文章编号：1001 - 4217 (2018) 01 - 0040 - 09

大数据下医保欺诈的有效识别模型

陈清凤，朱　宁，朱亩鑫
（桂林电子科技大学数学与计算科学学院，广西　桂林　541004）

摘　要　针对现在社会医保诈骗问题，提出了大数据下医保欺诈的有效识别模型. 首先运用 excel 对数据进行预处理，建立数据挖掘有效识别数据集；其次通过主成分分析构建欺诈识别的有效指标体系；再次由 K-Means 聚类得到可疑的医保欺诈行为的类别，并由判别分析中的交叉确认估计来确认可疑行为判断类别的准确性. 随后，由因子分析中的数据映射关系找到与欺骗行为有关的科室、医生、医嘱子类，并把欺诈行为归为医疗保险服务供应方的诈骗行为、医疗保险需求方的诈骗行为和医疗保险服务供应方与需求方合谋的诈骗行为这三大类；最后把模型用于由样本经验分布的反函数生成的大数据中，解决了统计分析中样本少而使统计分析出现误差这一问题.

关键词　数据挖掘有效识别数据集；主成分分析；K-Means 聚类；判别分析；因子分析；大数据
中图分类号　R195.1　　**文献标志码**　A

0　引　言

随着参保覆盖面和基金规模的迅速扩大，定点服务机构的大量增加，我国的医保信息系统也得到了广泛的应用，如何利用海量的医疗数据建立有效的医保欺诈预警模型，为医保中心实施监管的工作提供决策支持，是当前所要解决的首要任务.

对于医疗保险欺诈的理论分析和实证研究，国外学者主要从社会心理学、博弈论以及数据挖掘的角度进行研究. Arrow[1]根据信息不对称理论，首次对健康保险欺诈问题进行了探讨和研究. 随后 Pauly[2]，Schiller，Moreno[3]分别从管控道德风险和剔除受投保方操纵信号的方式反制欺诈. 在此基础上，Artis[4]，Chiappori[5]，Brocket[6]等人分别采用 Probit、AAG、Pridit、logit 等统计模型，对具体的欺诈行为进行识别. 但由于这些模型对数据有一定的要求，加上欺诈的复杂性，这使得传统的单一模型在实际的应用中受到很大的限制. 为此 Marisa S[7]，Sokol[8]，Lious[9]，等人把人工智能识别模型和统计回归模

收稿日期：2017 - 05 - 11
作者简介：朱　宁（1957—），男，湖南宁乡人，教授，研究方向：线性统计模型；
通讯作者：陈清凤（1992—），女，广西梧州人，E- mail：1572828324@qq.com.
基金项目：广西区大学生创新项目（201510595278）

型进行有效的组合，分别建立了基于 BP 神经网络模型、遗传算法、贝叶斯网络、糊集聚类算法、数据挖掘的欺诈识别模型，并用于特定的例子中，识别效果较好. 除此之外基于启发式和机器学习的电子欺诈识别技术也被广泛的应用于医疗保险欺诈识别.

国内学者对医疗保险欺诈问题主要是运用信息不对称和博弈论，围绕欺诈的类型、表现形式、欺诈的成因分析和反欺诈措施等三个方面进行理论研究，关于社会医疗保险欺诈的识别和度量的研究还较少[10]. 对于社会医疗保险欺诈的识别，较早应用的是徐远纯[11]根据粗糙集理论的特征属性提出的欺诈风险识别方法，随后陈辉金、韩元杰[12]基于数据挖掘和信息融合技术建立孤立点集来挖掘可疑数据；梁子君[13]利用贝叶斯网络建立了识别、评估和管控欺诈风险的概念模型；叶明华[14]把统计回归和神经网络进行有效融合，建立了基于江、浙、沪机动车保险索赔数据构建了欺诈识别的 BP 神经网络模型. 杨超[15]在叶明华的研究的基础上，运用嵌入 logistic 回归分析的 BP 神经网络模型研究识别被保险人道德风险引致的欺诈. 总的来说，如何从海量的复杂隐秘的医疗保险数据中识别出具有欺诈行为的信息还没有得到具体的解决，为此把统计方法与大数据相结合的识别模型的研究是有意义的.

