融合排序学习算法的局部回归

王晓倩

1 背景

大数据环境下，数据量骤增，数据维度庞大，且数据的分布情况往往比较复杂，在回归分析中，如果采用所有的样本数据进行估计，往往难以得到理想的效果。整体分布复杂的数据集在局部往往呈现出较规律的分布特性，这就应用到了局部线性的思想。其中，在流形学习、高等数学和局部建模等相关内容中就体现了局部线性的思想。

- 流形[1]学习中：若低维流形嵌入到高维空间中，则数据样本在高维空间的分布虽然看上去非常复杂，但在局部上仍具有欧氏空间的性质。在流形学习中，如果数据分布很复杂，可以有一个局部流形跟它同胚。
- 高等数学中：泰勒公式是一个用函数在某点的信息描述其附近取值的公式。如果函数足够平滑，在已知函数在某一点的各阶导数值的情况之下，泰勒公式可以用这些导数值做系数构建一个多项式来近似函数在这一点的邻域中的值。
- 局部建模中：局部建模即任意点X的一定邻域内，模型m()都线性的围绕X，于是将线性回归技术应用于围绕X的部分数据。范剑青提出局部建模方法是弥补多项式回归缺陷的一种途径。

因此得出结论：整体分布呈现非线性的数据集，局部可以用线性做近似。因而，局部回归算法具有更好的估计效果，对于局部回归算法，局部范围的确定是一个关键问题。而现有方法中，局部的范畴基本上都是基于点X的邻域范围来确定。这一方法对

[1]流形（Manifold），是局部具有欧氏空间性质的空间。欧氏空间就是流形最简单的实例。像地球表面这样的球面是一个比较复杂的例子。一般的流形可以通过把许多平直的片折弯并粘连而成。

于分布复杂的数据集而言，确定的局部范围准确性不高。如当数据总体分布为分段性函数形式时，间断点处的左右两端分布形式差异很大，但现有局部回归方法仍同时选取间断点两端的数据作为局部采样数据进行回归分析和预测，会造成较大的误差。因而，如何准确的描述局部的范畴，以提高局部回归算法的准确度和性能是当前研究的一个重要问题。

2 研究内容

本文在学习研究回归模型相关理论发展的基础上，为解决以自变量邻域为局部取样点的局部回归并不对所有分布情况总体（例如非连续型分布总体）都适合的问题，将排序学习算法融入局部回归的建模思想中，提出通过选择适当的排序学习算法作为局部回归模型选择局部点的取样方法，在这种情况下，局部回归不再只是通过简单的自变量邻域范围而是根据样本点之间的相似性来选取局部取样数据。

