不确定性关联分类与人才资源配置的研究

闵华清，田祥庆
（华南理工大学 计算机科学与工程学院，广东 广州 510009）

摘 要：采用关联规则分类的方法，根据个人所在的行业和岗位的不同，对管理胜任力相关数据进行分类。结合不确定性问题，用概率来表示胜任力的隶属度，使对管理胜任力素质的分类更加符合人们的思维习惯。并且利用新的规则启发知识，对建立的模型进行了精确度优化，使之对胜任力素质类型的预测更加有效。

关键词：数据挖掘；不确定性分析；关联分类；管理胜任力
中图分类号：TP301 文献标识码：A

The research about uncertain associative classification and the association of human resources

MIN Hua Qing，TIAN Xiang Qing
(School of Computer Science and Engineering, South China University of Technology, Guangdong 510009, China)

Abstract: The article uses the method of associative classification to build a scientific model of management competence for different industries and positions. For resolving the problem of uncertainty, it describes the degrees on which the sample belongs to a class with the membership. It makes the classification of competency be more accordant to people's thinking habits. All these above make the model more effective.

Key words: data mining; uncertainty associative classification; management competency

现代企业的发展对人才管理的要求越来越高，企业如何制定一套适合自身行业特色的人力资源战略，决定了企业能否吸引、留住人才，能否在竞争激烈的市场中保持企业的竞争优势。随着人力资源管理在企业中的地位日益重要，其能否在企业中发挥重要作用，很大程度上取决于人力资源管理人员的管理胜任力素质，即他们能否让企业员工工作在合适的岗位上。

1 胜任力的定义与评估

1.1 胜任力的定义

自 McClelland(1973)提出“胜任力”概念，中西方学者纷纷提出自己对胜任力(胜任特征)的理解。通过研究众多学者给胜任力所下的定义，可以发现，胜任力有 3 个特点：(1)与特定工作相关；(2)可以在特定工作中创造高绩效；(3)包含一些个人的特征，如特质(Traits)、动机(Motives)、自我概念(Self-image)、社会角色(Social-role)、态度(attitude)、价值观(Value)、知识(Knowledge)、技能(Skill)等。

本文采用 Spencer 等人(1994)对胜任力的定义，即胜任力是指特质、动机、自我概念、社会角色、态度、价值观、知识、技能等能够可靠测量并可以把高绩效员工与一般绩效员工区分开来的任何个体特征。其中，较容易通过培训、教育来发展的知识和技能是对任职者的基本要求，被称为基准性胜任力(Threshold Competency)；而在短期内较难改变和发展的特质、动机、自我概念、社会角色、态度、价值观等高绩效者在职位上获得成功所必须具备的条件，被统称为鉴别性胜任力(Differentiating Competency)[1]。

1.2 胜任力的评估

传统的胜任力评估主要以专家打分法确定胜任力素质指标，有别于此，本文的研究始于开放式问卷收集

与胜任力素质相关的条目，编制预试问卷，然后筛选掉重要度或区分度不高的条目，形成最终问卷。得到相关数据后，运用关联规则分类方法对胜任力的评价与评价可能性进行建模，获取精确度较高的胜任力评估预测模型。

2 关联分类及其算法

2.1 基于关联的分类方法

关联分类规则挖掘的第一步就是发现所有的频繁和准确的可能规则，它们是类别关联规则[3]。若一个规则项目包含 k 个项目，就称这一规则项目集为 k-ruleitems。算法利用与 Apriori 算法类似的循环过程，只是用规则项目集替代了其中的项。

CBA(Classification-Based Association)算法就是一种在关联分类规则挖掘中发掘类关联规则的算法[2]。它是在 Apriori 算法的基础上去发掘频繁集和分类规则的。

关联规则挖掘的第二步就是对所获得的 CAR 进行处理以便构造一个分类器。由于为了获得最准确的规则集而要对所有的规则子集进行检查，这样所要处理的规则数目极为庞大，因此必须采用启发知识[3]。根据启发规则，分类器对所选的规则按优先值从高到低排列。当进行分类时，使用优先值大且满足条件的规则进行分类。此外，分类器还应包含一个缺省规则(具有最低优先值)，当其他规则都不满足时，利用这一缺省规则对数据对象进行分类。