本文在大数据背景对医疗保险欺诈这一课题进行研究，首先对给定的医疗数据进行预处理，通过主成分分析构建欺诈识别的有效指标体系；其次由 K-Means 聚类得到可疑的医保欺诈行为的类别；再次，利用因子分析方法，根据特征因子分析诈骗类的特征确定其诈骗方式；最后把模型用于由样本经验分布的反函数生成的大数据中. 具体流程如图 1.

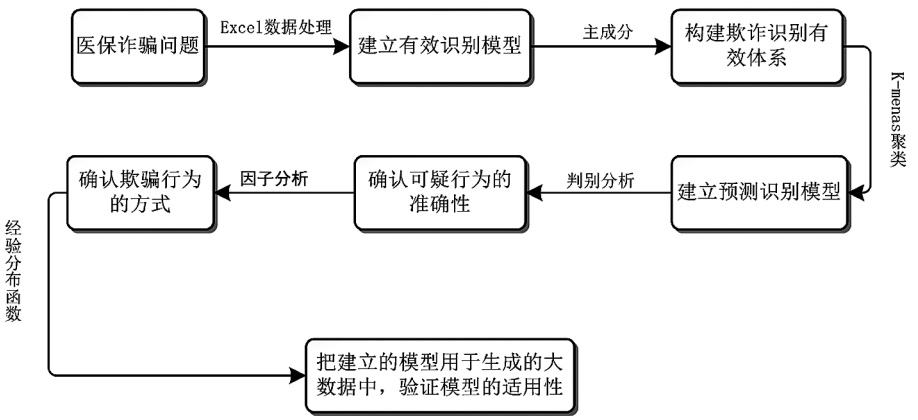

图 1　医保欺诈模型流程图

1　数据预处理

本文以 2015 年“深圳杯”数学建模夏令营 A 题：医保欺诈行 C 医保数据为研究数据，共 289 001 条记录. 为了构造医保诈骗有效识别的数据集，本文利用大数据挖掘技术对参保人信息进行数据预处理，利用 Excel 软件中的 vlookup 函数对原始数据进行定性筛选，去掉不必要的数据.

数据清洗　基于课题的研究意义和方向，结合给出的 6 个表格的医疗数据，进行数

据清洗. 首先利用 Excel 中的透视表剔除缺失值个数大于列数 20%的行，并删除对于本次数据挖掘没有意义数据，保留相关数据列，观察得到的数据集中没有重复记录，省去了对重复记录的处理. 其次是对于缺失的必要数据，例如刷卡次数缺失的数据，其占总样本的 25.5%，采用数据归约中多项式回归的方法填补空缺，其他指标也如此.

数据的转换 清洗得到的数据转换为便于处理的形式，日期采用“年 - 月 - 日”格式，医嘱 ID 号精简成数字型.

生成有效识别数据集 从给定的数据中提取出用于描述样本的指标，从而解释医疗数据的标签和分类的来由. 根据参保人信息数据集和医保交易记录数据集中的属性对数据进行适当处理，进而派生出所需要的识别指标. 对医保交易记录数据集中的重要属性进行不重复计数处理，派生出总费用、刷卡总次数、一次性消费最高额、平均消费金额以及医嘱子类、开嘱医 ID、下医嘱科室、核算分类、执行科室和病人科室的不重复计数这 10 个指标.

本文选取了具有代表性的属性，并根据参保人信息数据集中的 PAPMI_PAPER_DR (身份证 ID) 和医保交易记录数据集中的 WorkLoad_PAPMI _DR (病人病历 ID) 将两数据集进行自然连接，从而生成目标数据集，即医保诈骗有效识别数据集，见表 1. 此时数据集已经从初始的 289 001 条原始记录整合成 58 014 条目标记录.

