1.2

在定义数据属性与基本评价规则基础上对用户答题数据进行基于谱聚类的分层过滤，标识出基本的群体簇；与传统的聚类算法相比，谱聚类具有能够在任意形状的样本空间上聚类，且收敛于全局最优解的优点。谱聚类算法可简要描述为：给定一个数据集*X={x1,x2,…,xn},xi*。根据数据集*X*建立加权图*G=(V,E)*。其中*V={vi,i=1,2,…,n}*是顶点的集合，*E={eij}*是连接顶点*（vi, vj）*的边。图*G*中每一个节点*vi*与数据集*X*中的*xi*相关。采用一个相似度准则构造图*G*的定点之间的相似度矩阵*W, W*。对于两个对象*xi*和*xj*,其相似度为*s（xi，xj）=exp（-|| xi-xj||2/2σ2）*。算法输入图*G*的相似度矩阵*W*和聚类个数*k*，输出*x1,x2,…,xn*的聚类结果。基于谱聚类的分层过滤算法根据考试答卷数据对象的特点在谱聚类基础上，增加了答卷数据对象的距离矩阵，得到适合答卷数据分析的谱聚类算法。假设研究的答卷数据对象*x*有*p*个测试点，表示为*（{x1,x2,…,xp）*，两个被试答卷数据对象*u*和*v*之间的距离*d(u,v)*可定义为:*d(u,v)*=。其中，|=1，或|=0，；*i=1，2，…,p*。根据答卷数据对象之间的距离，两个被试的答卷数据对象*u*和*v*之间的相似度*s(u,v)*可定义为：*S(u,v)=exp(-d(u,v)/ 2σ2)。*通常情况下，σ=1。答卷数据的谱聚类算法对答卷数据进行分层过滤分析算法包括以下几步：

1. 读入考试数据库记录。答卷数据对象集合表示为*X={x1,x2,…,xn}*,*xi*。其中，*xi*为第*i*个被试的答题信息，*P*为试题数。计算被试的答卷数据对象的距离矩阵*A。*
2. 计算被试的答卷数据对象的相似度矩阵*W*，矩阵*W*的阶数与矩阵*A*的阶数相同。
3. 计算扩展邻接矩阵*L。*根据相似度矩阵*W*建立一个对角矩阵，记为*D=*，*dii=*，令矩阵*L=D-1/2W D-1/2。*
4. 计算矩阵*L*的特征值和特征向量，选择前*k*个特征向量*u1,u2,…,uk*为列的矩阵*U，*并对矩阵*U*的每一行进行单位化处理。
5. 设*yi*（*i*=1,2,…,*n*）是对应矩阵*U*的第*i*行的向量。用k均值算法，将向量*yi*（*i*=1,2,…,*n*）聚类到*C1,C2,…,Ck。*
6. 建立映射*xi*|→*yi*，*i*=1,2,…,*n。*根据*xi*和*xj*之间的对应关系，利用5）中矩阵*U*的行向量*yi*的聚类结果，确定点*x1,x2,…,xn*在聚类中的结果。

1.3

基于谱聚类的分层过滤是对被试者进行的第一次分类过程，该分类过程中仅仅用到被试者的考分数据信息，但是在一些特定的选拔过程中，仅仅依靠分数来衡量被试者的能力显得科学性不足，不能真实反映被试者综合能力以及应试过程中因特殊情况导致的特殊群体。因此，在第一次针对考试分数的分层过滤基础上，对聚类结果进行二次校正。以单选客观选择题为例，假设某试题有四个答案*A、B、C、D*。其中：*A*为正确答案，得分为*x*；*B*为强相关答案，得分为*k1\*x；C*为弱相关答案，得分为*k2\*x*；*D*为不相关答案，得分为*k3\*x*。*k1*> *k2*> *k3*且取值区间为[0,1),假设考试题目数量为*N*,某学生答错的题目数为*M，MN*。答错题目中权重为*k1，k2*，*k3*的错题数目分别为*M1*、*M2*、*M3*，*M1*+*M2*+*M3*=*M，*则该被试者最终校正得分*X*记为：

*X=*( *M1 k1*/ *M*+ *M2 k2*/ *M*+ *M3 k3*/ *M*+*(N-M))x* (1)

同理，对于多项选择题，假设某题目存在*A、B、C、D、E、F*六个选项，对应答案出错的题目其正确答案的选项个数为*N\**,1*N\**6, 被试者选对的答案选项数为*M\**,0*M\*N\**,，被试者选错的答案选项数为*K\**，0*K\*N\**,则被试者对该多选题的最终得分*Y*定义为：

(2)

基于以上描述，可以在被试者基本考试成绩分类的基础上对其实际能力进行一次新的校正和衡量，二次校正算法描述如下：

1. 输入单选与多选题目总数分别为*N*、*N\**以及单选多选题目的开始与结束题号，每个单选与多选题得分分别为x、x\*，，并输入多选题答案得分标准表，表中的每个记录是一个三元组(*M\**，*N\**，*K\**);
2. 将答题卡进行扫描;
3. 对于单选题，根据式(1)计算单选题校正得分*X*；
4. 对于多选题，根据式(2)计算多选题校正得分*Y*；
5. 计算总校正得分*X+Y*。