



中国地质大学

CHINA UNIVERSITY OF GEOSCIENCES

北京 · BEIJING



数理学院

机器学习中数据特征提取与提升算法的研究

学院：数理学院

专业：计算机技术专业



学生：孙金龙

指导老师：高世臣 教授

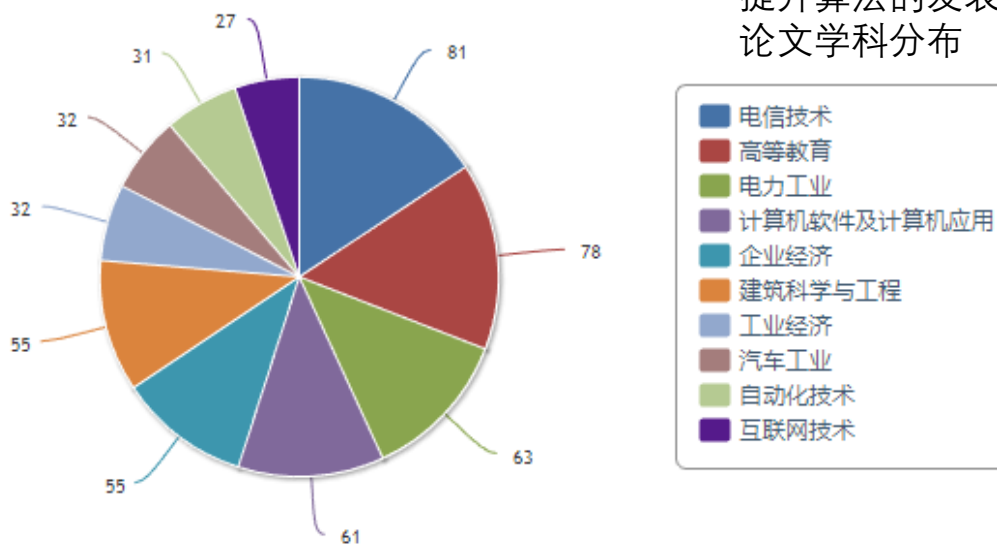
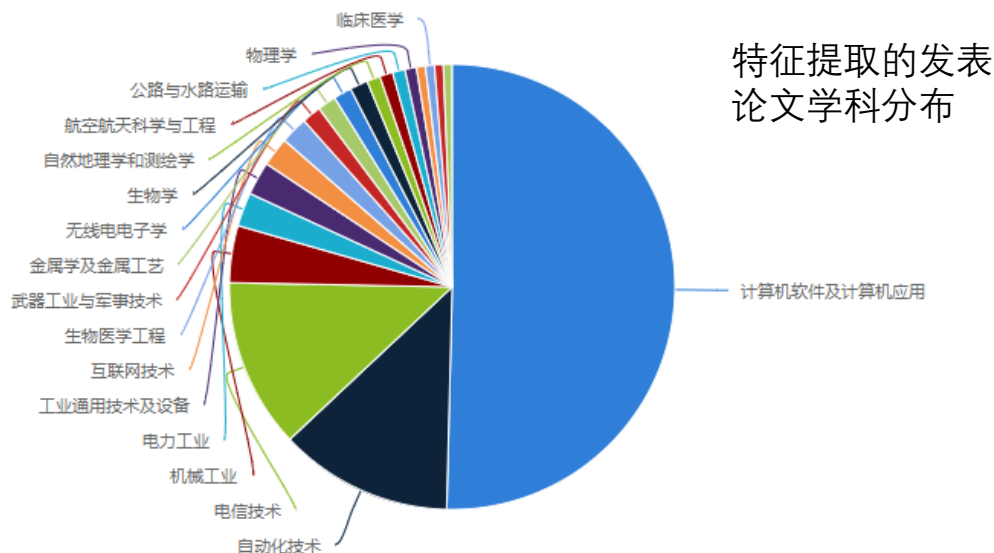
目录

contents

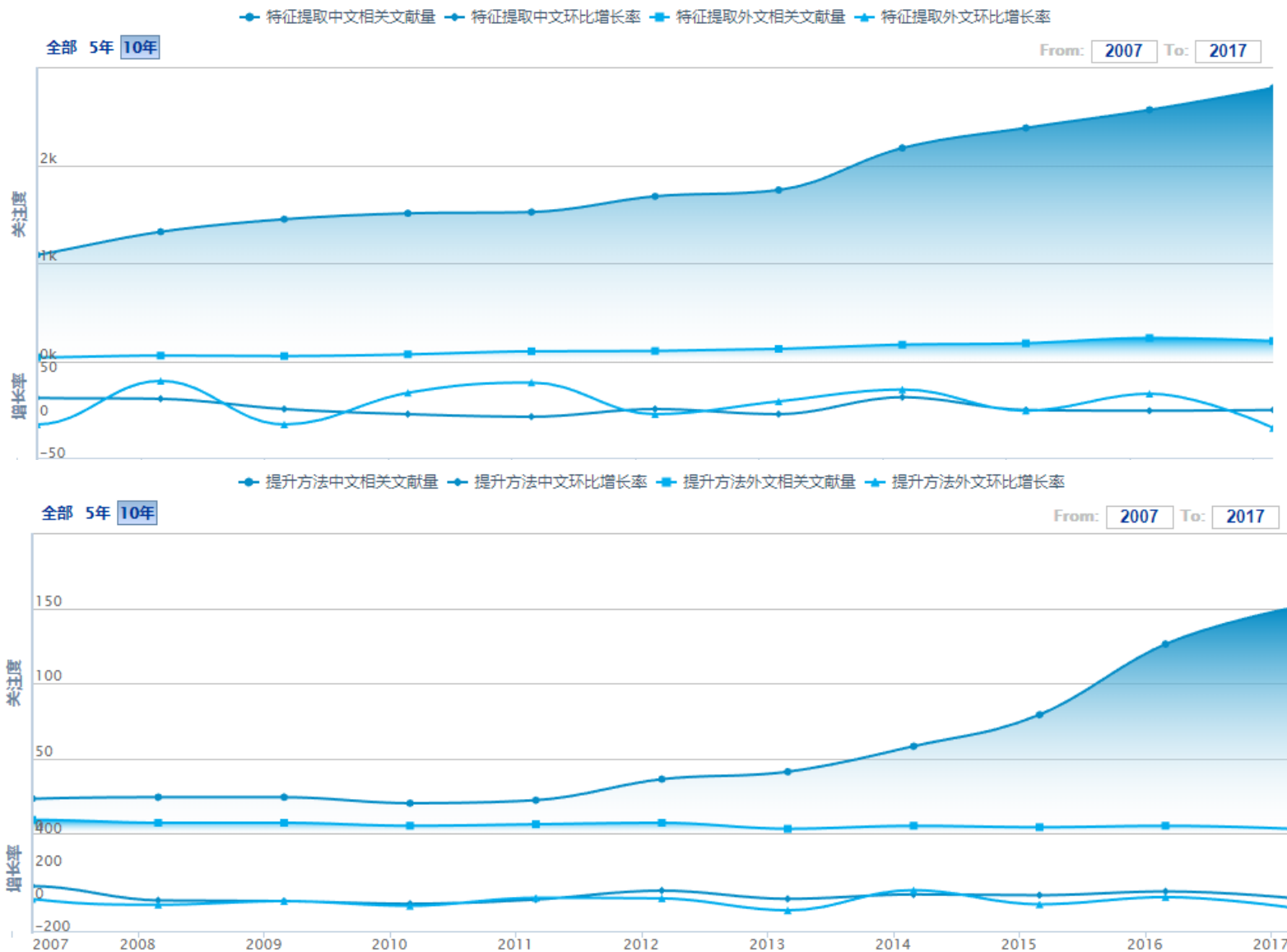
- PART 01/ 选题背景及意义
- PART 02/ 主要研究工作
- PART 03/ 研究流程与阶段性成果
- PART 04/ 进阶研究工作
- PART 05/ 预期目标和成果
- PART 06/ 时间安排

1-1 / 选题背景以及意义

人工智能技术高度发展的今天，其方法和理论均来自于机器学习的核心研究成果，而机器学习中最重要最具有研究价值的就是特征的分析与提取，特征提取作为数据科学的核心任务，如右图所示，是计算机技术专业必须掌握的数据分析技能。已经有很多学者在此领域取得了很多的成果，与此同时，如右图所示，近几年来，算法提升理论也逐渐备受关注，算法性能的提升也在机器学习理论中占据着很大的价值，二者并驾齐驱才能使得机器学习理论有着从量到质量的转换。



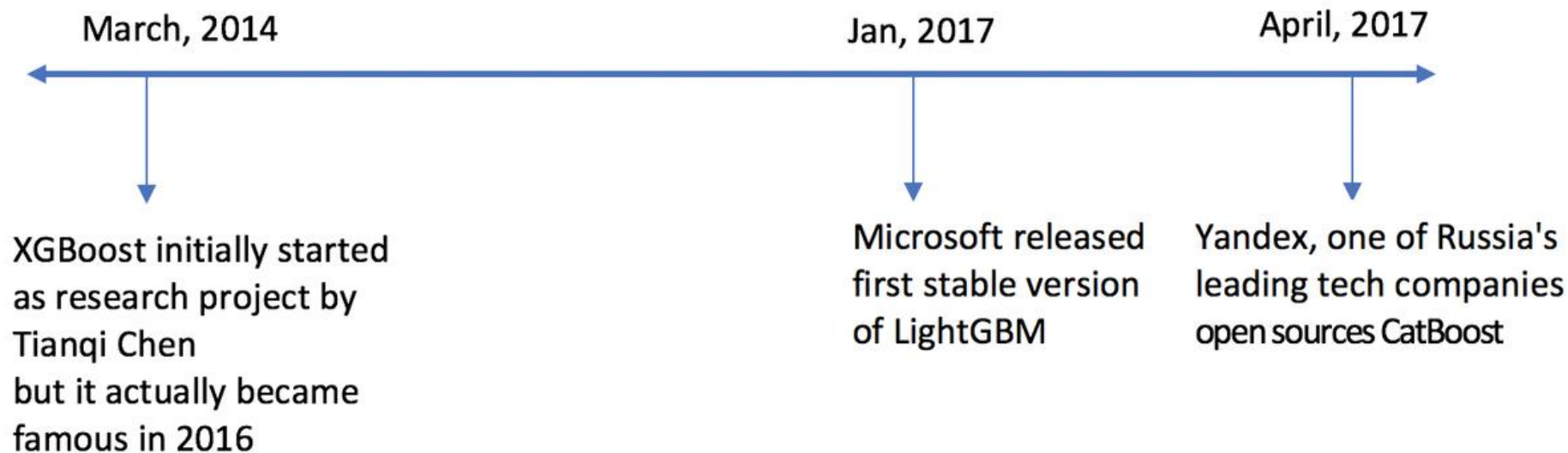
1-3 / 近10年特征提取和提升方法的关注度



特征提取

提升算法

1-2 / 提升算法背景调研



1995年	Freund 和Schapire 将Adaboost算法提出
1999年2月	Greedy Function Approximation A Gradient Boosting Machine论文发布，GBDT雏形产生
2014 年 3 月	XGBOOST 最早作为研究项目，由陈天奇提出
2017 年 1 月	微软发布首个稳定版 LightGBM
2017 年 4 月	俄罗斯顶尖技术公司 Yandex 开源 CatBoost