通常，关联分类方法要比 C4.5 等普通分类算法更加准确，且以上两个步骤都具有线性可扩展性。

2.2 利用关联分类解决分类的不确定性问题

计算机要模拟人的思维和判断过程，就必须将人的语言中所具有的多义和不确定信息定量地表示出来，即不确定性问题。这种方式更加自然，更加接近人的表达方式。目前利用贝叶斯网络、模糊神经网络都能够解决不确定性分类的问题。

关联规则挖掘中将规则信任度表示为：$c(A \geq B)=P(B/A)=s(A \cup B)/s(A)$，其意义就是在 A 发生的前提下出现 B 的概率。如果把 A 看成条件，B 看成一个类，则可以表达为：在具备条件 A 的情况下，样本属于 B 类的概率。本文以此利用关联分析来解决不确定性分类的问题。

3 数据样本的预处理

3.1 预测问卷的因子分析

因子分析是从众多的原始变量中构造出少数几个具有代表意义的因子变量，这里有一个潜在的要求，即原有变量之间要具有比较强的相关性，否则无法从中综合出能反映某些变量共同特性的少数公共因子变量来[3]。因此，在因子分析时，需要对原有变量作相关分析。本文用 KMO 和球形 Bartlett 检验，对变量进行相关分析。

(1) KMO(Kaiser-Meyer-Olkin)检验

KMO 统计量用于比较变量间简单相关和偏相关系数，计算公式如下：

$$KMO=\frac{\sum\sum_{i\neq j} r_{ij}^2}{\sum\sum_{i\neq j} r_{ij}^2+\sum\sum_{i\neq j} p_{ij}^2}$$

其中，r_{ij}^2 是变量 i 和变量 j 之间的简单相关系数[4]，p_{ij}^2 是变量 i 和变量 j 之间的偏相关系数。已知 $0 \leqslant KMO \leqslant 1$，如 KMO 的值越接近于 1，则所有变量之间的简单相关系数平方和远大于偏相关系数平方和，因此越适合于作因子分析；反之，则不适于作因子分析。

Kaiser 给出了一个 KMO 的标准[5]，可表示为：

0.9<KMO：非常适合；

0.8<KMO ≤ 0.9：适合；

0.7<KMO ≤ 0.8：一般；

0.6<KMO ≤ 0.7：不太适合；

KMO ≤ 0.6：不适合。

(2)巴特利特球形检验（Bartlett Test of Sphericity）

巴特利特球形检验是基于变量的相关系数矩阵的检验方法[5]。它的零假设为相关矩阵是一个单位阵，即相关系数矩阵对角线上的所有元素都为 1，所有非对角线上的元素都为零。巴特利特球形检验的统计量是根据相关系数矩阵的行列式得到的。如果该值较大，且其对应的相伴概率值小于给定的显著性水平，就拒绝零假设，认为相关系数矩阵不可能是单位阵，即原始变量之间存在相关性，适合于作因子分析；相反，如果该统计量值比较小，且对应的相伴概率大于显著性水平，则不能拒绝原假设，此时不宜作因子分析。

3.2 公共因子的提取

本文以旺旺集团、广州百事可乐集团等八大现代企业的管理人员为研究对象，从发放开放式问卷出发，收集可能与胜任力特征相关的条目形成预试问卷，应用因子分析和方差分析方法对试卷进行检验和优化，筛选掉荷载低或区分度低的问题，最终生成包含 40 个问题的问卷。

用主成分法对最终问卷中 40 个问题的数据进行因子分析。首先应判断数据是否适合进行因子分析，此处仍然采用 KMO 和球度 Bartlett 检验，检验结果如表 1 所示。

表1 最终数据的KMO和球形Bartlett检验

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		0.875
Bartlett's Te 胜任特征 of Sphericity	Approx. Chi-Square	3 093.681
	df	990
	Sig.	0.000