2.1 研究融合排序学习算法的局部回归模型

将机器学习算法－－排序学习引入局部回归分析，构建融合排序学习算法的局部回归。即通过排序学习算法确定局部回归的局部样本选择。

2.2 研究局部范畴参数K的取值方法

局部范畴参数K即确定局部回归的局部取样点数量，这直接影响到了回归模型的构建。如何确定局部范畴参数K的取值，以构建最优的局部回归模型是本文要研究的一方面内容。

3 方法设计

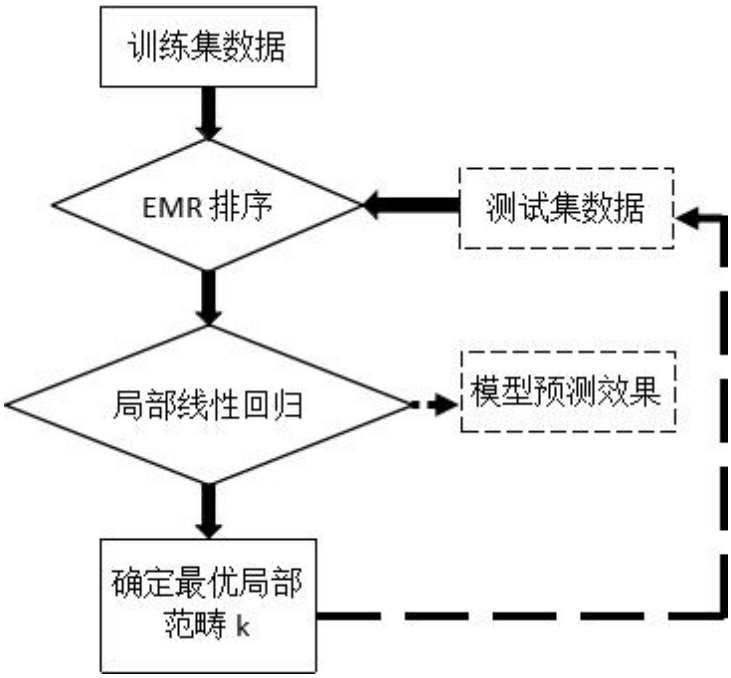
该方法的主要步骤如下所示：

1. 根据EMR算法对训练数据集进行排序；

2. 根据排序结果，选取不同的局部范畴k，即取点个数，进行局部线性回归，计算训练数据集的预测误差；

3. 根据不同局部范畴k对应的预测误差，确定最优局部范畴参数k；

4. 利用训练得到的排序模型，对测试集数据进行EMR排序。并按最优局部范畴参数k得到相应数量的局部点，进行局部线性回归，得到测试集数据的模型预测误差。



3.1 EMR算法

EMR（Efficient Manifold Ranking）算法，即高效的流形排序算法，是流形排序算法（MR, Manifold Ranking）的改进方法，因此首先介绍流形排序（MR）算法。

Manifold Ranking与传统的欧式空间上直接计算查询之间的相似性不同，该方法通过利用大规模数据内在的全局流形结构来计算排序得分。直观的解释为:首先构造一个带权网络，并且分配给源节点一个正的得分，其他待排序节点的得分设为0，然后每一个节点将其自身得分传播到邻居节点直到整个网络达到平衡状态。除源节点外的所有节点按照最终得分大小进行排序（得分越大,排序越靠前）。

算法1详细描述了MR算法的实现过程。

算法1: 传统的流形排序（MR）算法

输入项：数据集$X = \{x_0, x_1, \cdots, x_n\} \in R^m$，其中$x_0$是查询样本，其他查询$x_i(1 \leq i \leq n)$为候选查询。$f : X \to R$为排序函数，其为每一个查询$x_i(0 \leq i \leq n)$计算一个排序得分$f_i$，将$f$看成是一个向量$f = [f_0, f_1, \cdots, f_n]^T$。同时定义向量$y = [y_0, y_1, \cdots, y_n]^T$，其中$y_0 = 1$（$x_0$是源查询），其余$y_i = 0(1 \leq i \leq n)$。

输出项：样本得分序列f^*。

1. 构建整个数据集X上的连通图。方法是将数据集中所有样本与其距离最近的样本连接，得到连通图$G = (X, E)$，E为边集；
2. 构建关系矩阵K。如果边$e(i, j) \in E$，则$K_{ij} = exp(-||x_i - x_j||^2/2\sigma^2)$；否则$K_{ij} = 0$。由于不存在自环边，因此$K_{ii} = 0$；
3. 计算矩阵L。$L = D^{-1/2}KD^{-1/2}$，其中D为对角矩阵，$D_{ii} = \sum_{j=1}^{n} K_{ij}$；
4. 对于所有待比较样本$x_i(1 \leq i \leq n)$，进行迭代：$f^{(t+1)} = \alpha L f^{(t)} + (1 - \alpha)y$，其中$\alpha \in [0, 1]$；迭代至$f$收敛。其中，$\alpha$用来控制来自于先验的得分和来自于结构上相邻的结点得分对最终排序得分的贡献，α值越大表示来自于相邻节点得分贡献所占比例越大。
5. 输出样本得分序列f^*。f^*为样本得分序列$\left\{f^{(t)}\right\}$的收敛值，样本x_i的得分越高，表示其与查询样本的关联程度越高。

