关于数据挖掘的一点想法（复杂网络、机器学习、群体智慧）

今天去听了遥感所的分享，也听了韩老师和刘老师的一些关于数据挖掘的心得，有些启发。之前刚进入数据挖掘这个领域，接触了复杂网络，信息物理和机器学习之后，有些迷失。

面对复杂网络，觉得这个不就是将复杂的数据抽象为以关系为边的网络，然后以网络的思维去挖掘信息。复杂网络相关研究成果目前已成功应用于推荐系统(物质扩散、热传导、三部图推荐、超图算法等)，社交网络的挖掘（圈子的划分、重要节点的识别、网络性质（小世界、无标度）等的研究），信息传播（谣言控制、疾病传播、信息源确定），链路预测（一些基于结构的经典算法）等等领域。**复杂网络，个人理解，就是以网络的视角去分析研究，找出可被解释的结果或者现象，相当于寻找隐藏在数据背后的规则，特别是关联关系的挖掘。几乎所有的数据都可以使用这种方式来做相关处理**。说实话，这些发现虽然简单，但其实并不简单。因为所研究出来的成果，一般都是数据所蕴含的主要的一些规则，就拿链路预测作为例子，简简单单的一个 common neighbor 就可以达到非常好的效果，其对数据的假设就是用户关注是因为他们之间的共同朋友多，这种现象在实际的数据中真的非常普遍。复杂网络具有物理界的性质，就是结果虽然简单，但就是能反映出本质，可以被解释，如同能量方程。

同时，因为学习推荐算法，也接触了一些机器学习相关知识。一般拿到的数据是用户对商品或者电影的评分数据。仅仅使用结构信息，总是缺少了可被利用的信息。而像传统的协同过滤，需要去计算用户或者商品的相似度。通过挖掘用户以及商品的潜在特征，或者结构，可以用于求解相似度。先是接触了 SVD 算法，一方面可以得到用户与商品的特征矩阵，另外又能通过该算法实现降维。后来又接触到了矩阵分解算法，通过矩阵分解，可以得到用户的潜在特征矩阵以及商品的潜在特征矩阵，然后可以得到其他评分信息。或者使用这种潜在特征矩阵做聚类，做相似性判别等等处理。说白了该算法是对原始值的一个拟合，因为损失函数就是使得预测评分与原始评分相差越小越好。后来又融入评分偏置，社交网络信息等等。但这样的损失函数设计真的好吗？真的是能反映出真实的规律吗？慢慢的，随着研究课题的原因，接触了聚类算法（Kmeans，层次分析），分类算法（logistic regression、朴素贝叶斯、K 近邻、SVM、神经网络、决策树、RF，boost 等）。后来因为链路预测以及推荐算法的原因，又接触了排序学习，其相对于原来的一些算法，不再是对具体数值的拟合，而是关注排序本身。

学了这么多，总体感觉就是其实机器学习与复杂网络的本质是一致，都是去挖掘数据内在的规律。但复杂网络偏向于寻找可被解释的本质规律，特别是可以被数学精确表达的或者是普适性的。机器学习则偏好于通过算法模型去训练学习出数据的本来分布，当然这种分布很可能是未知的，需要尝试多种算法模型才可能可以找到，另一方面则是特征的抽取，寻找到潜在的能够表达出本质的因素。但现在看过很多的比赛，一般都采用机器学习，不过在特征选择上偏向于暴力，不优美，感觉像医生在用杀猪刀做手术。

上次是参加了阿里的竞赛决赛，看到有几位在处理特征上真的让人大跌眼镜，就是拍脑袋觉得可行或者完全解释不了的东西放到模型里去训练。当然工业界也流行这样，不怎么提挑选特征，有没有用先放进去试，有种非科学的味道。当然，把你所有能够想到的或者能利用的信息都当成特征放到模型中，处理得好（比如去除一些异常的数据信息）也许会得到好的结果，可是这非常的不可控。因为并没有找到真正的本质，当新的数据来了，难道又这样乱七八糟搞一通吗？难到大数据时代，我们只需要机器学习，而不需要人的学习吗？当机器真的会学习的时候，人的思考能力将不堪设想，如果这样下去。不过，**如果我们对数据进行深入分析理解，挖掘出真正的可被解释或者可被验证的规律时，再使用机器学习算法，将会得到更加高效且准确的结果。**之前看了 BPR-opt 的算法，其算法核心就是使得训练得到的评分矩阵，被评分的高于没有评分的，该分差越大越好，当然其算法也融合了协同过滤的思想。我觉得这样的假设是经过思考，并且是可以反映出数据的本质，因为推荐或者检索，我们更关注排序靠前的结果的正确性。当然后面可以考虑加入原始评分大的要比评分小的排名更靠前的知识。

此外，看过一篇说为什么 google 的算法就是比其他公司牛逼的文章，它提到规则库的构建，google 的规则库可能是世界上数一数二的。举个例子，比如在淘宝上我们要对商品聚类，通过分析用户浏览商品的轨迹，很可能就可以得到超过一些现有算法的效果了。因为用户浏览商品的过程中，连续的商品轨迹很可能就是同一类商品。另外，比如地图软件，要规划 AB 两点的最短路径，从算法角度来说，可以从时间最短或者路径最短来考虑，但是万一 AB 之间的路发生了堵车或者正在施工无法通行呢？软件是不知道的。但是通过观察分析用户在 AB 之间的运动轨迹，当然是很多次的轨迹，也许就能找到最优的线路，应该还可以记起来迪斯尼乐园因为道路规划而获奖吧，群体智慧是非常重要的可被利用的信息。如果算法能够整合群体智慧知识，我相信算法的想过一定会超越以前。不妨，使用复杂网络的知识去挖掘潜在的可被证明或者普适的规律，再融合群体智慧知识，然后构建相关的特征，最后使用相关的机器学习算法去解决问题（当然模型的损失函数需要好好考量，必须符合你的任务）。

另外，关于大数据的一点思考。为什么大数据可行？韩老师说关键在于数据的打通，不同来源的数据如果能够打通，就可以得到 1+1>2 的效果。如果阿里、百度、腾讯、微博这四家数据可以实现互通，将可以对每个人构建档案的构建，我们将知道每个人的需求，爱好，社交圈、什么时间要做什么事等等，智慧城市将提前实现。盲人摸象一样，一层数据可能构建了你的一个片面，然后另一层补上一个片面，当多元异构数据融合到一个，就是完整的一个人。我们人类无时无刻都在产生数据，数据也无时无刻都被保存，你的一切其实都可被数字化，预测将只是一个时间问题。