由表 1 数据可知，KMO 检验值为 0.875，根据 Kaiser 给出的标准 0.8<KMO<0.9，表明数据适合于进行因子分析。Bartlett 球度检验的相伴概率为 0.000，小于显著性水平 0.05，故拒绝 Bartlett 球度检验的零假设，即认为适

合因子分析。

主成分分析研究如何通过原来变量的少数几个线性组合来解释随机向量的方差－协方差结构[3]。其作用为：(1)简化数据；(2)揭示变量间的关系。所谓主成分是指原来变量的线性组合，它们互不相关，且方差达到最大。采用主成分法，设定提取特征值大于1的因子，共提取了7个因子，其中特征值最大为15.810，最小为1.198。

根据因子的特征，本文提出7个公共因子的对应解释。7个公共因子的解释如图1所示。

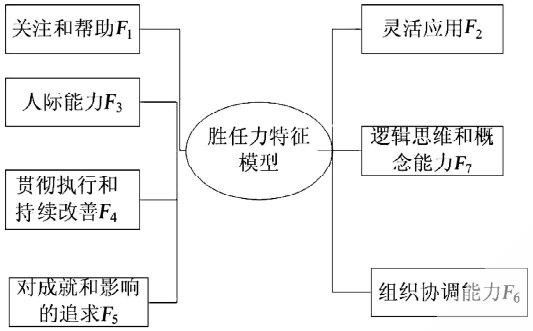

图1 管理人员胜任力特征

F_k取值为1～5，对应了5个重要性级别，1最低，5最高。

得到7个公共因子得分结果以后，接下来根据公共因子得分对所有样本数据进行聚类分析，以便确定如何对样本的评判等级进行分类。

通过基于EM（基于期望最大化）算法的聚类分析，样本数据集聚成了三类。本文把评判数据样本的优秀等级分为三级：优秀，良好，普通。

管理胜任力素质的评价预测模型，是多因素、多指标综合评价。在某一工作岗位上非常重要的知识和技能，在另外一个工作岗位上可能会成为制约其发展的阻碍因素。在一个组织中不同职务和不同管理层级所要求员工具备的胜任力内容和水平也是不同的。因此，需要建立能适用科学可行的管理胜任力素质评价体系，使企业做到人－岗匹配，发挥员工的最大能力。

本文把样本的多级管理胜任力水平与样本所在的岗位联系起来，从而解决了长期以来针对管理胜任力的研究没有结合具体岗位的问题。

4 建立基于关联分类的管理胜任力模型

首先，对2/3的样本数据建立管理胜任力预测模型。把F_1～F_7七个属性作为规则的左边，并利用岗位和胜任力水平两属性的值共同决定一个类别，设置最小支持度阈值和最小信任度阈值分别为0.3、0.6，对样本数据进行关联分类。

得到频繁集后，进而得到分类规则。根据启发知识对分类规则排序并建立胜任力模型，表2是分类规则的基本形式。

表2 胜任力模型中的部分规则

规则序号 i	规则	岗位 P	隶属度 C
1	F_7=4 ⇒Y=1	A	0.83
2	F_2=5, F_4=5 ⇒Y=2	A	1
3	F_1=3, F_2=3, F_3=3, F_6=4 ⇒Y=0	B	0.87
⋮			

4.1 检验模型精确度

本文用于建立模型的训练样本是总样本的2/3，为了检验管理胜任力模型的预测精度，需要使用剩下的1/3的样本作为测试数据集，对已经建立的胜任力模型的精确度进行评估。预测精度的检验公式为：

$$error=\frac{\sum_{i=1}^{n}(O_i-y_i)^2}{n} \quad (1)$$

其中O_i为输出值（预测值），y_i为真实值。

经过与测试样本的对比，本文得到的模型对测试数据集检验的精确度为89.341%，预测成功率较高。部分检验结果如表3所示。

表3 部分检验样本的输出结果与实际结果比较

样本序号 i	因子得分 F_1	F_2	F_3	F_4	F_5	F_6	F_7	岗位P	实际评价 y_i	计算评价 O_i
1	1	3	5	3	3	5	5	A	0	0
2	5	2	2	4	3	2	5	A	1	0
3	3	3	2	5	4	3	4	B	0	1
4	3	3	2	1	4	2	5	B	1	1
⋮										