表 1 参保人信息和医保交易记录交叉数据集

指标	数据类型	指标	数据类型
病人 ID	主键	执行科室非重复计数 (x6)	离散性
刷卡次数 (x1)	离散值	病人科室非重复计数 (x7)	离散值
一次性消费最高金额 (x2)	连续值	医嘱子类非重复计数 (x8)	离散值
总费用 (x3)	连续值	下医嘱科室非重复计数 (x9)	离散值
平均消费金额 (x4)	连续值	核算分类非重复计数 (x10)	离散值
开嘱医生 ID 非重复计数 (x5)	离散值		

数据标准化 根据 $z_{ij}=(x_{ij}-x)/s_i$ 对提取出的数据集进行标准化处理，其中 z_{ij} 为标准化后的变量值，x_{ij}为实际变量值.

2 欺诈识别的有效指标体系的构建

由于得到的识别指标过多，如果对所有的指标进行分析可能会存在信息重叠，对部分个体的欺诈识别因子进行主成分分析，提取综合指标来消除指标间相关性. 首先，对指标进行了相关分析，运用 SAS 统计软件导入包含 58 014 个医保人信息的数据集，计算出各指标之间的 Pearson 相关系数，结果如表 2.

由表 2 可以看出，部分指标之间存在着严重的相关性，如病人科室不重复计数和下医嘱科室不重复计数间的相关系数高达 0.999，接近于 1；一次性消费最高数额和总费用的相关系数也达到了 0.758，说明原指标变量间有一定的相关性. 此时如果直接对原来的指标进行分析就会造成信息的重复使用而使得结果不准确.

表 2　指标之间 Pearson 相关系数

相关矩阵										
	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
x1	1.000 0	0.150 8	0.383 8	- 0.024 2	0.671 6	0.133 2	0.451 0	0.464 8	0.451 0	0.192 0
x2	0.150 8	1.000 0	**0.853 8**	**0.943 8**	0.097 7	0.117 4	0.101 8	0.084 2	0.102 2	0.095 2
x3	0.383 8	**0.853 8**	1.000 0	**0.771 0**	0.263 0	0.139 2	0.219 5	0.155 8	0.219 7	0.124 8
x4	- 0.024 2	**0.943 8**	**0.771 0**	1.000 0	- 0.016 1	0.076 1	- 0.001 8	- 0.002 3	- 0.001 7	0.058 7
x5	**0.671 6**	0.097 7	0.263 0	- 0.016 1	1.000 0	0.110 9	**0.674 8**	0.476 0	**0.675 2**	0.173 6
x6	0.133 2	0.117 4	0.139 2	0.076 1	0.110 9	1.000 0	0.176 5	0.227 7	0.177 2	0.501 9
x7	0.451 0	0.101 8	0.219 6	- 0.001 8	**0.674 8**	0.176 5	1.000 0	0.326 8	0.999 2	0.193 1
x8	0.464 8	0.084 2	0.155 8	- 0.002 3	0.476 0	0.227 7	0.326 8	1.000 0	0.326 5	0.375 5
x9	0.451 0	0.102 2	0.219 7	- 0.001 7	**0.675 2**	0.177 2	**0.999 2**	0.326 5	1.000 0	0.192 6
x10	0.192 0	0.095 2	0.124 8	0.058 7	0.173 6	0.501 9	0.193 1	0.375 5	0.192 6	1.000 0

随后，通过主成分分析来消除指标之间的相关性，提取出欺诈识别模型的综合指标，结果如表 3.

