛平局；当 Y 小于零，比赛败北。然后根据式 (12) 计算出每场比赛双方的综合评价得分 F，绘制综合评价得分 F 与得分差 Y 的散点图，如图 3 所示。

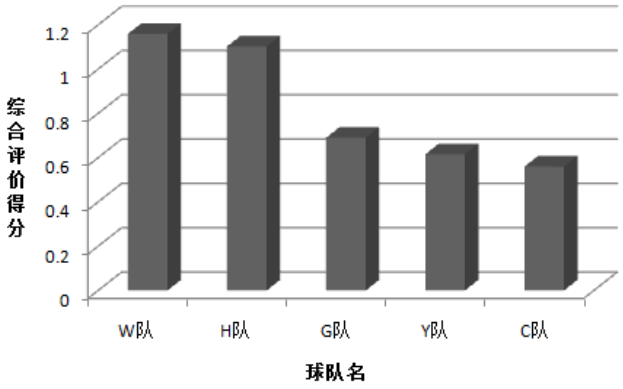


图 2 综合评价得分排名前五

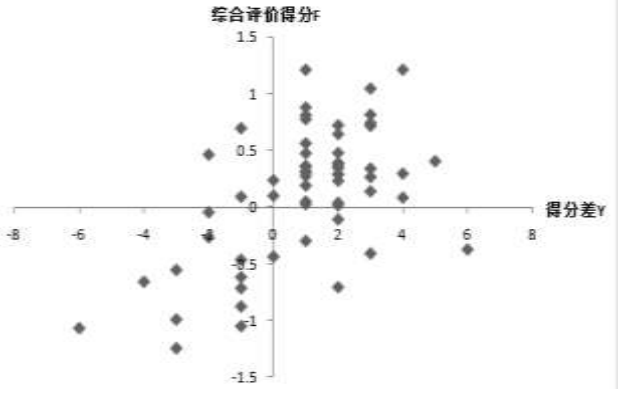


图 3 F-Y 散点图

最后，结合式 (11) 和式 (12) 可推出：

$$F = 0.1615X_1 + 0.1371X_2 + 0.1140X_3 + 0.0014X_4 + 0.1985X_5 - 0.1457X_6 + 0.0469X_7 - 0.0122X_8 \tag{13}$$

3.5 结论

由表 6 所示的中场因子、禁区因子、边路防守因子，在结合图 1 后可以更加形象直观地理解公因子的意义，三个公因子在整体上决定了一支球队进攻与防守的综合实力；由图 2 可知，此方法分析得出的综合评价

得分排名前 5 的都是近几年在 RoboCup 世界杯上排名顶尖的球队，验证了该模型的合理性；由图 3 可知，当一队的综合评价得分 F 高于另一队时，比赛获胜的概率达到 81.7%，即图 3 中分布在第一、三象限内的点占所有点的 81.7%；由式 (13) 可知，影响综合评价得分 F 大小的最重要的因素是 X_5，即图 1 中的 E 区域，也就是本方中场的侧翼区域。理论研究表明，中场球员在靠近本方中场的侧翼位置时，既可以组织本队的进攻又可以及时回来防守，能够很好地处理比赛中进攻与防守之间的关系，符合实际比赛情况。

理论结合实际比赛后可得出结论，对中场球员活动的区域来说，本方中场侧翼位置的跑位对比赛胜负的影响最大。

4 结束语

本文将因子分析的方法应用到 RoboCup 仿真 2D 科研领域中，最终得出本方中场的侧翼位置关乎比赛胜负的结论，并且结合一些图表对模型的合理性进行了阐述。下一步工作就是将此方法应用到拥有 2 名防守中场、2 名进攻中场的阵型研究中，考虑 4 名中场球员同时上场的情形，尝试去分析起决定性作用的关键区域，以期望 YuShan 队球员在跑位上有更大提高，并在今后赛事中取得更加骄人的成绩。

本研究存在着两点不足：1、本文研究了 3 名中场球员的阵型，暂时未考虑 4 名中场球员的阵型，对结果可能有一定影响；2、球场区域的划分是基于底层的，球场状态划分不够详细，可能有待于进一步改进，以期得到更加精准的结论。尽管存在着这些不足，但本研究为 RoboCup 仿真 2D 跑位上提供了一个较新的研究思路——因子分析。

REFERENCES

[1] Peter,Stone.Veloso M.Multiagent Systems:A survey from a Machine Learning Perspective[J].Automous Robots,2000,8(3):345-383

[2] Guo ye-jun, Xiong yong, Wu tie-jun. RoboCup simulation's development and design[J].Computer engineering and application, 2003, 29: 146-149

[3] Yang Yu-pu, Li Xiao-meng, Xu Xiao-ling. A survey of technology of multi-agent cooperation[J]. Information and Control, 2001, 30(4): 337-342

[4] Pan ling-han, Chu wei, Cheng xian-yi. Based on the role of the RoboCup soccer strategy [J]. Computer engineering and application, 2004, 26: 66-69

[5] Guo lan, Zhang yong, Li zhi-juan. Based on factor analysis of listed tourism companies efficiency evaluation of DEA method [J]. Journal of management, 2008, 5(2): 258-264

[6] Zhu ruo-xu. Our country's commercial bank competitiveness research--case analysis based on factor analysis method [D].Cheng du, 2012: 39-42

[7] Su li-yun, Chen cai-xia. The basis of statistical analysis and case application tutorial of SPSS19 [M]. Beijing hope electronic press, 2012

[8] Liao wei-kun, Cai guo-liang, Tu wen-tao.Urban economic development evaluation based on factor analysis method [J]. Statistics and decision, 2005

[9] Ran lun, Li jin-lin. Factor analysis method in the small and medium-sized enterprise plate application in comprehensive performance evaluation of listed companies [J]. Mathematical statistics and management, 2005, 24(1): 75-79

[10] Jiang tong-tong, Wu xiu-guo. Based on factor analysis/data envelopment analysis (dea) efficiency of 985 engineering colleges and universities of science and technology innovation evaluation research [C].2013 3rd International Conference on Education and Education Management, 2013: 407-412

[11] Chen wen-cheng.Based on factor analysis of unbalanced regional economic development [J]. Mathematical statistics and management, 2010, 29(3): 490-501

[12] Wang xue-min.Application of factor analysis in the stock evaluation [J]. Mathematical statistics and management, 2004,23(3):6-10

[13] Quan tao, Sun xue-han. Scores of China football super league in 2008 with the statistic analysis of the factors [J].Mathematical

statistics and management, 2011, 30(2): 370-380

【作者简介】

[1] 程泽凯（1975-），男，汉，硕士，副教授，人工智能，数据挖掘，机器学习，本科就读于合肥工业大学，研究生就读于广西师范大学。Email: chengzk@ahut.edu.cn

[2] 聂亮（1986-），男，汉，硕士，人工智能，数据挖掘，本科就读于西华大学，研究生就读于安徽工业大学。
Email: nlh200717@sina.com

[3] 秦锋（1962-），男，汉，硕士，教授，机器学习，数据挖掘，本科、硕士就读于合肥工业大学。