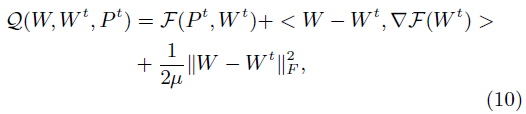
**关于《missing value learning》论文的学习报告**

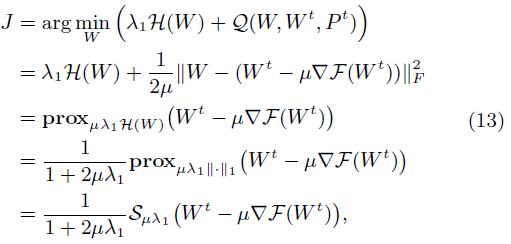
论文中提及的算法的主要思想是：无论是对有标签的还是无标签的数据集，对于给定的数据集，某一个特征可能与其他的一个或多个特征有关系，因此可以通过一个初始矩阵不断迭代地计算，用权重矩阵表征特征之间的相关性并通过矩阵乘法实现特征的相关性组合，通过训练矩阵与初始矩阵已知数据的相似性表征训练数据与原始数据的拟合程度。

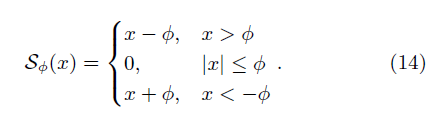
算法合理性：算法的主要依据是根据数据样本的某些特征之间有较强的相关性的特点，正如人的身高与脚长的关系：正常来说，人的身高是脚长的6.725倍，因此假如给定关于人类的统计数据，对于某个给定样本，已知身高数据，但是缺失了脚长的数据，那么就可以根据这两个特性的相关性对缺失值进行训练学习，得到的结果的可信度也会较高。因此，基于这样的思想，这个算法的提出是有一定的合理性的。

算法的细节分析：

1. 对于权重矩阵W的要求，首先要求这个矩阵的对角线元素都为0，这样做的目的是根据矩阵乘法的运算规则，，等价于P矩阵的第i行根据权重矩阵中的第j列的元素的加权求和能得到新矩阵的一个元素，对应于算法思想中某个特征与其他的一个或多个特征相关的思想，因此数据矩阵中的某一个特征的数值是根据同一个样本其他特征确定的（加权求和）而与自身特征无关，因此矩阵对角线元素为0这一特性能避免误差的积累，也就是说当某次迭代过程中某一特征估计错误不会对下一次迭代产生影响。
2. 在目标函数中加上矩阵W的L1范数和frobenius范数的目的是：在机器学习方法中，很多时候要避免过拟合的现象，为了算法训练中过拟合产生的不好效果以及使算法泛化，一种主要使用的方法是正则化，而基于L1范数和frobenius范数的正则化是常用的正则化方法。L1范数能产生稀疏解，而通常稀疏解的泛化能力会比较好，之所以需要产生稀疏解是因为原始数据集中的某个特征只会和少数的几个特征有较强的相关性，因此稀疏矩阵中多数的0项就能把与待求特征相关性较弱的特征的贡献程度降低。而由于基于L1的正则项并不是处处连续的，而frobenius正则项具有连续可微的特性，在求导时特别方便，因此能增加算法的稳定性。
3. 关于class relationship这一约束条件：在论文中，这一约束条件是通过K-Means的算法思想给定的，K-Means算法的主要思想是每一个聚类都用一个聚类中心表征，聚类中心的计算是通过该聚类中所有点的相同维度的均值求出。因此在每一次训练矩阵P的时候通过判断每一个数据样本与其对应的聚类中心的距离表征聚类的合理程度。
4. 在算法训练矩阵的过程中，用的是初始化一个完整的矩阵P进行训练而不是用不完整的原始数据矩阵R进行训练的原因是避免训练出来的新矩阵只适应于R矩阵的已知数据，因此算法是首先通过一个与原始数据矩阵R无关的P矩阵，通过把矩阵作为一个整体进行训练从而使训练矩阵能整体适应原始数据，目的是训练出一个与原始数据矩阵的已知数据相近而不是相同的新矩阵。
5. 对于约束条件,如果严格限制该约束条件等于0可能导致训练出的矩阵与R一样而缺失值并没有通过训练学习出来，同时，由于数据中存在的噪声，松弛了这个约束条件能增加噪声容忍度从而避免了过拟合的现象，提高算法的泛化程度。
6. 对于参数的选择，根据目标函数，当的取值太大，函数的值基本上取决于第二项（判决器），而判决器采用的是K-Means算法，因此当缺失值等于该特征的平均值时判决器函数的值能达到最小，缺失值学习算法就变成了MEAN算法。
7. 关于目标函数的优化，由于目标函数还附带了约束条件，因此可以通过拉格朗日乘子法进行求解，拉格朗日乘子法的主要思想是把约束条件与目标函数联系到一起从而生成新的函数，对新函数进行求解可得到满足约束条件的目标函数解。于是得到拉格朗日方程（论文中的式子（7）），对矩阵P的优化是通过梯度下降法求解，梯度下降法的主要思想是要找到某函数的最小值，最好的方法是沿着该函数的梯度方向搜寻，是步长，即沿着梯度方向，每次走的距离，然后在该点重新计算梯度方向，再走的距离，一直搜寻下去。由于P的每次更新都是按照搜寻目标函数最小值的方向进行，所以最终能找到一个使目标函数尽量小的P。这里有一个疑问是对于式子（7）的梯度计算没能理解清楚怎么求解。
8. 关于矩阵W的优化，涉及到两条公式（10）和（13）：







一开始看到公式（10）的时候没能理解后面两项的意义以及用公式（10）替代函数F的作用，但是看到公式（13）以后就清楚了，要求出公式（13）就要求出满足使目标函数最小的权重矩阵W，注意到与拉格朗日公式（7）相比，这里的目标函数的少了几项，由于这里是针对W进行优化，因此与W无关的项可以忽略。而公式（10）中后面两项的意义在于：第一项是添加一个约束条件保证搜寻函数最小值的时候往梯度的方向搜寻；第二项是为了保证每次搜寻的时候到达的是一个不太远的点，这样才能使第三项的值不至于太大（proximal算法的思想）。这里也有一个疑问是关于公式（13）的后面的推导没能理解。可以看到，关于S的函数（14），我们看到当x满足一定条件是函数值为0，这是稀疏矩阵的0项的来源，这是前面提到的正则化的作用。

1. 关于拉格朗日乘子Y和s的优化，在满足约束条件的情况下根据P和R的拟合（/偏离）程度调整Y和s，这个算法思想与机器学习方法中的logistic回归算法很相似。