由于原始的MR算法在图的构建阶段和得分计算阶段开销很大，导致MR算法具有很高的计算复杂性，因而严重限制了其在大型数据集上的应用。基于此，Xu等人提出高效的流形排序算法（EMR）。EMR算法改进了MR算法中图的构建过程（即改进算法 1 中步骤 1 ），通过使用聚类算法构建样本到聚类中心的关联关系图取代原始MR算法的k近邻关系图，大大简化了图的构建过程。同时形成一种新形式的邻接矩阵来加速整个得分计算过程（改进算法 1 中步骤 2 ）。实验结果证明同原始的MR算法相比，EMR算法能够在相近准确度的基础上大大提高计算效率，因而对于大规模数据集具有更好的适用性。

- EMR算法改进了MR算法中图的构建过程。假设数据集$X = \{x_1, \cdots, x_n\} \in R^m$，首先我们需要使用$k - means$聚类算法计算得到若干个聚类中心$U = \{u_1, \cdots, u_d\} \in R^m$，我们的目标是构建关系矩阵$Z \in R^{d \times n}$，其中元素$z_{ki}$表示样本$x_i$与聚类中心$u_k$的关联关系。

- 关系矩阵$Z \in R^{n\times d}$相当于将数据集$X \in R^{n\times m}$映射到d维向量空间。EMR算法中新形式的关系矩阵为$K = Z^T Z$，含义为如果两个样本是相关联的$(K_{ij} > 0)$，那么它们至少有一个共同关联的聚类中心，否则$K_{ij} = 0$。EMR算法通过计算数据集在d维空间上的关联关系来衡量样本间的相似性。

3.2 最优局部范畴的确定

在本文中，通过计算不同局部范畴k对应的训练集数据的平均绝对误差，以平均绝对误差较小且相对稳定为标准确定最优局部范畴k。即通过计算平均预测误差来衡量回归模型的优劣。平均预测误差越小，表明回归模型越好。其中，平均绝对预测误差计算公式如下：

$$mae = \frac{1}{m}|y_{pred} - y|$$

4 方法应用及检验

通过模拟四组实验数据：严格的分段线性函数、添加扰动项后的分段线性函数、严格的双月亮数据以及添加扰动项后的双月亮数据，分别对这四组实验数据进行融合排序学习算法的局部回归，并展示其局部线性回归取样点的选择结果及回归预测效果，并将其与普通线性回归得到的拟合结果进行比较，从而说明融合排序学习算法的局部回归方法的有效性。

