一种面向海洋大数据的质量检验选择算法

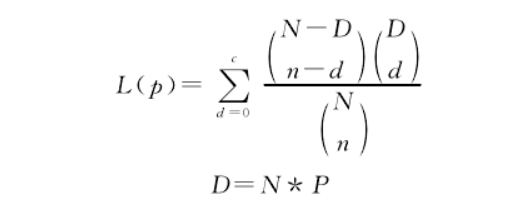
《海洋数据质量检验方案中残差优化选择算法》一文中，作者提出了分批排序选择算法进行海洋数据的质量检验。在此方法中，首先选用百分比抽样方案对海洋数据建立抽样检验方案，记为S（N，n，c），然后对其进行抽样检验；然后基于接收质量限（AQL）和极限质量限（LQL）假定二者对应的不合格品率，利用超几何分布公式，求解二者对应的接受概率值；然后求出两种情况下的残差Ea和Eb；最后依照Skyline计算，将每一个抽样检验方案中的一系列Ea和Eb作为一个点集合，对集合中的点进行筛选，判断筛选之后的集合中的点之间的相互支配关系，经过计算得到该抽样检验方案的Skyline集合，依据Skyline计算的累加性，将多种百分比抽样所得的优化选择方案归并在一起，再判断点与点之间的相互支配关系，从而得到最优的抽样方案或方案集。

反思：

1. 本文采用百分比抽样对海洋数据进行分批，是否可以选择其他的分批方法进行对比
2. 使用了接收质量限和极限质量限，其中需要设置两个参数α和β，这两个的参数设定是否获得更好的值（机器学习？），见附录（2）（3）
3. 超几何分布可否使用泊松分布？

**附录**

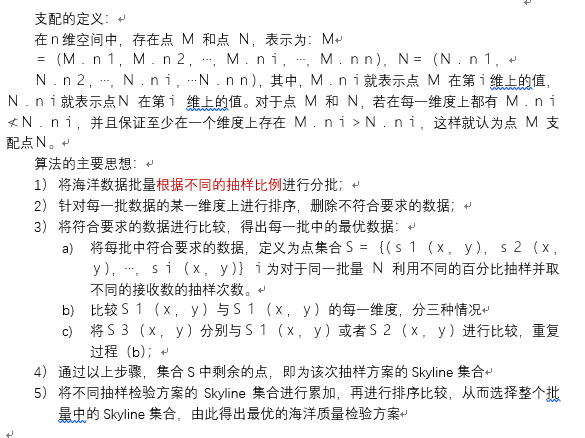
（1）超几何分布公式：



（2）AQL：接收质量限，可允许的最差过程平均质量水平，它是可以接收和不可以接收的过程平均界限值，不合格率的最大值；但是由于抽样检验的随机性，达到质量精度要求的检验批，也可能会被拒收，所以当检验批的质量水平等于或小于AQL时，其被判为不合格的概率应保证不大于α，则接收概率应该大于或等于1-α。

（3）LQL：极限质量限，为了抽样检验，限制在某一低接收概率β的质量水平，它是在抽样检验中对不应接收的批质量的最小值。将不合格的海洋数据检验误判为合格。

（4）Skyline计算是从一个数据库中提取不被其它任何数据对象所支配的数据对象集合；



面向海洋大数据采集的协同计算模型

由于海洋大数据的数据类型多，需要信息采集算法能够综合集成多种数据源，故采用基于多种数据、方法、只是的协同计算来实现这一过程。

专题信息提取方法，提取用户感兴趣的一种专题信息而忽略其他地物信息，要求该类信息提取的精度及自动化程度高

协同计算：

1. 多分辨率协同计算：多分辨率协同能够增加信息提取的精准度，综合考虑低中高分辨率卫星，实现多数据源、多分辨率数据的协同计算。高分辨率图像能够提高更精准的图信息。
2. 多知识协同计算：知识的融入能提高遥感信息提取的精度及效率。需要一些先验知识的辅助，通过多种知识的协同辅助提高信息提取的精度；另外知识模型的加入也会提高识别精度或降低其他干扰信息。
3. 多算法协同计算：不同算法针对同一数据可能会得到不同的信息提取结果，而综合多种算法可在一定程度上提高成果的可靠性。
4. 多计算资源协同计算：综合应用计算机提供的计算资源能够提高信息提取与计算的效率。

结论：采用专题信息提取方法，根据不同的信息特征建立规则，实现自动化提取信息。

面向海洋大数据的副本放置策略方法

副本技术是大数据存储和管理的关键技术，其目的是提高系统的效率和可靠性，对于创建好的副本，选择一个合适的位置来存放副本同样至关重要。

云环境：存储结点数量大、用户复杂多变、节点性能差异大是其代表性特点。

优化放置策略：基于节点评价，使用多目标优化算法，用目标加权法找出满足约束条件的解。所要进行优化的4个子目标是节点存储负载率、节点可用率、节点传输带宽占比率、机架中节点使用率。然后根据相关文献找出合适的权重并赋予四个优化目标，根据优化算法计算出各个节点的评价值，将评价值由高到低进行排序，在节点的放置过程中优先选择排序靠前的高评价值节点来存储。

副本数的动态决策：不同的数据受欢迎程度有所不同，甚至是同一数据，在不同的时间段中其受欢迎的程度也可能不同，通过计算每个数据的热度属性，以衡量数据受欢迎程度，针对受欢迎程度高的数据建立更多的副本，多个副本可以分散数据的请求操作，放置产生系统热点，最大可能地减少网络传输的时延，提高系统的响应速度；若数据的受欢迎程度低，则为其建立较少的副本，这样不仅能保证数据的可靠性，而且能尽量减少副本的冗余度，在一定程度上节省系统的存储空间。

兼顾负载均衡的副本快速恢复策略：对数据块的恢复排好先后顺序，热点文件的数据恢复优先级高一些，这样就能加快热点数据文件的恢复速度。在源节点和目标节点的选择中，要考虑哥哥节点的负载程度；设计准确的节点负载计算方法，然后计算并存储各节点的负载值和机架的平均负载值。源节点选择标准是存有失效副本的备选节点中负载最小的，目的是提高副本的复制效率。为了快速复制，减少带宽竞争的压力，目标节点的选择也最好是与源节点相同的机架上。再考虑到失效节点的不同数据副本是存储在不同的节点上的，这样在恢复重构数据时可以引入并行过程，多个数据的恢复在不同节点上同时进行，能够分散负载压力，提高重构效率。