表 3　主成分分析结果

	$\hat{t}_1$	$\hat{t}_2$	$\hat{t}_3$	$\hat{t}_4$	$\hat{t}_5$
特征值	3.731 230 06	2.496 450 93	1.402 224 97	0.912 311 35	0.559 715 42
方差	1.234 779 13	1.094 225 96	0.489 913 63	0.352 595 93	0.109 611 15
贡献率	0.373 1	0.249 6	0.140 2	0.091 2	0.056 0
累计贡献率	0.373 1	0.622 8	0.763 0	0.854 2	0.910 2

由表 3 的数据可以看出，前五个主成分的累计贡献率已达到 91.02%，可以认为它们能较好地概括原始指标的大部分信息，即用前五个主成分作为欺诈识别指标.

3　欺诈识别的统计模型

3.1　随机样本的类平均聚类

为了更好的识别出医保数据中的欺诈行为，根据收集到的六万人的消费交易记录，利用类平均聚类对其进行聚类获取先验信息，将主成分分析得到的前五个主成分作为综合指标，通过无放回简单随机抽样方法抽取 5 组样本 (每一组容量 5 000) 进行聚类，下面对其中一组建立医保诈骗识别模型. 聚类的信息如表 4.

从 R^2 统计量来看，当 NCL (聚类数) ＞5 时下降较缓慢，且 NCL＝5 时下降较大，半偏相关统计量达到最大；从伪 F 统计量来看，NCL＝5 时，取得极大值，且 NCL＝5 时，$PST2$ (伪 F 统计量) 取得极大值. 由此可知，随机样本分成 5 类较合适.

表4 随机样本类平均聚类结果

聚类数	频数	半偏R方	R方	近似期望R方	立方聚类条件	伪F统计量	伪t方	Norm RMS distance
10	25	0.001 7	0.873	0.935	-23	755	23.9	0.613 3
9	14	0.000 7	0.872	0.927	-19	845	8.8	0.624 4
8	20	0.003 4	0.869	0.918	-16	939	61.1	0.647 3
7	39	0.008 5	0.860	0.906	-14	1 019	63.5	0.774 2
6	59	0.012 3	0.848	0.890	-12	1 109	39.2	0.863 7
5	3	0.002 7	**0.845**	0.868	-5.9	**1 359**		1.412 6
4	992	**0.226 7**	**0.619**	0.835	-33	539	1 477	1.536 6
3	994	0.028 2	0.590	0.779	-26	719	73.5	2.692 1
2	997	0.103 2	0.487	0.655	-15	948	251	4.229 7
1	1 000	0.487 2	0.000	0.000	0.000		948	9.034 1

重复以上步骤，再对随机抽取的其他4组样本进行K-Means聚类分析，过程与上面样本类似. 通过对利用无放回简单随机抽取方法抽取到的5组样本量为5 000的样本依次进行主成分聚类分析，其中有3组样本认为聚成5类最合适，其余2组比较分散，将这些信息作为先验信息，根据最大似然函数的原理认为全部样本聚成5类是合适的. 聚类结果如表5.

表4 K–Means动态聚类

聚类	频数	均方根标准差	从种子到观测值的最大距离	最近的聚类	聚类质心间的距离
1	263	235.6	964.1	3	1 452.8
2	4	383.3	1 072.3	4	2 153.7
3	7 612	148.1	969.9	5	613.4
4	24	185.4	657.0	1	1 617.2
5	50 111	54.297 6	566.0	3	613.4

由表4看出第五类包含的样本最多，共有50 111条记录，其次是第三类，而第1、2、4类的个数较少. 由于医疗保险诈骗事件属于小概率事件，且诈骗的形式有多种，比如拿着别人的医保卡配药、在不同的医院和医生处重复配药等，可以表现为单张处方药费特别高、一张卡在一定时间内反复多次拿药等. 由表4的数据可直观的认为第1、2、4类属于医保诈骗的可能性较大，因为它们组内均方根的标准差和从凝聚点到各类内观测值的最大距离都比较大，说明这些类之间有一定的差异，存在着问题，需要谨慎对待.

