论文名：《基于多数据源排序和整合的计算药物重新定位》

这篇论文是一个整合的计算框架，为每个数据源定义了不同的相似性度量，并利用加权k-最近邻算法将最近邻的相似性转换为给定化合物的预测分数。作者在摘要中，有一个大边际法，将多个来源的单个指标组合为一个全球指标。同时作者对数据源对预测的贡献进行排序。

作者方法名：基于相似性的多源学习（Similarity-based Large-margin learning of Multiple Sources, SLAMS）,方法对多个药物信息源进行排序和整合，以促进预测任务，同时作者提出的方法是可以扩展的，可以纳入其他类型的药物信息源。

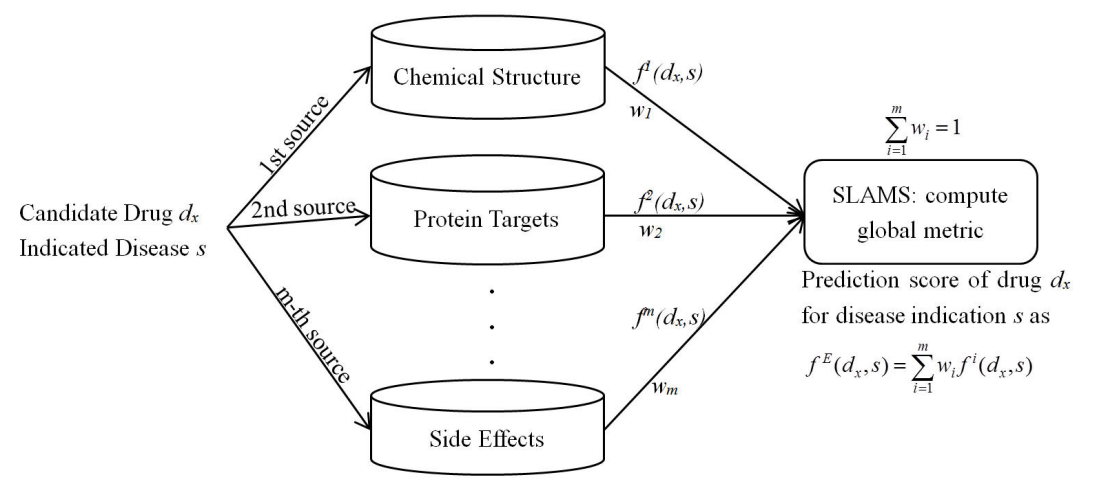
这篇论文的优点：

1. 作者同时考虑了多种化学特性、生物特性和表型特性。SLAMS算法还可以纳入其他类型的药物属性。
2. 相比于粗暴的方式尝试了多个数据源的所有代表性权重，SLAMS为所有数据源分配了权重，无需人工调整。
3. 作者使用大边际方法（即最小化铰链损失）来整合多个数据源，从机器学习理论的角度来看，这种方法通常比逻辑回归方法（即最小化对数损失）更为理想。同时，大边际方法得出的权重向量也更具有可解释性。
4. 作者使用典型相关分析（CCA）来估算副作用资料的缺失值。然后，作者用预测的副作用资料来增强已知的副作用资料，以建立一个新的副作用来源。
5. 作者使用多种衡量标准（如精确度、召回率、F分数）来评估药物重新定位实验的结果。以前的许多方法只用ROC曲线下的面积（AUC）来评价其性能，但在一个高度不平衡的分类任务中，AUC的高分并不意味着什么，不幸的是，药物重新定位就是这样一项任务。

**SLAMS方法步骤：**

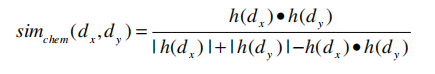
m个数据源参与了整合过程。每个候选药物dx查询第i个(i=1,,m)数据源，并获得指定疾病s的预测分数为f i (dx,s)。然后，来自多个数据源的m个预测分数被合并为一个单一的、最终的分数fE(dx,s)。

作者的整合多个预测分数的方法：k-近邻分类器对单个数据源进行评分以及通过大边际方法整合多个预测分数的细节。



1. 相似性计算
2. 化学相似性计算

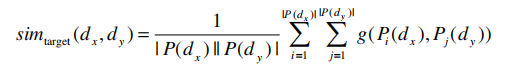
基于每种药物化学结构的二维化学指纹描述符计算成对相似度simchem(dx,dy)的方法，使用化学开发试剂盒将每个化学成分编码成PubChem中定义的881维化学子结构载体。也就是说，每个药物d用一个二进制指纹h(d)表示，其中的每个比特表示预定义的化学结构片段的存在。将两种药物dx和dy的化学相似性计算为其指纹的谷本系数：



|h(dx)|、|h(dy)|分别为药物dx、dy中的结构片段数。h(dx)\*h(dy)点积表示两种药物共用的结构片段数。相似度得分在[0,1]范围内。

1. 蛋白质靶点相似度计算

基于两个目标蛋白集的序列相似性的平均值：

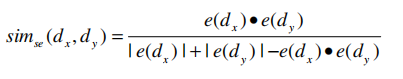


在给定药物d的情况下，我们将其目标蛋白组表示为P(d);| P(d)|为药物d的目标蛋白集的大小。两个蛋白g的序列相似函数计算为Smith-Waterman序列比对评分。相似度得分在[0,1]范围内。

1. 副作用相似度计算

作者利用已知数据集对副作用数据集进行预测，因为有些药物没有副作用数据集描述，存在缺失。

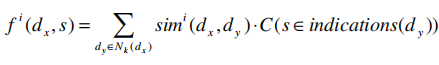
副反应关键词从SIDER数据库中获取，该数据库包含已上市药物的信息及其记录的不良反应。每个药物d用1385维二进制副作用谱e(d)表示，e(d)的元素分别用1或0编码每个副作用关键字的存在或不存在。用两药指纹的谷本系数计算两药dx和dy的两两副作用相似度：



|e(dx)|和|e(dy)|分别为药物dx和dy的副作用关键字的计数。e(dx)\* e(dy)点积表示两种药物共有的副作用数。simse得分在[0,1]范围内。

1. 单一数据源计算预测分数

为了计算药物dx具有指示的可能性，我们使用了k-最近邻(kNN)算法的加权变体。模型参数k的优化采用交叉验证设置(本研究中k=20)。对于第i个数据源，某一适应症s对药物dx的预测评分f计算为：



其中 sim i (dx,dy) 表示来自第 i 个数据源的两个药物 dx 和 dy 之间的相似性分数，C 是一个特征函数，如果 dy 有适应症 s，则返回 1，否则返回 0，Nk(dx) 是药物 dx 的 k 个最近的邻居，根据由第 i 个数据源的类型决定的度量 simi。衡量标准simi可以是前面几个小节中定义的相似性之一（即化学结构、蛋白质目标和副作用），也可以是任何其他类型的药物信息源。因此，SLAMS算法是可扩展的。作者为药物重新定位任务提出了一个k-NN评分组件，因为它在多个数据源上的实现很简单，可以直接使用多个分数，而且其准确性与更复杂的算法相比具有竞争力。

1. 结合多种策略

考虑多个数据源，并为每对(d,s)获得多个预测得分。给定一个药物-疾病对的m个分数(d,s)(即，有m个不同的数据源)，我们提出一个大边际方法来计算最终分数f E作为个人分数的加权平均值：