2-1 / 研究工作内容

本题将在研究过程中将运用统计机器学习的基础方法：决策树、KNN、SVM，集成方法：随机森林、GBDT、XGBoost和深度学习方法对数据进行数据特征挖掘与算法识别方面的研究，一方面对数据的空间性、邻域特征、卷积特征、条件概率特征进行分析、挖掘和提取；另一方面从算法的稳定性、性能、适用范围以及拟合情况进行对比。希望通过大量的实验，结果的横向和纵向对比分析，从而加深对数据的阅读，理解，以及空变特性的直观认识。

2-2 / 主要研究思路

特征选择

基于模型的特征选择方法:例如基于树的特征选择以及不同提取方法的对比。

特征提取

结合特征数据的空间特性领域特征、数据空间分布特性对数据的特征进行分析和理解，主要提取的邻域特征、卷积特征、以及基于贝叶斯先验信息条件概率特征。

提升算法的研究

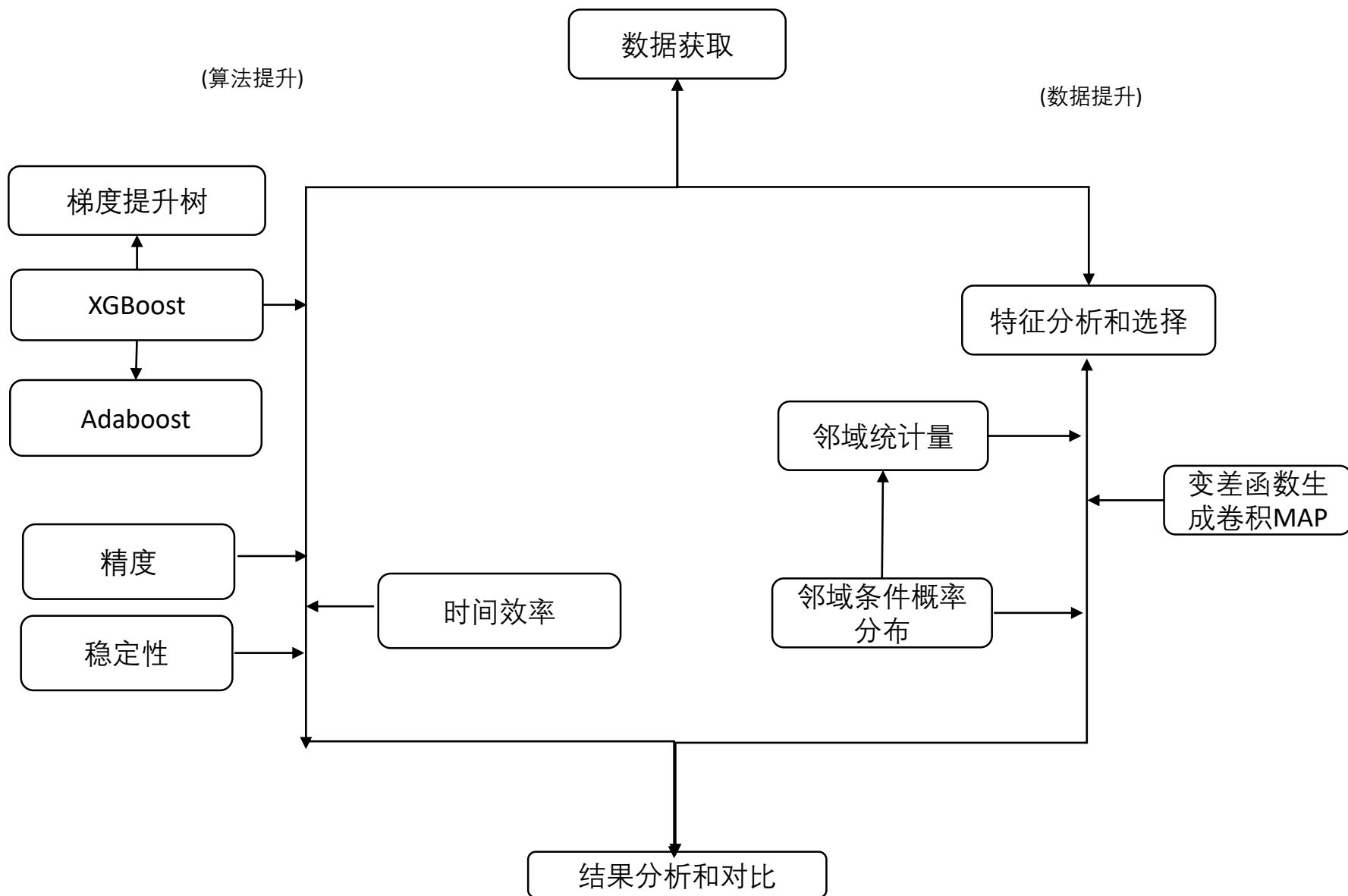
微观上从算法拟合数据空间的过程、方式、损失函数的优化路径、算法在拟合数据过程出现的一些问题而对数据局部梯度的搭建。

宏观上与数据特性进行结合从算法的拟合策略、优化目标、提升目标等方面对提升类算法进行横向对比，纵向上与其他统计学习算法进行对比，从而到达理想的拟合效果。

实验结果分析和对比

完成数据的准确率、kappa检验系数、混淆矩阵、学习矩阵、loss下降趋势、数据本身的效果几个部分进行分析和验证。


2-3 / 研究流程



2-4 / 提升模型

提升(Boosting)方法是一种多模型集成的模型，本质上就是简单基函数的加法模型，以决策树为基函数提升方法称为提升树(Boosting Tree)。提升树算法可以表示为决策树的加法模型：

$$f_M(x) = \sum_{m=1}^M T(x; \Theta_m)$$

$$\hat{\Theta}_m = \arg \min_{\Theta_m} \sum_{i=1}^N L(y_i, f_{m-1}(x) + T(x; \Theta_m)) \longrightarrow f_m(x) = f_{m-1}(x) + T(x; \Theta_m)$$


2-5 / GBDT (梯度提升算法)

梯度提升决策树(Gradient Boosting Decision Tree)算法是近几年来被应用比较频繁的模型, 这主要得益于其在数据挖掘竞赛中的优秀表现

The diagram illustrates the iterative process of the Gradient Boosting Decision Tree (GBDT) algorithm. It consists of four mathematical expressions arranged in a cycle, connected by arrows indicating the flow of the algorithm:

- Top Equation:**
$$f_M(x) = \sum_{m=1}^M T(x; \Theta_m)$$
- Right Equation:**
$$f_m(x) = f_{m-1}(x) + T(x; \Theta_m)$$
- Bottom Equation:**
$$\theta^t = \sum_{t=0}^T -\alpha_t L'(\theta^{t-1})$$
- Left Equation:**
$$\hat{\Theta}_m = \arg \min_{\Theta_m} \sum_{i=1}^N L(y_i, f_{m-1}(x) + T(x; \Theta_m))$$

Arrows indicate the flow of the algorithm: from the top equation to the right equation, from the right equation to the bottom equation, from the bottom equation to the left equation, and from the left equation back to the top equation.

2-6 / XGBoost(eXtreme Gradient Boosting)

XGBoost (eXtreme Gradient Boosting)是一个托管在Github上的开源软件框架，其算法借鉴了GBDT的基本思想，并对拟合误差进行泰勒公式二阶极端展开，同时支持分布式的高性能计算，是一种优秀的提升算法，广泛的应用于数据挖掘竞赛领域

$$f_M(x) = \sum_{m=1}^M T(x; \Theta_m)$$

$$\begin{aligned} Obj^{(t)} &= \sum_{i=1}^n l(y_i, \hat{y}_i^{(t)}) + \sum_{i=1}^t \Omega(f_i) \\ &= \sum_{i=1}^n l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) + c \end{aligned}$$

正则化参数 $\Omega(f_t) = \gamma T + \lambda \sum_{j=1}^T w_j^2$

$$\Delta \theta = -\frac{g}{h}$$

$$Obj^{(t)} \approx \sum_{i=1}^n \left[l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) + c$$

$$= \sum_{i=1}^n \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) + c \quad \text{二阶精度展开}$$

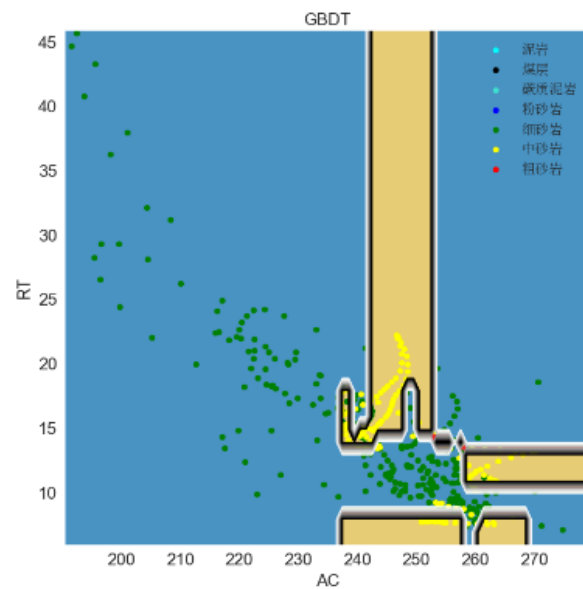
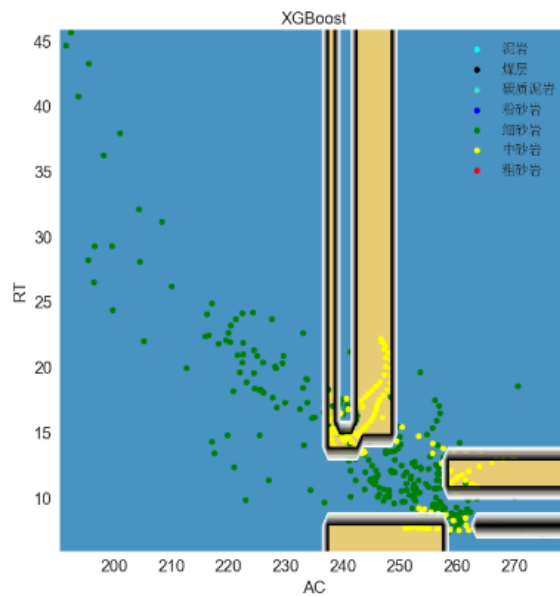
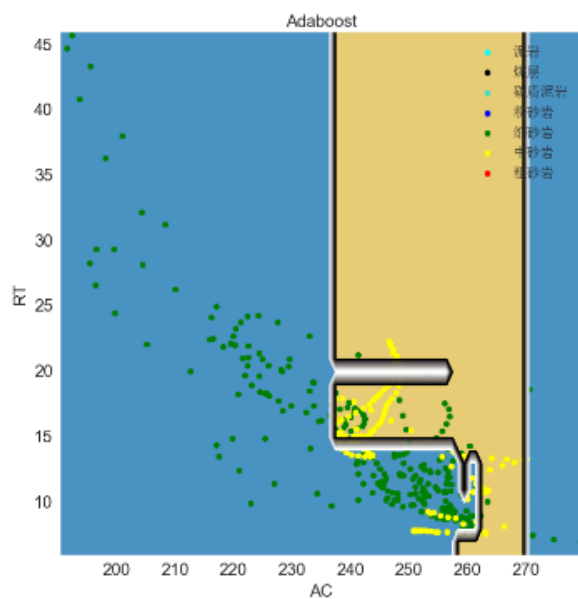
$$f_m(x) = f_{m-1}(x) + T(x; \Theta_m)$$