4.2 模型的优化

普通的关联分类算法(CBA)在建模过程中采用的是一种基本的启发知识，如表4所示。这种启发知识主要考虑支持度和信任度的不同来对规则进行排序，然而当两条规则的支持度和信任度都相同时，启发知识规定产生时间早的规则拥有优先权。显然，越早产生的规则所含的属性越少，这说明基本的启发知识中含有这样一条隐含规则：当信任度和支持度相同时，规则左边所含属性少的规则的优先权高。

表4 基本规则启发知识

有两条规则r1、r2，r1优先于r2，如果：
①r1的信任度大于r2的信任度
②两规则的信任度相同，但r1的支持度比r2更大
③两规则的支持度和信任度相同，但在数据集中r1所指的类别比r2的类别出现得更频繁
④以上都相同，但r1产生得比r2早

（下转第64页）

（上接第60页）

然而，当遇到大数据集时，这种方法并不是非常有效。例如，在大数据集时，关联分类方法可能产生上万条分类规则，其中会有几千条具有相同的支持度和置信度。根据以上的启发知识，只能随机选择这些规则的优先权，而对于那些拥有优先权，它们有可能并不是最优规则，所以会影响模型的准确率。

在这里，根据以上的分析，本文提出了一项新的启发知识，使得规则的优先权确定更加完善合理，如表5所示。在两条规则的支持度、信任度相同时，赋予所指的类别在数据集中出现得更多的规则较高的优先权。当且仅当它们都相同时，分类器才选择产生得早的规则。

表5 改进的启发知识

有两条规则 r1、r2，r1 优先于 r2，如果：
①r1 的信任度大于 r2 的信任度
②两规则的信任度相同，但 r1 的支持度比 r2 更大
③r1 和 r2 的信任度和支持度都相同，但 r1 产生得比 r2 早

经过改进后，新的启发知识使规则与规则的关系更加明显，同时也保证了好的规则拥有更高的优先权，这就使得模型的准确率有可能进一步提高。

4.3 模型优化后准确率的对比

为了能说明以上提出的模型优化方法的有效性，本文同时也对 WEKA 3.5.5 所自带的几个数据集进行了关联规则分类建模与优化后精确度的对比，结果如表6所示。

表6 模型优化后准确率的对比

数据集	优化前的精度/%	优化后的精度/%
胜任力数据	89.34	90.86
soybean	67.76	70.97
weather	94.61	94.25
labor	81.30	82.04
Segment-test	88.86	88.17

通过以上对比可以看出，总的来说，数据集在经过优化后的模型精度都是有所上升的。这说明本文对启发规则的优化的确改进了建模的精度，从而能够提高模型的预测精度。

本文针对管理胜任力素质，以旺旺集团、广州百事等企业的管理人员为研究对象获取数据，考虑分类的不确定性问题，尝试采用关联规则分类来建立管理胜任力预测模型，把对样本胜任力的预测与岗位相联系，在预测中增加了隶属度的表示来帮助决策者做出决定，并使用了多级的评判标准，最后根据分析建模过程中规则的优先级排序提出了优化的规则启发知识，使规则的排序更加完善，进而使样本的分类准确度更高，提高了模型精确率和效率。

参考文献

[1] SPENCER L M, SPENCER S M. 才能评鉴法：建立卓越的绩效模式[M]. 魏梅金，译. 汕头：汕头大学出版社，2003.

[2] LIU B, MA Y. Integrating classification and association rule mining [C]. Proc of the 4th International Conference on Knowledge Discovery and Data Mining, New York, 1998.

[3] HAN Jia Wei. Data mining-concepts and techniques[M]. 北京：机械工业出版社，2006.

[4] TAN Pang Ning, MS V. Introduction to data mining [M]. 北京：人民邮电出版社，2006.

[5] 梁之舜，邓集贤. 概率论及数理统计(第二版)[M]. 北京：高等教育出版社，1988.

（收稿日期：2009-03-17）