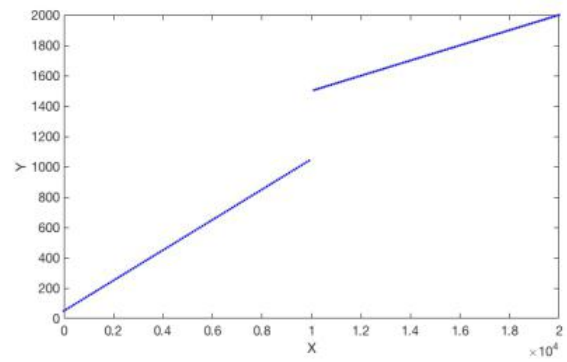
4.1 严格的分段线性函数

样本量$n = 200$

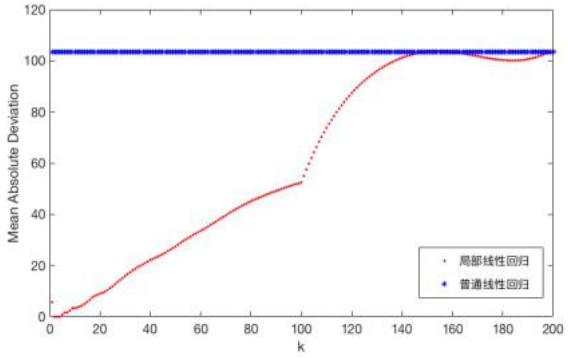
间断点$x = 1 \times 10^4$

间断点左侧函数形式$y = 50 + 0.1x$

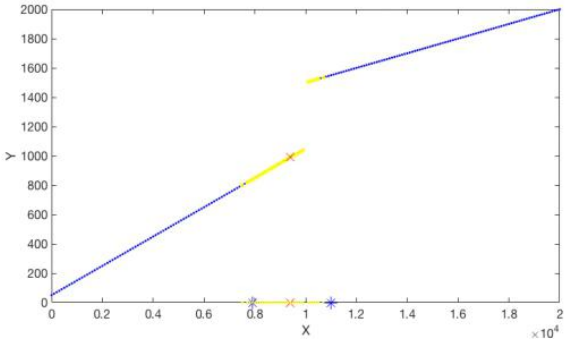
间断点右侧函数形式$y = 1000 + 0.05x$



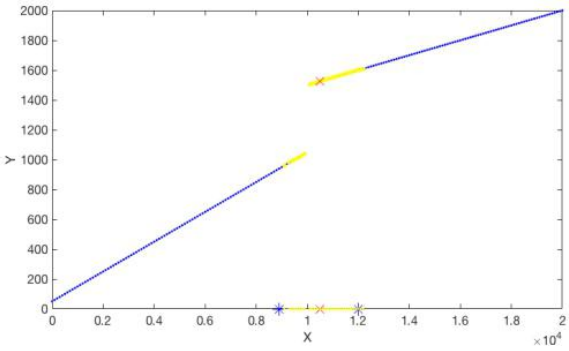
(a)散点图



(b)确定最优局部范畴k=4



(c)断点左侧样本选取结果(x=9401)



(d)断点右侧样本选取结果(x=10500)

图1　严格的分段线性函数

表1 线性回归与局部线性回归预测误差比较

	普通线性回归	局部线性回归(k=4)
平均绝对误差	99.5634	1.3443×10^{-13}

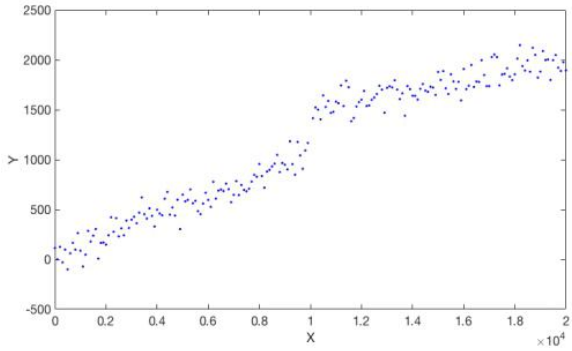
4.2　添加扰动项后的分段线性函数

样本量$n = 200$

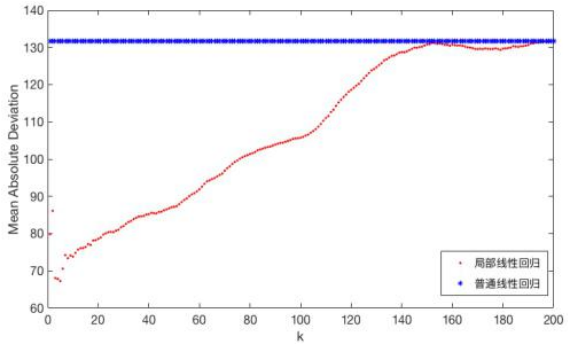
间断点$x = 1 \times 10^4$

间断点左侧函数形式$y = 50 + 0.1x + \mu$，其中$\mu \sim N(0, 100)$

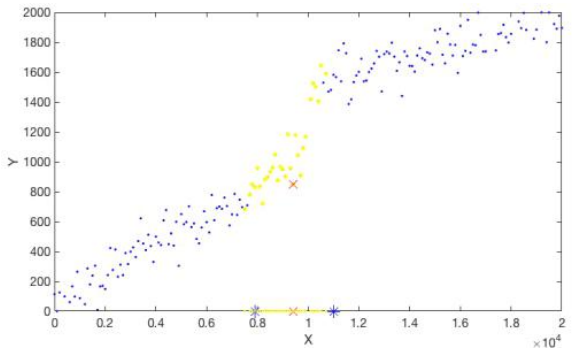
间断点右侧函数形式$y = 1000 + 0.05x + \mu$，其中$\mu \sim N(0, 100)$