2-7 / 岩性分类实验效果

2000余条测井曲线，6个测井变量

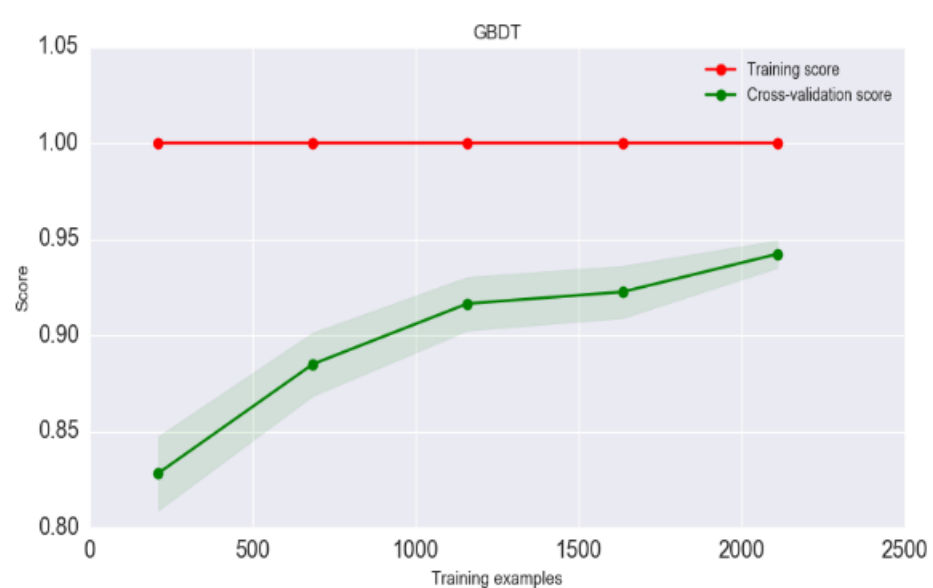
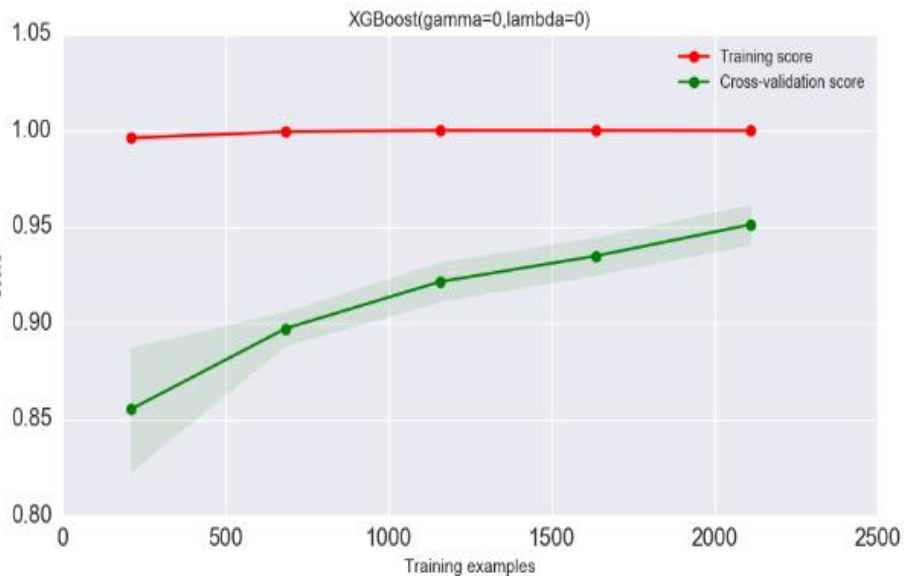
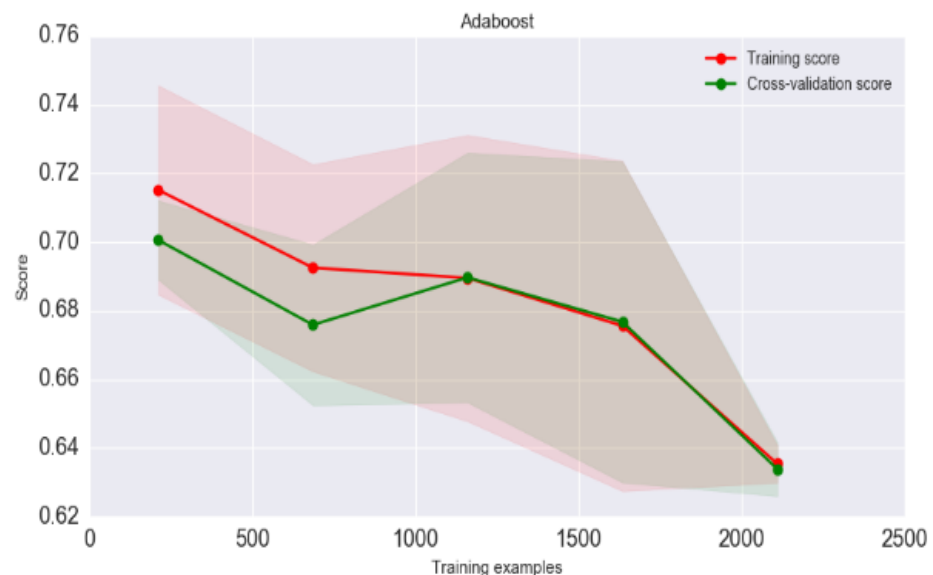
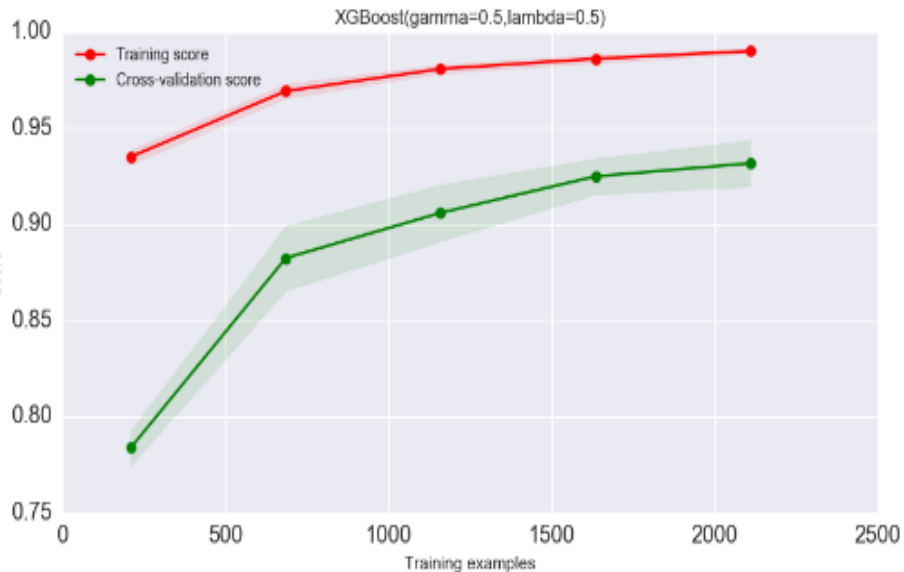
	泥岩	煤层	碳质泥岩	粉砂岩	细砂岩	中砂岩	粗砂岩	含砾粗砂岩	recalls
泥岩	149	0	0	0	0	0	0	0	1
煤层	0	23	0	0	0	0	0	0	1
碳质泥岩	0	0	42	0	0	0	0	0	1
粉砂岩	0	0	0	53	0	0	0	0	1
细砂岩	0	0	0	0	58	5	0	2	0.892308
中砂岩	0	0	0	0	0	34	0	0	1
粗砂岩	0	0	0	0	0	0	66	12	0.846154
含砾粗砂岩	0	0	0	0	3	2	7	204	0.944444
precisions	1	1	1	1	0.95082	0.829268	0.90411	0.93578	0.95303

2-9 / 提升算法的对比

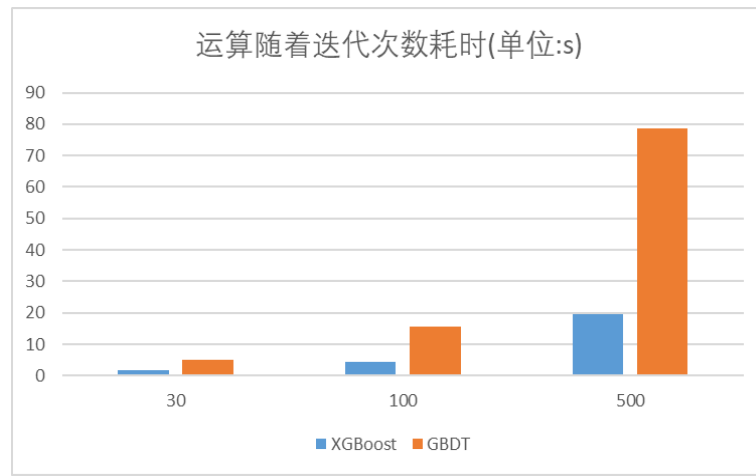
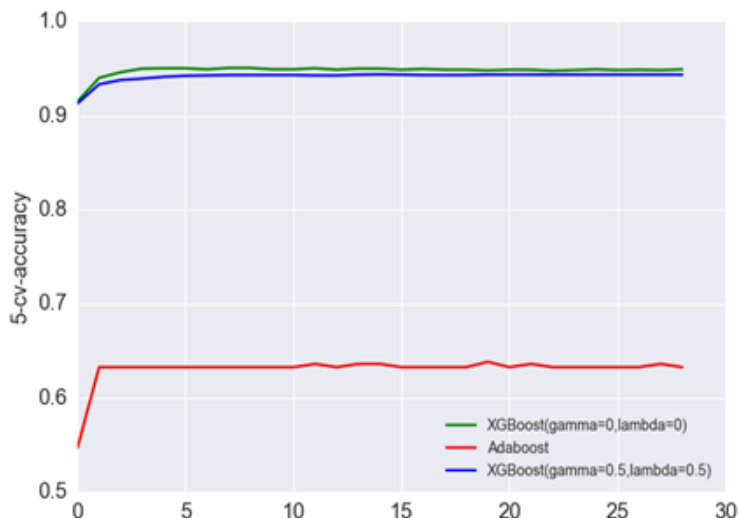
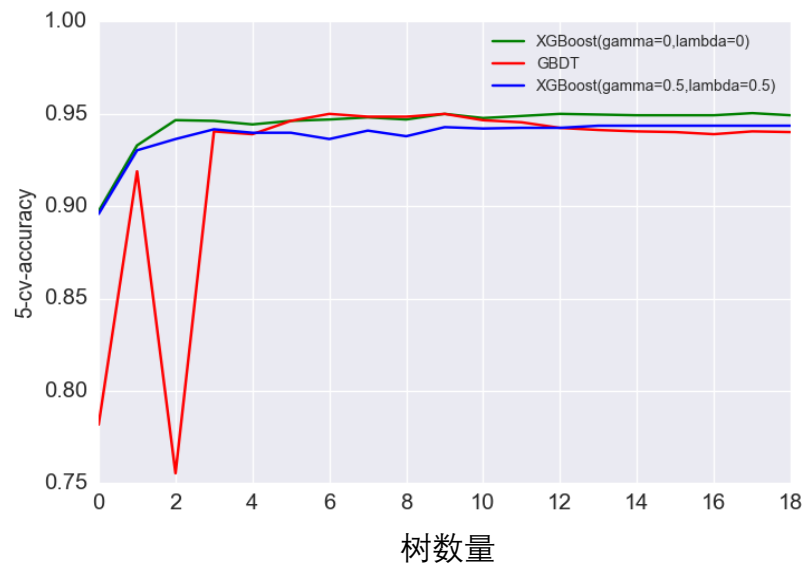
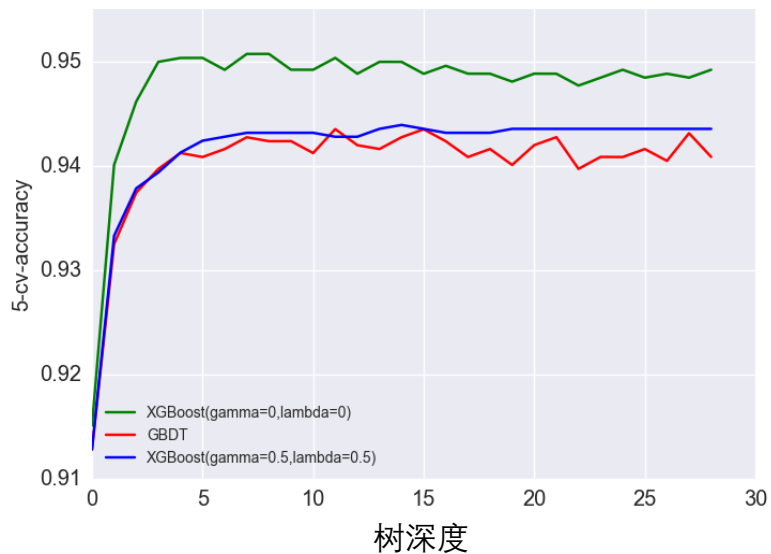