3.2 模型检验—判别分析

为了验证K-Means动态聚类结果的合理性，利用判别分析中的交叉确认估计来判断聚类准确性，结果如表5和表6.

表 5　各组错判具体情况

分入“group”的观测数和百分比						
组别	1	2	3	4	5	合计
1	254	0	0	0	0	263
	96.58	0.00	0.00	3.42	0.00	100.00
2	0	**2**	0	**2**	0	4
	0.00	50.00	0.00	50.00	0.00	100.00
3	387	0	7 224	0	**1**	7 612
	5.08	0.00	94.90	0.00	0.01	100.00
4	**2**	0	0	**22**	0	24
	8.33	0.00	0.00	91.67	0.00	100.00
5	**4**	0	3 560	0	46 547	50 111
	0.01	0.00	7.10	0.00	92.89	100.00
合计	647	2	10 784	33	46 548	58 014
	1.12	0.00	18.59	0.06	80.24	100.00
先验	0.2	0.2	0.2	0.2	0.2	

表 6　错判概率

“groun”的出错估计						
	1	2	3	4	5	合计
比率	0.034 2	0.500 0	0.051 0	0.083 3	0.071 1	0.147 9
先验	0.200 0	0.200 0	0.200 0	0.200 0	0.200 0	

由表 5 和表 6 的数据可知，聚类时总体的错判概率为 0.147 9. 其中第 1 组中错判的样本量为 9 个，错判概率为 0.034 2，且这 9 个错判的样本都被错判到第 4 组；第 2 组中错判的样本量为 2，错判概率高达 0.500 0，且这 2 个错判的样本都被错判到第 4 组；第3 组中错判的样本量为 388，错判概率为 0.051 0，其中 387 个样本被错判到第 1 组，1 个样本被错判到第 5 组；第 4 组中错判的样本量为 2，错判概率为 0.083 3，且这 2 个错判的样本都被错判到 1 组；第 5 组中错判的样本量为 3 564，错判概率高达 0.071 1，其中 4 个样本被错判到第 1 组，3 560 个样本被错判到第 3 组.

结合 K-Means 聚类的结合和判别分析的结果可知，在 57 723 个非欺诈个体中有 391 个可能属于欺诈个体，错判概率为 0.677%；而初始判断为欺诈类别的 291 个样本中有 0 个被错判，此时错判概率为 0%. 由此可以初始确定的诈骗类别是合理的.

3.3　医保欺诈识别的特征模型—因子分析

利用因子分析找出潜在的对医疗数据中较为可疑的医疗数据的特征进行分析，通过公共因子来查找出 K-Means 聚类中的第 1，2，4 类可疑诈骗的基本特征，最终确定诈骗方式. 设特征值（Eigenvalues）、贡献率（Contribution rate）和累计贡献率分别用（Cumulative contribution rate）Eig、CR、CCR 表示，则进行因子分析后的统计量如表 7.

表7 因子分析统计量

指标	因子载荷				指标	因子载荷			
	f_1	f_3	f_3	f_4		f_1	f_2	f_3	f_4
x1	0.392	0.779	0.178	-0.378	x8	0.186	0.436	0.667	0.141
x2	-0.09	0.048	0.097	0.937	x9	0.396	0.182	0.173	0.098
x3	0.184	0.909	0.038	0.008	x10	-0.002	0.072	0.928	0.017
x4	-0.304	-0.412	-0.203	0.744	Eig	4.609	1.612	1.174	1.094
x5	0.768	0.340	0.037	0.232	CR	0.461	0.161	0.117	0.109
x6	0.460	-0.144	0.669	0.170	CCR	0.461	0.622	0.740	0.849
x7	0.936	0.182	0.173	0.098					

从表7可以看出，在以100%的累计方差贡献率确定的10个因子中，前4个因子特征值大于1，累计方差贡献率高达84.9%，故考虑提取4个公因子. 又从最大方差旋转的因子载荷矩阵可知，公因子f_1主要在病人科室非重复计数、开嘱医生ID非重复计数、执行科室非重复计数上具有较大的正载荷，故命名为科室分类因子；公共因子f_2主要在刷卡次数、费用有很大的正载荷，故命名为刷卡费用因子；公共因子f_3主要在执行科室非重复计数、医嘱子类非重复计数有较大的正载荷，故命名为医疗服务因子；公共因子f_4主要在一次性消费最高金额、平均消费金额有很大的正载荷，故命名为费用因子.