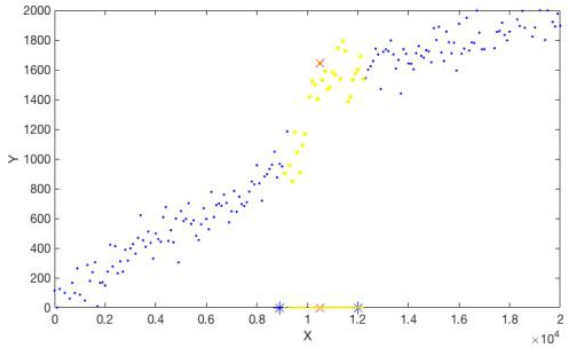
(a)散点图



(b)确定最优局部范畴k=4



(c)断点左侧样本选取结果(x=9401)



(d)断点右侧样本选取结果(x=10500)

图2　严格的分段线性函数

表2 线性回归与局部线性回归预测误差比较

	普通线性回归	局部线性回归(k=4)
平均绝对误差	137.9259	81.3227

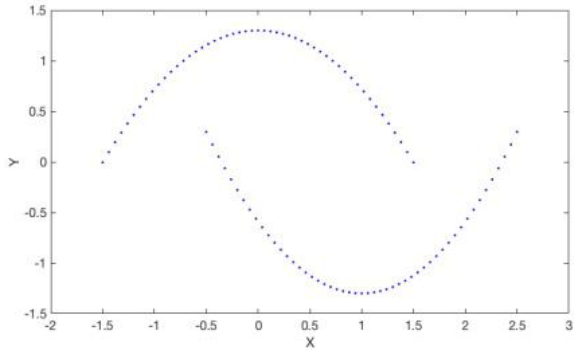
4.3　严格的双月亮数据

上月亮：$z = e^{x+y} + 100$，其中$y = -0.58x^2 + 1.3$

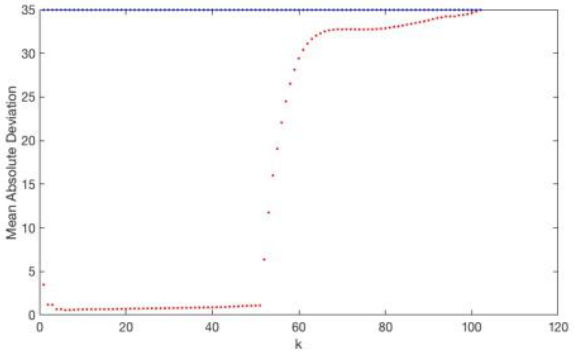
下月亮：$z = e^{x+y} - 100$，其中$y = -0.71x^2 - 1.42x - 0.59$

表3 线性回归与局部线性回归预测误差比较

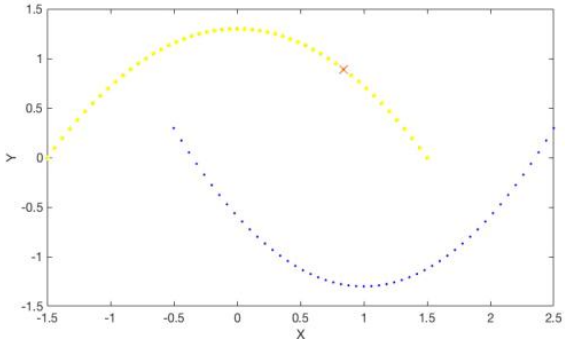
	普通线性回归	局部线性回归(k=50)
平均绝对误差	101.2432	97.8303



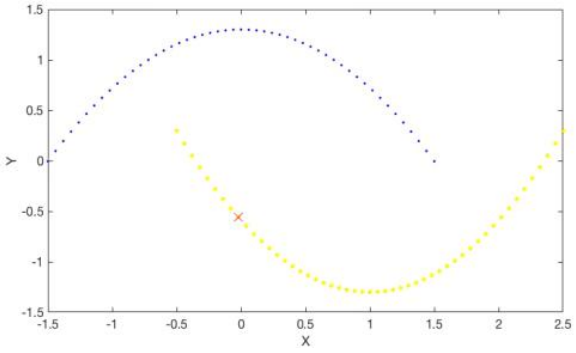
(a)散点图



(b)确定最优局部范畴k=50



(c)上月亮样本选取结果(x=0.8400,y=0.8908)