两种有代表性的类别和变量

2-8 / 学习曲线



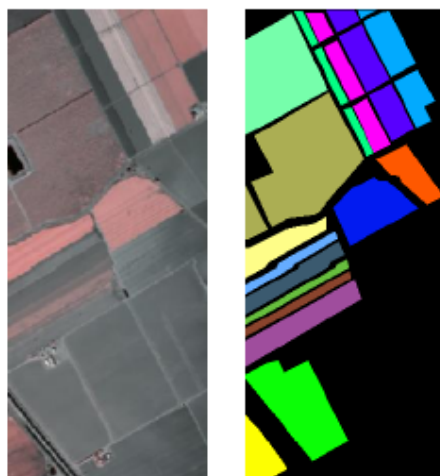
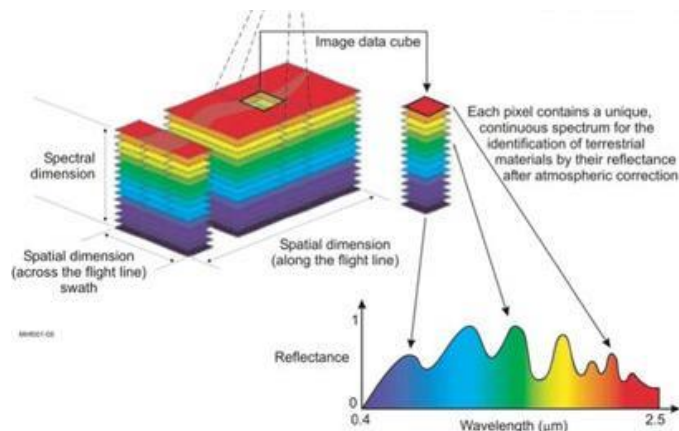
2-9 / 稳定性对比



3-1 / Salinas

数据特征：224个波段→24 1：1的训练和测试样本

数据类别



(a)

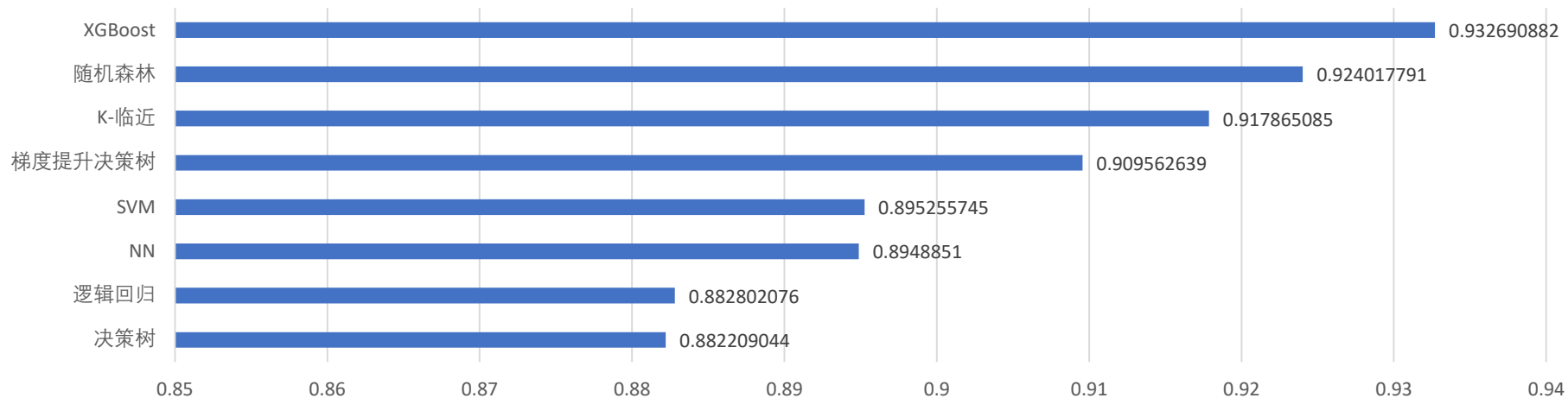
(b)

图2.3 Salinas 图像

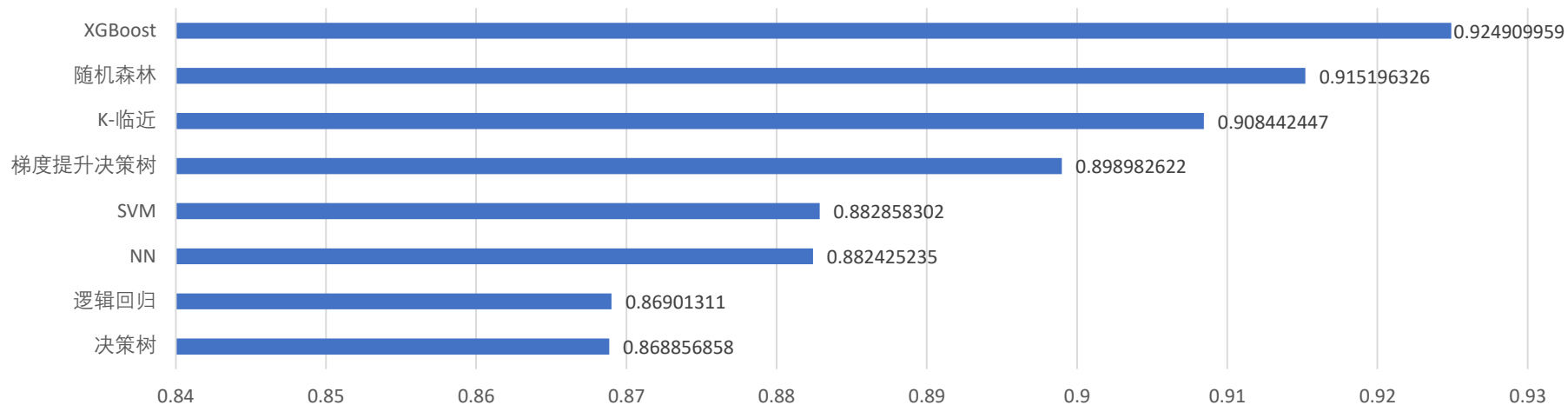
类别	类名	颜色	每类样本数
1	Brocoli_green_weeds_1	黄色	2009
2	Brocoli_green_weeds_22	蓝色	3726
3	Fallow	红色	1976
4	Fallow_rough_plow	青色	1394
5	Fallow_smooth	品红色	2678
6	Stubble	深蓝色	3959
7	Celery	浅蓝色	3579
8	Grapes_untrained	棕色	11271
9	Soil_vinyard_develop	亮绿色	6203
10	Corn_senesced_green_weeds	紫色	3278
11	Lettuce_romaine_4wk	天蓝色	1068
12	Lettuce_romaine_5wk	灰蓝色	1927
13	Lettuce_romaine_6wk	浅绿色	916
14	Lettuce_romaine_7wk	深棕色	1070
15	Vinyard_untrained	青色	7268
16	Vinyard_vertical_trellis	亮黄色	1807