通过上述分析可发现此类有个共同特点就是一次性消费平均消费最高金额，病人科室非重复计数所占比率最高，存在故意串通医生开大处方行为，购大量药品等来套取统筹医保基金的嫌疑，属于医疗保险服务供方与需方合谋的诈骗行为.

以此类推可以得到第2、第4类的诈骗方式. 其中，第2类欺诈的方式可定义为贩卖药品诈骗，是指医保患者通过医保卡去不同的医保定点医院多次重复看病、取药，然后再将多取的药品贩卖，从而达到骗取医保基金的目的；第4类诈骗方式定义为分解收费诈骗，即定点医疗机构在为参保患者提供医疗服务过程中，人为地将一个完整的连续的医疗服务项目分成两个或两个以上的医疗服务项目，并按分割后的项目进行收费，从中获取差价进行医疗诈骗.

综上所述，可将欺诈行为分成三大类：

1. 医疗保险服务供应方的诈骗行为；

2. 医疗保险需求方的诈骗行为；

3. 医疗保险服务供应方与需求方合谋的诈骗行为.

结合各类的具体特征，又可以将各欺诈行为分别定义为分解收费诈骗、贩卖药品诈骗、提供虚假证明或伪造病历诈骗、冒名顶替诈骗.

3.4 大数据下的模型的优越性

为了验证模型的适用性，将识别模型应用于生成的海量数据中运行. 首先，把第一个指标的数据(刷卡次数(x1))由origin软件拟合出样本的分布函数为：

$$F(x)=\frac{1.004-0.929}{1+\left(\frac{x}{87.15}\right)^{1.746}}.$$

其次，产生符合该分布随机，通过分布 $F(x)$ 反函数求出随机数对应的样本 x 值，重复以上步骤便可得其他各指标的数据的样本的分布函数，最后把提出的识别欺诈模型带入求得的样本值中，再利用上述方法重新运行一遍，以便验证之前所用方法是否正确.

4　结论

研究结果表明：基于主成分 K-Means 聚类和因子分析的数据挖掘方法对医保欺诈行为能够进行较为准确的预警，与直接进行聚类相比，文中提出的模型运行速度较快、效率较高，并适用于大数据中的欺诈行为的识别. 在设计思路上从统计分析的角度出发，定量地研究了如何从大量数据中识别出少数的可疑的医保诈骗行为.

参考文献

[1] ARROW K J. Uncertainty and the welfare economics of medical care[J]. Uncertainty in Economics，1978，82(2)：141-149.

[2] PAULY M V. Taxation，health insurance，and market failure in the medical economy[J]. Journal of Economic Literature，1986，24(2)：629-675.

[3] SCHILLER J. The impact of insurance fraud detection systems[J]. Journal of Risk and Insurance，2006，73(3)：421-438.

[4] ARTÍS M，AYUSO M，GUILLÉN M. Detection of automobile insurance fraud with discrete choice models and misclassified claims[J]. Journal of Risk and Insurance，2002，69(3)：325-340.

[5] CHIAPPORI P A，SALANIE B. Testing for asymmetric information in insurance markets[J]. Journal of Political Economy，2000，108(1)：56-78.

[6] BROCKETT P L. Fraud classification using principal component analysis of RIDITs[J]. Journal of Risk and Insurance，2002，69(3)：341-371.