(d)下月亮样本选取结果（x=-0.0200,y=-0.5613）

图3　严格的双月亮数据

4.4　添加扰动项后的双月亮数据

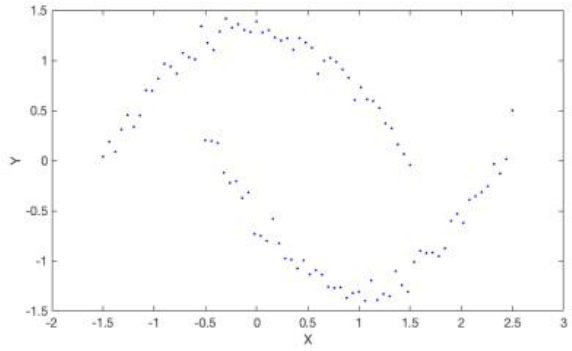
扰动项：$\mu_1 \sim N(0,1)$　$\mu_2 \sim N(0,0.1)$

上月亮：$z = e^{x+y} + 100 + \mu_1$，其中$y = -0.58x^2 + 1.3 + \mu_2$

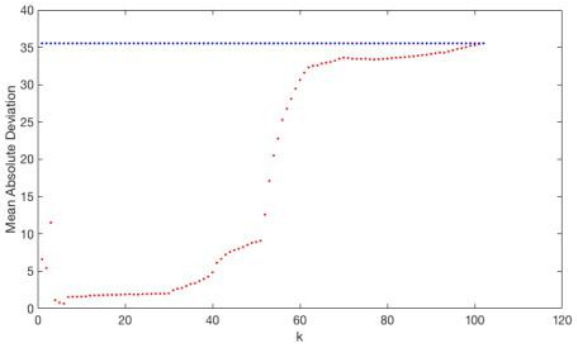
下月亮：$z = e^{x+y} - 100 + \mu_1$，其中$y = -0.71x^2 - 1.42x - 0.59 + \mu_2$

表4 线性回归与局部线性回归预测误差比较

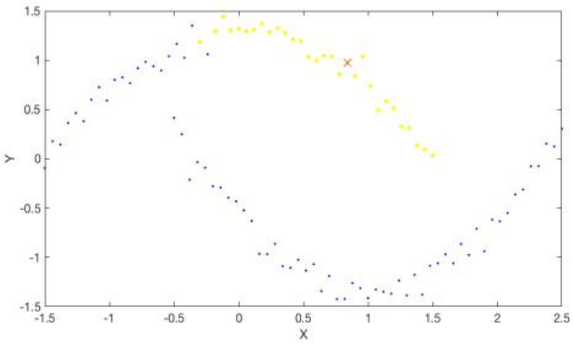
	普通线性回归	局部线性回归(k=30)
平均绝对误差	99.4997	91.0994



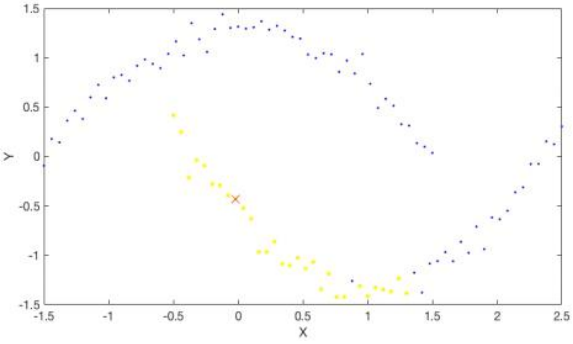
(a)散点图



(b)确定最优局部范畴k=30



(c)上月亮样本选取结果(x=0.8400,y=0.9720)



(d)下月亮样本选取结果（x=-0.0200,y=-0.4307）

图4　添加扰动项后的双月亮数据