3-2 / 数据的初步识别情况

原始特征准确率

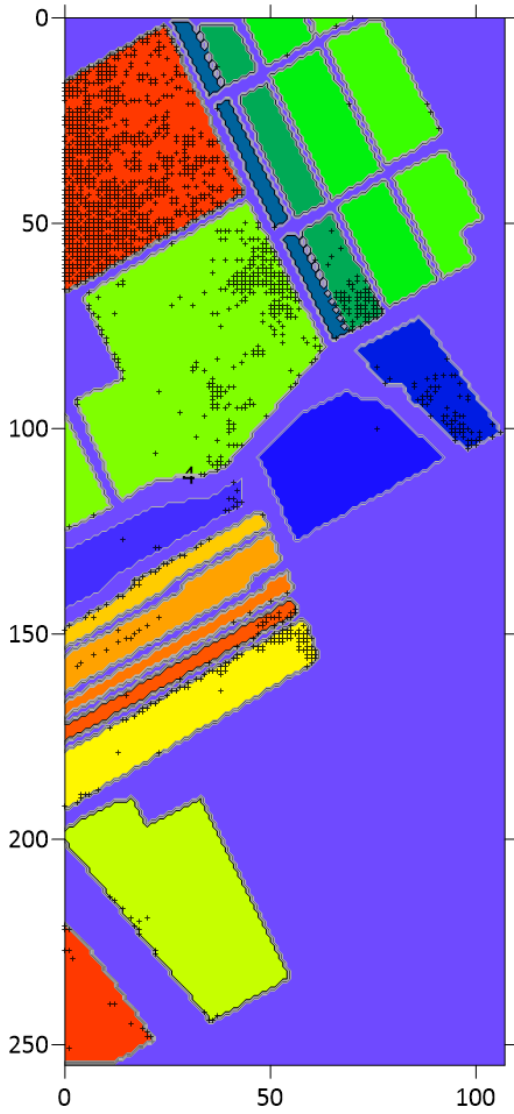


kappa

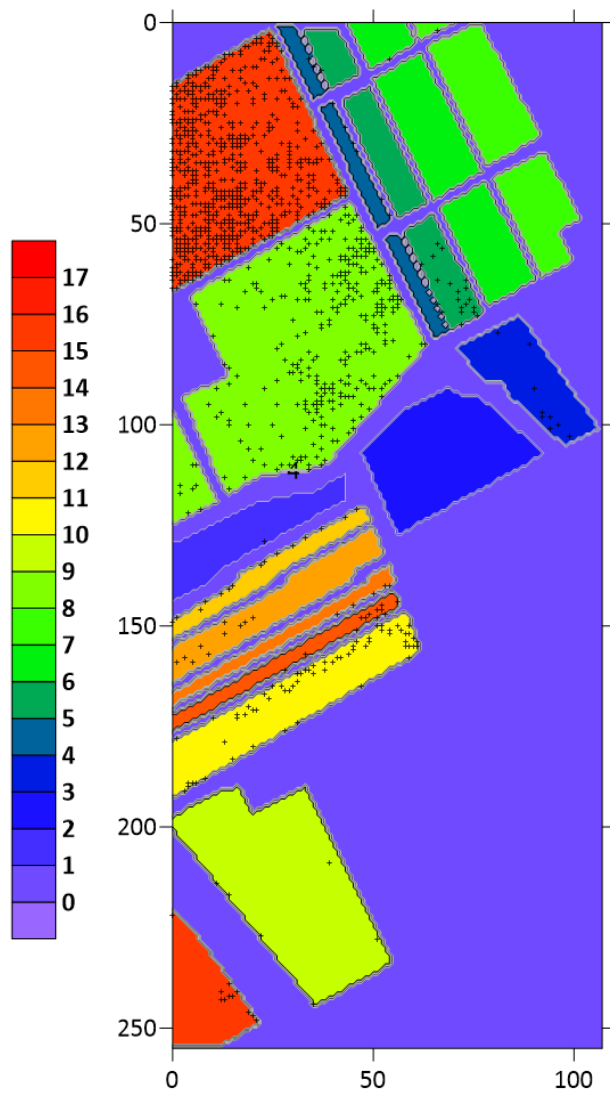


3-3 / 各方法对比

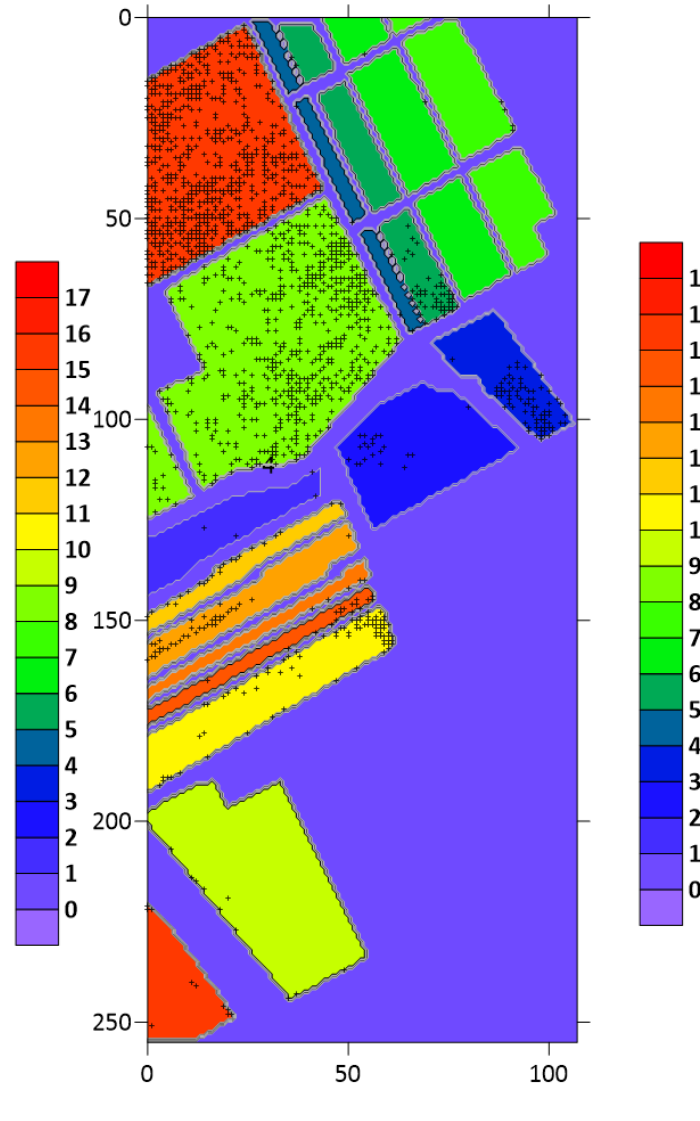
SVM



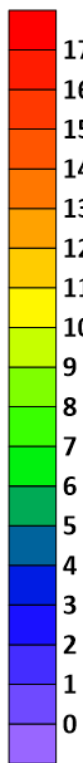
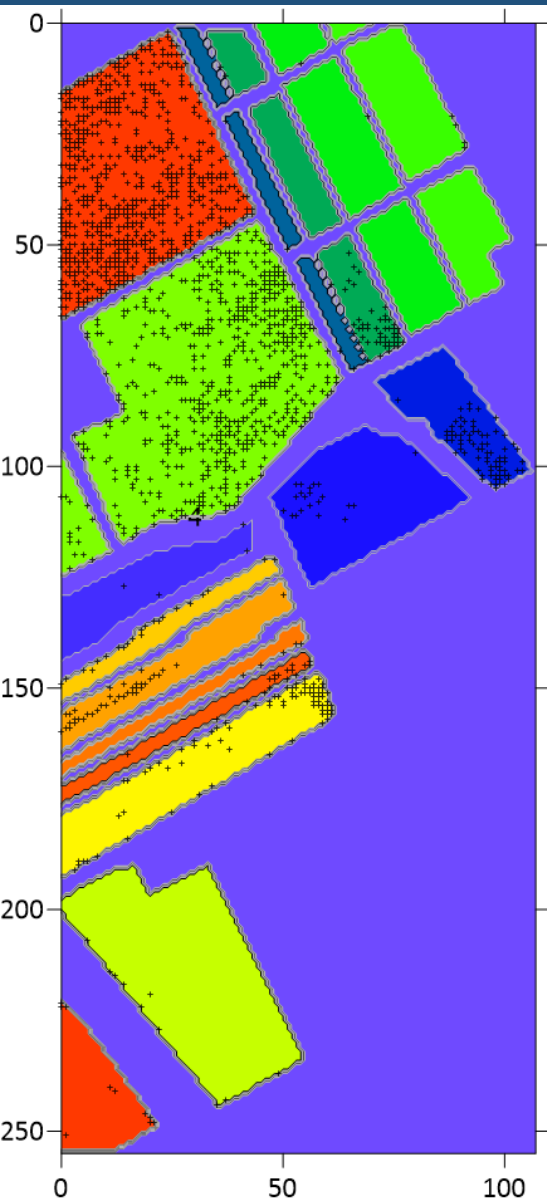
随机森林



KNN



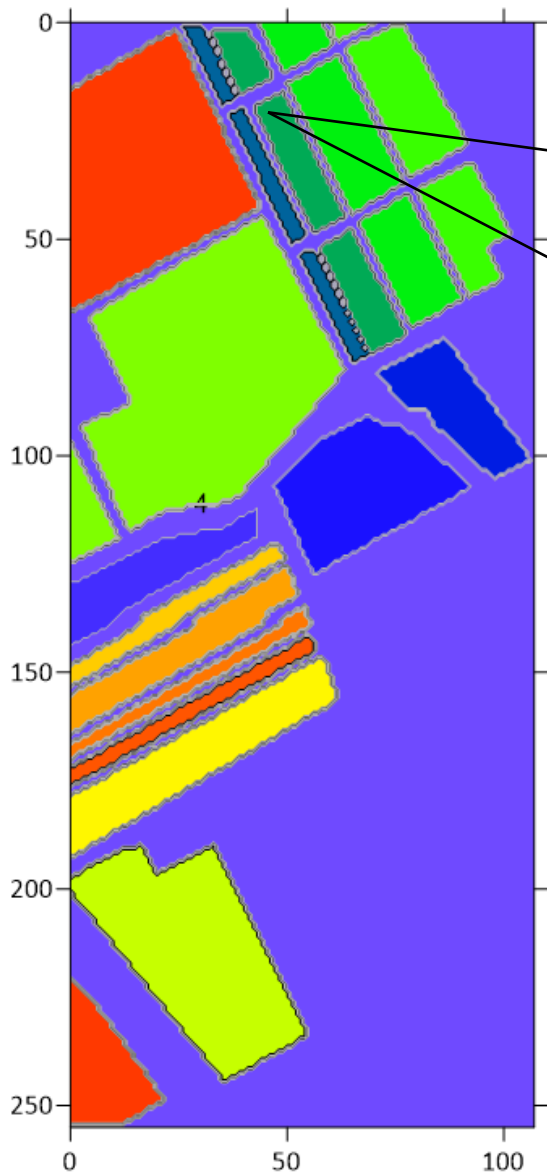
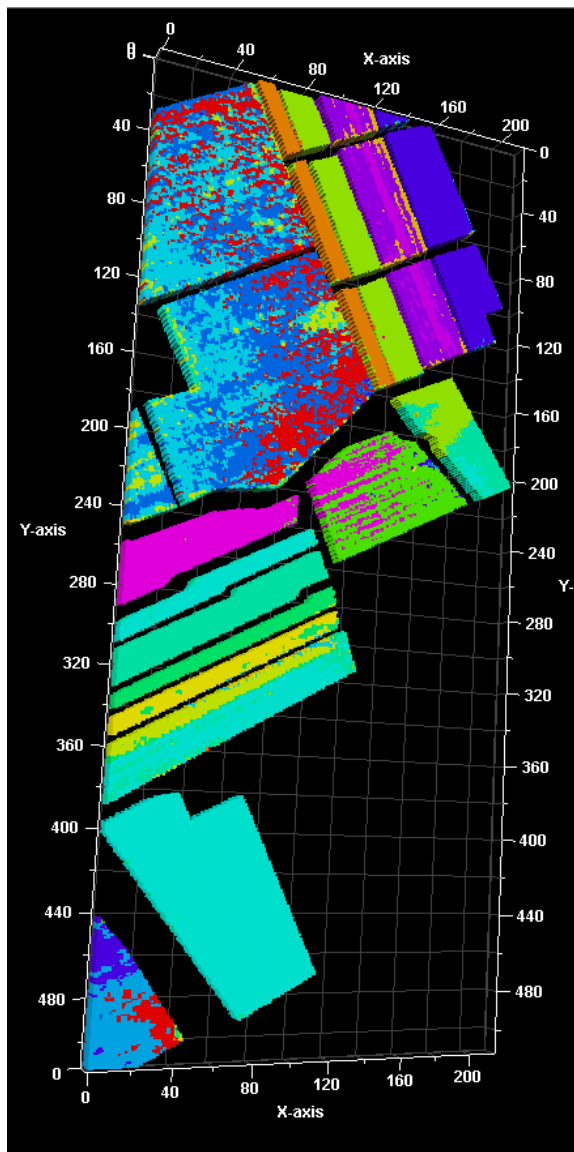
3-4 / XGBoost提升树模型



XGBOOST交叉矩阵

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	504	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	34	897	0	0	0	0	0	0	0	0	0	0	0	1	0	0
3	0	0	484	1	6	0	0	0	0	2	0	0	0	0	0	0
4	0	0	0	341	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	3	5	652	0	0	0	0	1	0	0	0	0	1	0
6	0	0	0	0	1	992	0	0	0	0	0	0	0	1	0	0
7	0	0	0	0	0	0	895	1	0	0	0	0	0	1	0	0
8	0	0	0	0	0	0	0	2577	0	12	0	0	0	1	219	0
9	0	0	0	0	0	0	0	4	1549	2	0	0	0	0	0	0
10	0	0	18	0	0	0	0	4	7	773	5	8	0	0	0	0
11	0	0	3	0	1	0	0	0	4	2	257	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	480	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	225	3	0	0
14	0	0	0	0	0	0	0	2	0	1	0	0	3	261	1	0
15	0	0	1	0	0	0	0	538	0	7	0	0	0	0	1258	0
16	0	0	0	0	0	0	0	1	0	1	0	0	0	0	2	437