[7] VIVEROS M S，NEARHOS J P，ROTHMAN M J. Applying data mining techniques to a health insurance information system[C]//VLDB '96 Proceedings of the 22th International Conference on Very Large Data Bases. San Francisco：Morgan Kaufmann Publishers Inc. 1996：286-294.

[8] SOKOL L，GARCIA B，RODRIGUEZ J，et al. Using data mining to find fraud in HCFA health care claims[J]. Topics in Health Information Management，2001，22(1)：1-13.

[9] LIOU F M，TANG Y C，CHEN J Y. Detecting hospital fraud and claim abuse through diabetic outpatient services[J]. Health Care Management Science，2008，11(4)：353-358.

[10] 林源. 国内外医疗保险欺诈研究现状分析[J]. 保险研究，2010(12)：115-122.

[11] 徐远纯，柳炳祥，盛昭瀚. 一种基于粗集的欺诈风险分析方法[J]. 计算机应用，2004，24(1)：20-21.

[12] 陈辉金，韩元杰. 数据挖掘和信息融合在保险业欺诈识别中的应用[J]. 计算机与现代化，2005(9)：110-112.

[13] 梁子君. 保险公司操作风险管理——用贝叶斯网络评估和管理保险欺诈[D]. 上海：上海财经大学，2006.

[14] 叶明华. 基于 BP 神经网络的保险欺诈识别研究——以中国机动车保险索赔为例[J]. 保险研究，2011(3)：79-86.

[15] 杨超. 基于 BP 神经网络的健康保险欺诈识别研究[D]. 青岛：青岛大学，2014.

Effective Identification Model of Medical Insurance Fraud for Big Data

CHEN Qingfeng, ZHU Ning, ZHU Muxin
(School of Mathematics and Computing Science, Guilin University of Electronic Technology, Guilin 541004, Guangxi, China)

Abstract Aiming at the problem of social health insurance fraud, an effective identification model for Medicare fraud in big data is proposed. First, the excel is used to preprocess the data, establishing an effective identification data mining data set. Secondly, the effective index system of fraud identification by principal component analysis is constructed. The category of suspicious medical fraud is obtained from K-Means clustering, and the accuracy of the suspicious behavior judgment category by cross validation of discriminant analysis is confirmed. From the data mapping relationship of factor analysis is used to find the deceptive behavior of the departments, doctors, medical sub-class. Then, the fraud is classified as medical insurance service provider, demander, collusion between supply and demand three categories. Finally, the model for big data which generated by the inverse of the empirical distribution of the samples, is used to solve the problem of statistical analysis of the error for the sample less.

Keywords effective identification data set; principal component analysis; K-Means clustering; discriminant analysis; factor analysis; big data.

(上接第39页)

Automatic Reconstruction of Fragments

LIAO Minyu, XIE Ruicheng, YU Shengyu
(Department of Mathematics, College of Science, Shantou University, Shantou 515063, Guangdong, China)

Abstract In order to improve the efficiency of fragment reconstruction and protect the information security, a strategy for fragment reconstruction is proposed, which mainly consists of two characteristics: ink feature extraction and image matching. An automatic restoring algorithm which is based on the baseline of text, advanced genetic algorithm (GA) and optical character recognition (OCR), is proposed for English documents. Firstly, the algorithm classifies fragments according to the same criterion of the lower baseline of the peer letters. Therefore, the problem is transformed to Traveling Salesman Problem after the classification. Secondly, the improvement genetic algorithm and the optical character recognition technology are used to solve the problem. Thirdly, the algorithm arranges the lines to page by the location of the baseline and greedy algorithm as an assistant. In addition, based on the characteristics of Chinese characters, the algorithm can be modified as an automatic reconstruction algorithm for Chinese documents. In the end, MATLAB programs are developed according to the auto-matching algorithm of fragments. Experimental results demonstrate that the algorithm is efficient.

Keywords fragments; automatic reconstruction; traveling salesman problem (TSP); improved genetic algorithm (GA); optical character recognition technology (OCR)