3-5 / 邻域统计思想

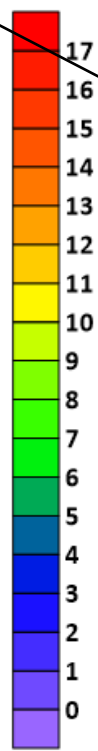


MEAN

用 $n*n$ 的邻域
均值来作为地
理位置信息的
邻域特征

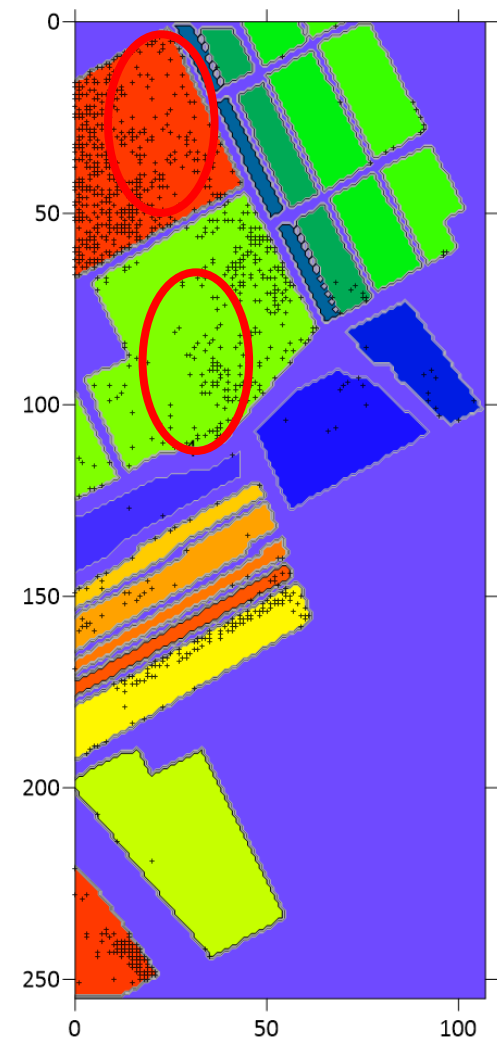
VAR

用 $n*n$ 的邻域
方差来作为地
理位置信息邻
域粗糙程度

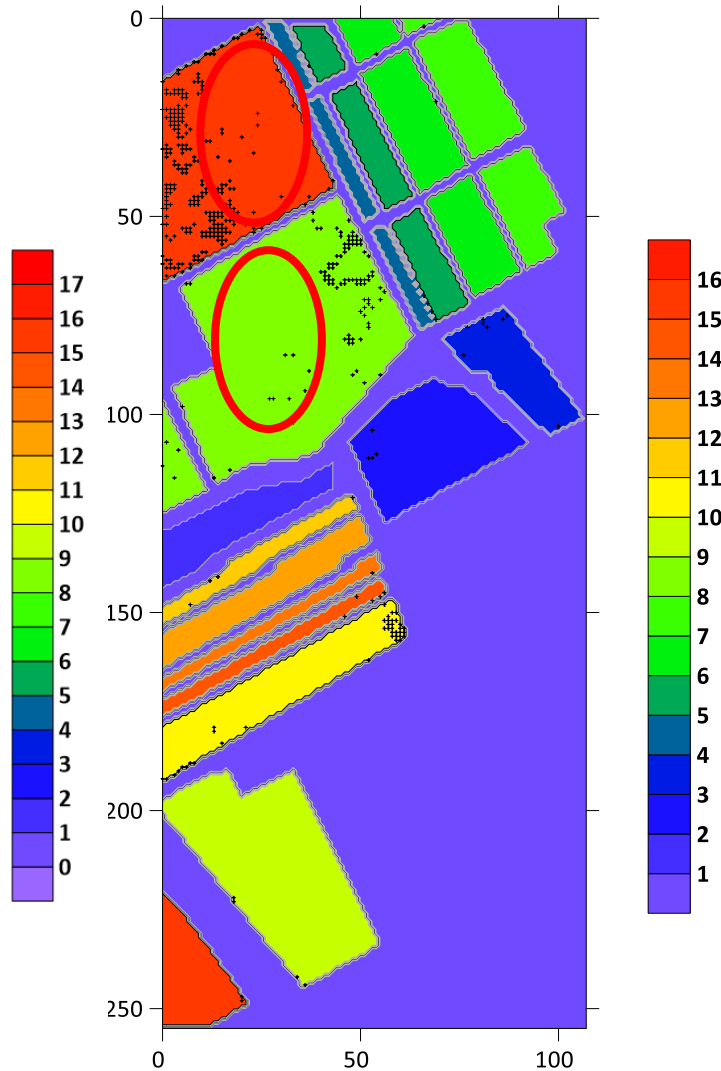


3-7 / 均值特征+XGBOOST

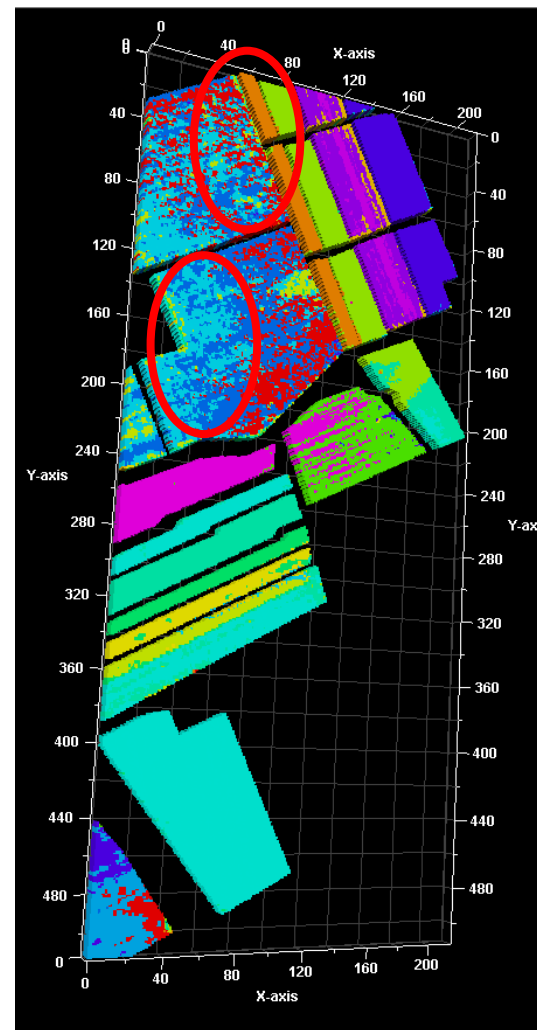
原始特征效果



邻域概率特征效果

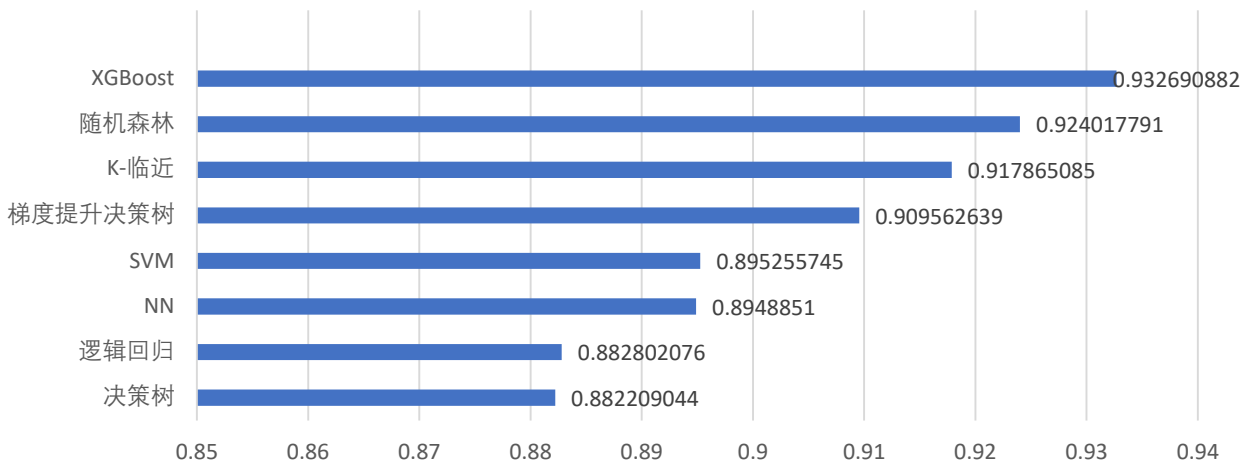


BP聚类效果

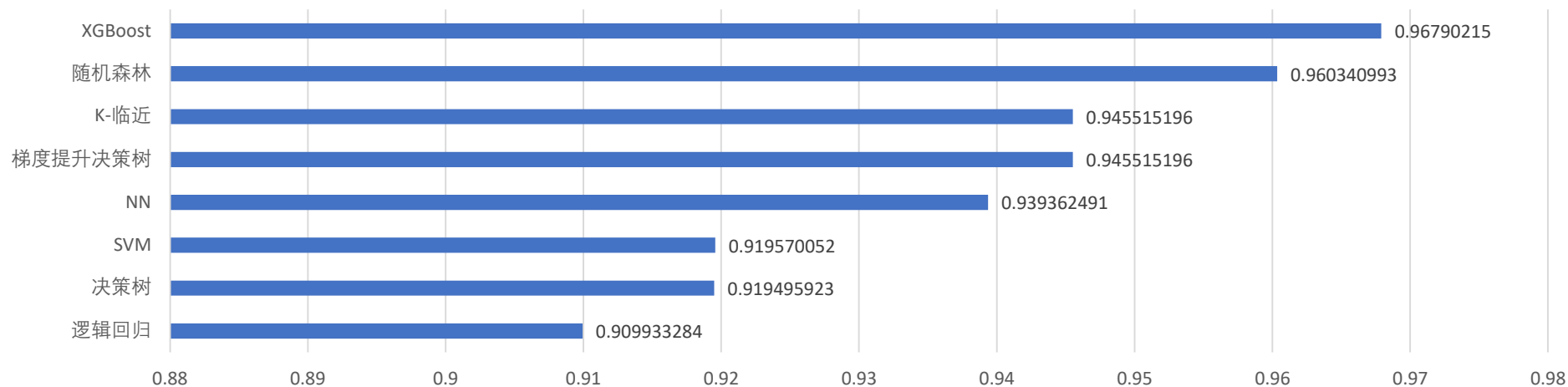


3-8 / 均值特征

原始特征



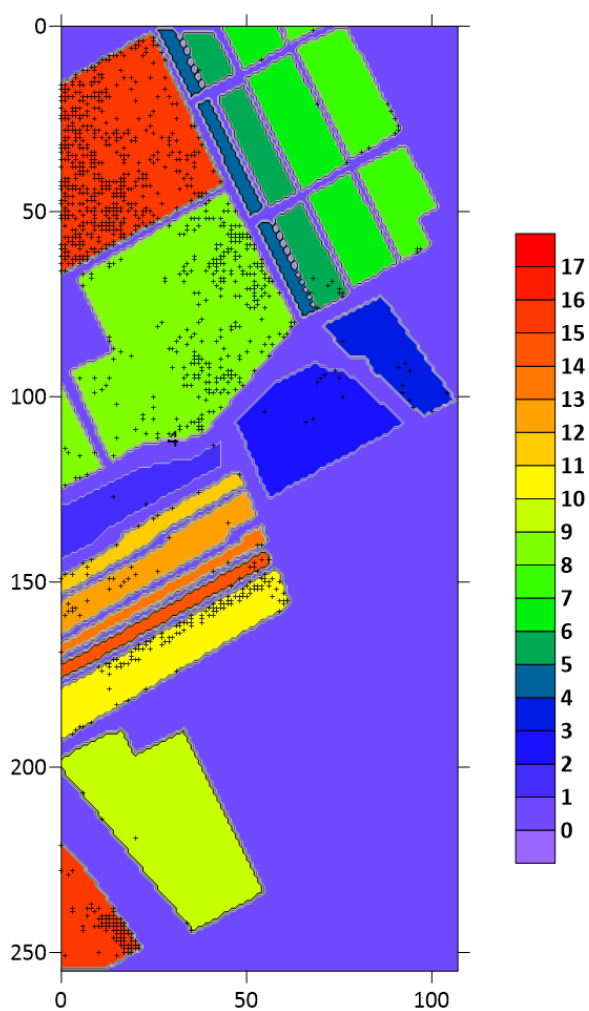
均值特征



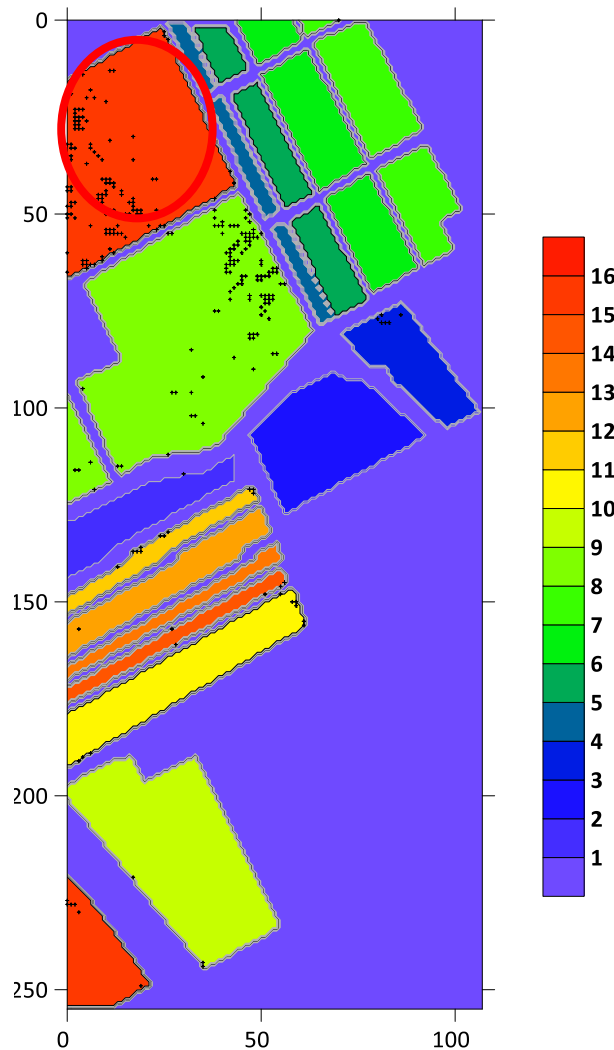
提升了3个百分点

3-9 / 均值特征+方差+XGBoost

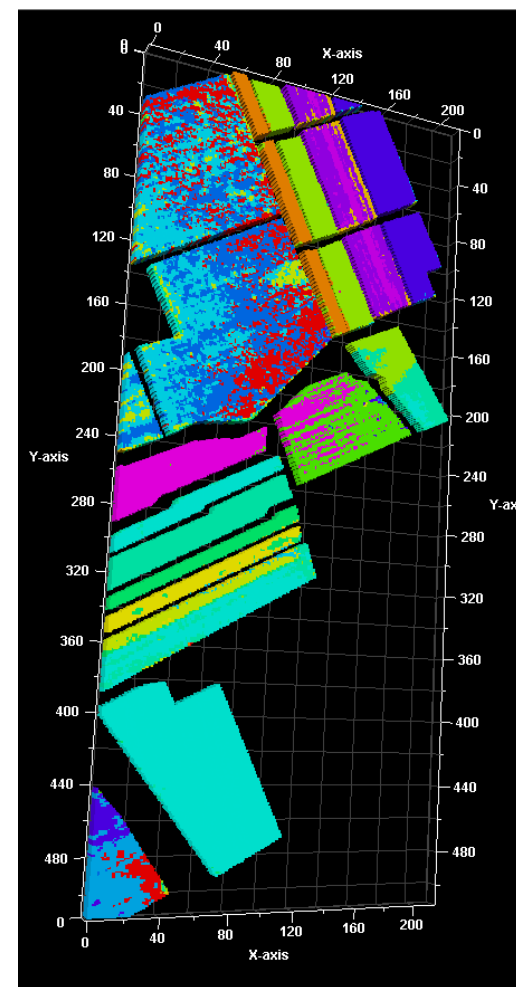
原始特征效果



邻域均值方差特征效果

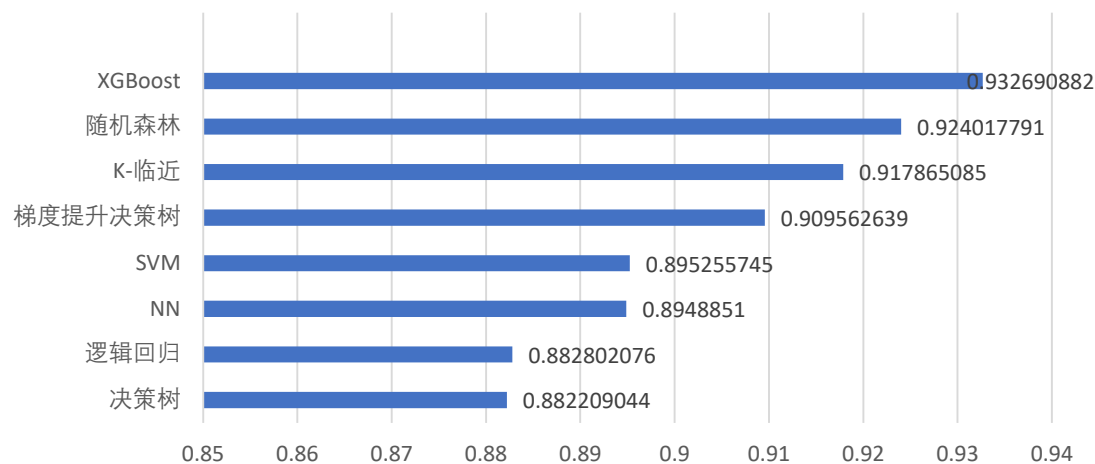


BP聚类效果

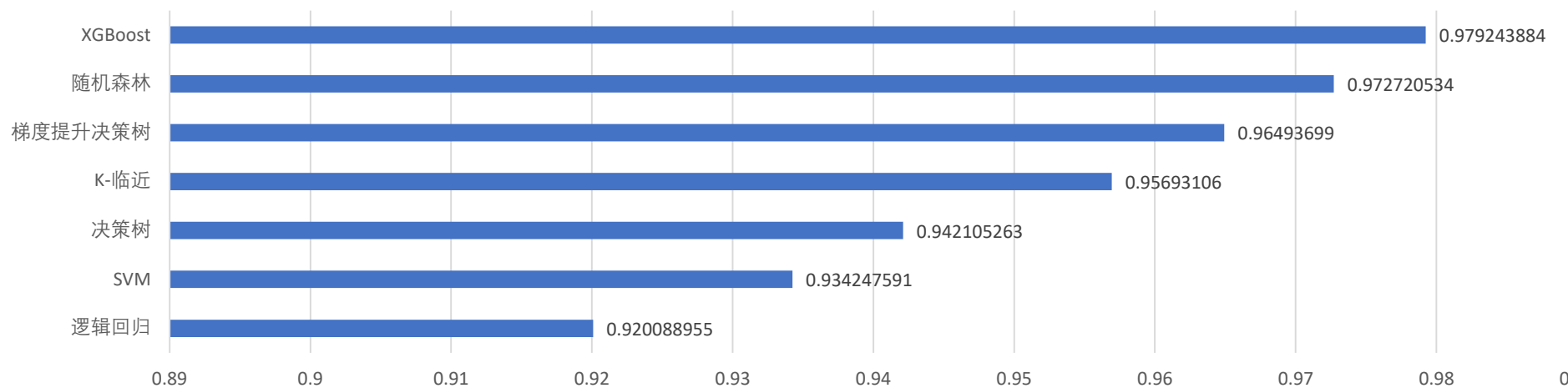


3-10 / 邻域均值+方差

原始特征

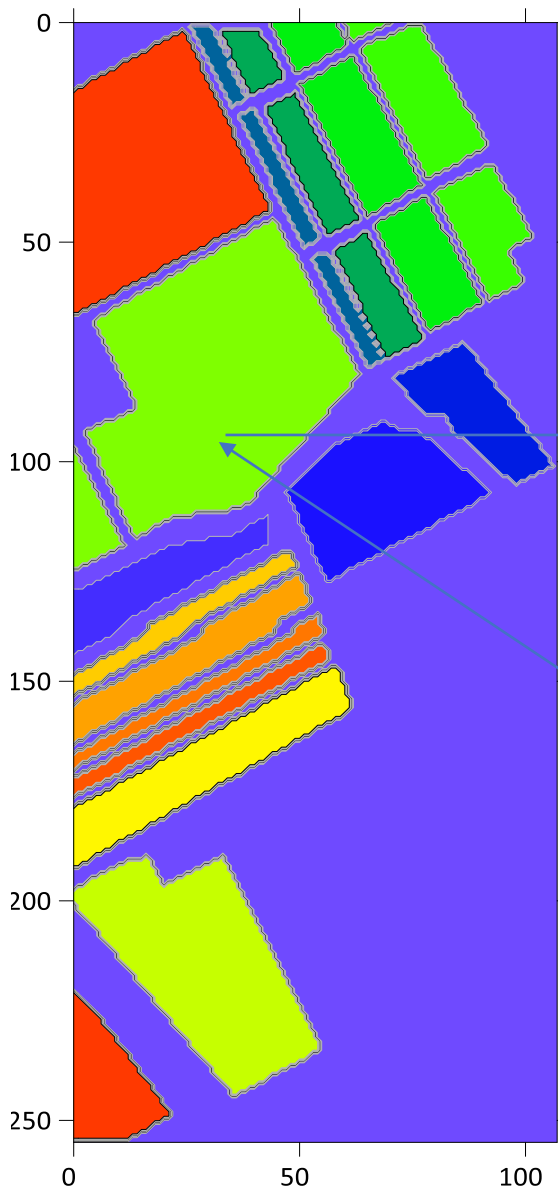


均值+方差

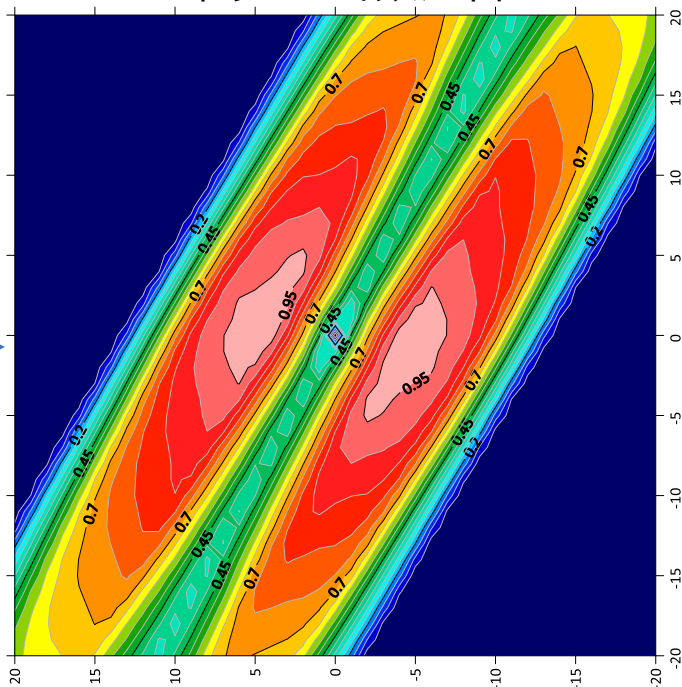


提升了4.6个百分点

3-12 / 特征卷积



二维变差函数矩阵



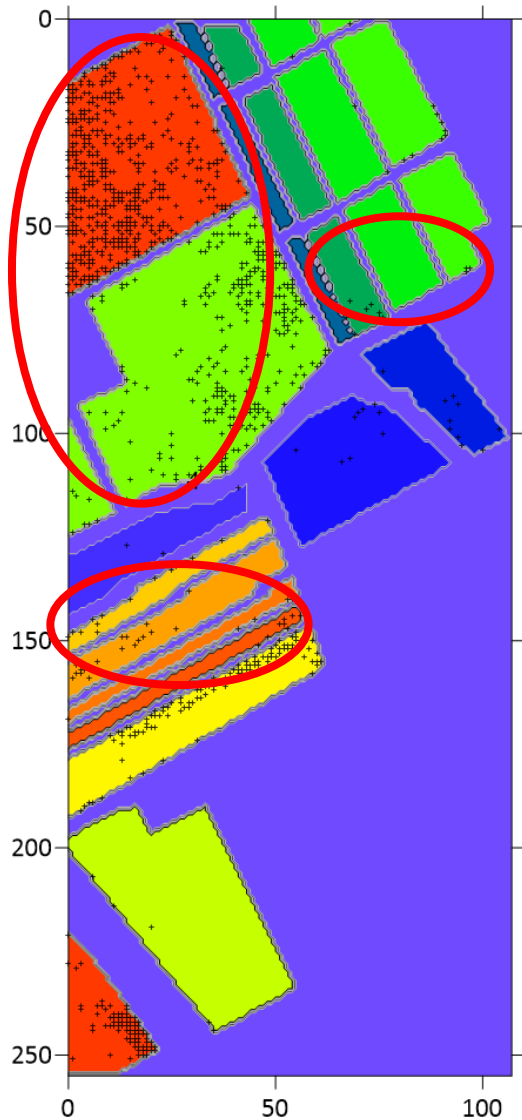
提取出了数据变化的方向和
数据变化的敏感度

$N \times N$

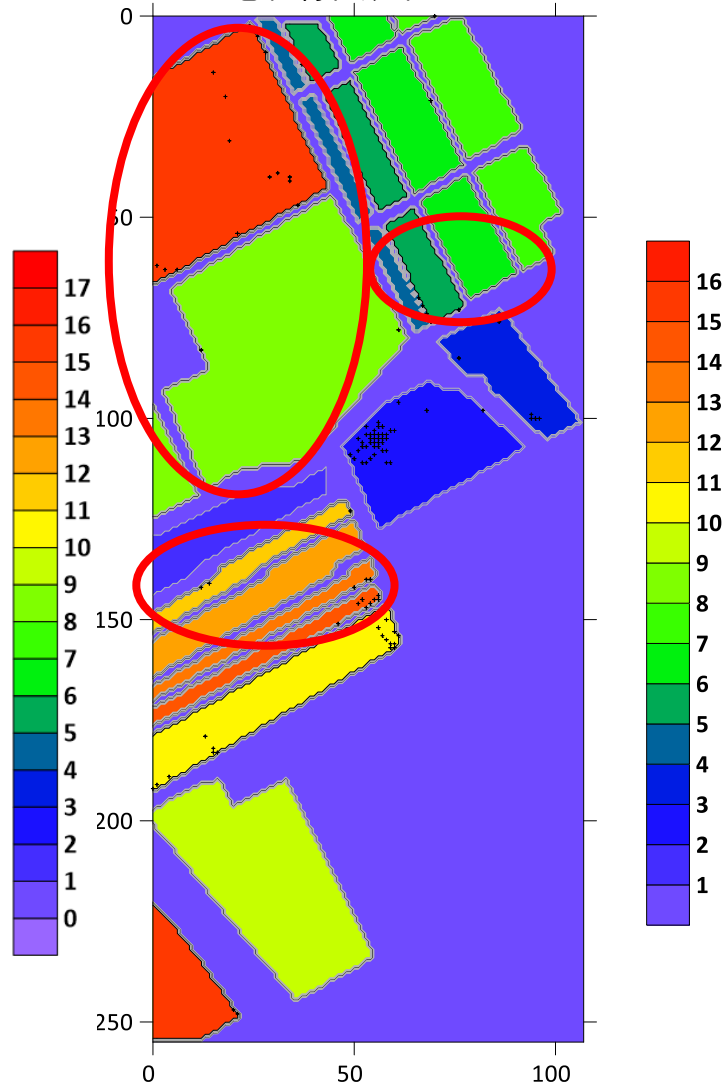
用 $N \times N$ 的矩阵去卷积每个波段

3-13 / 卷积特征+XGBOOST

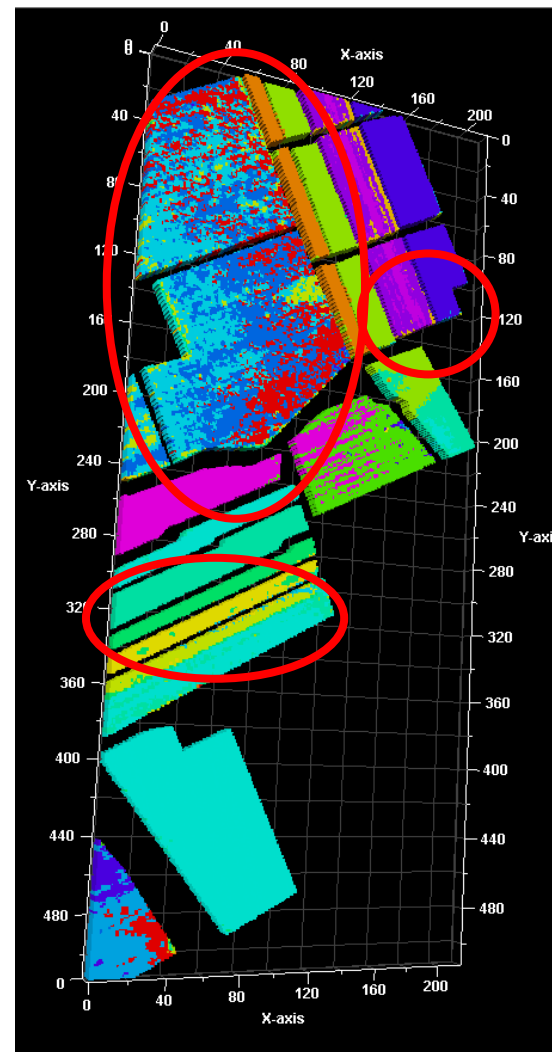
原始特征效果



卷积特征效果

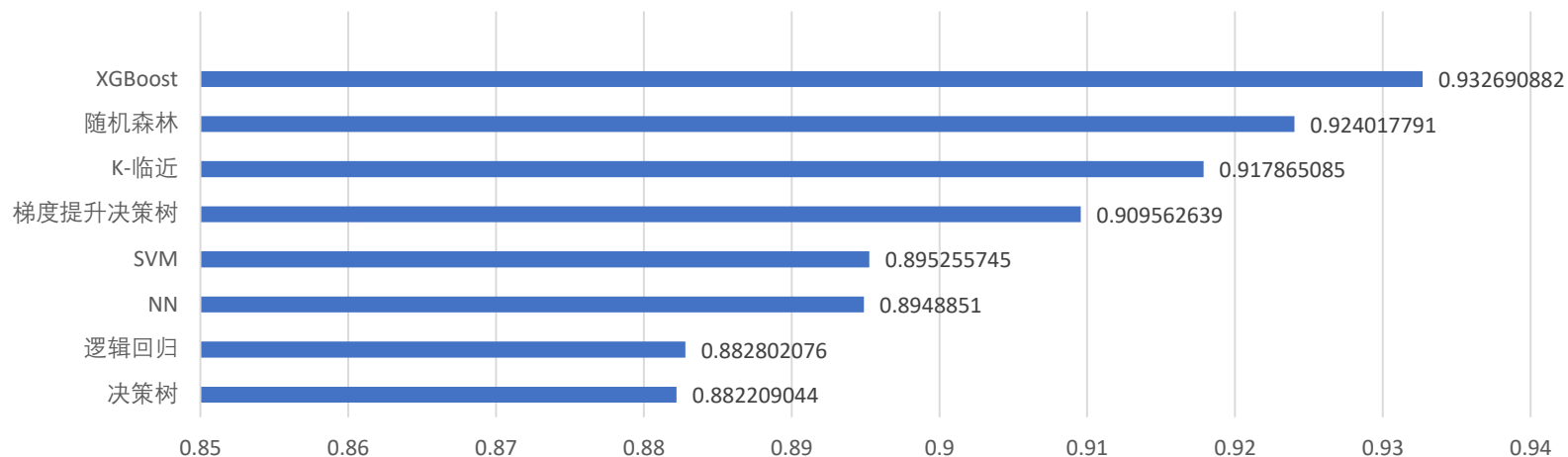


BP聚类效果

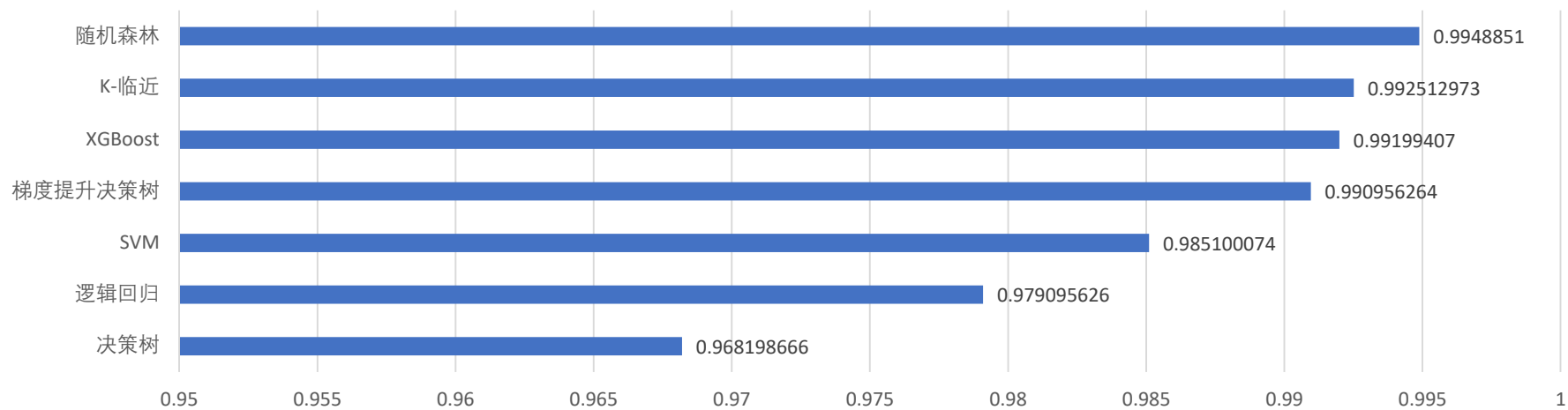


3-13 / 结果对比

原始特征

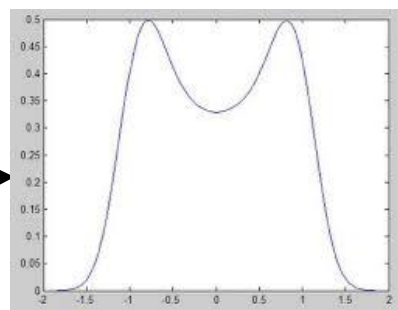
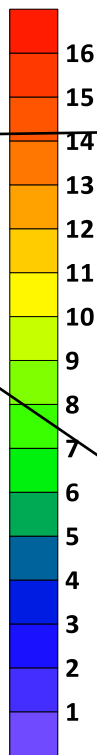
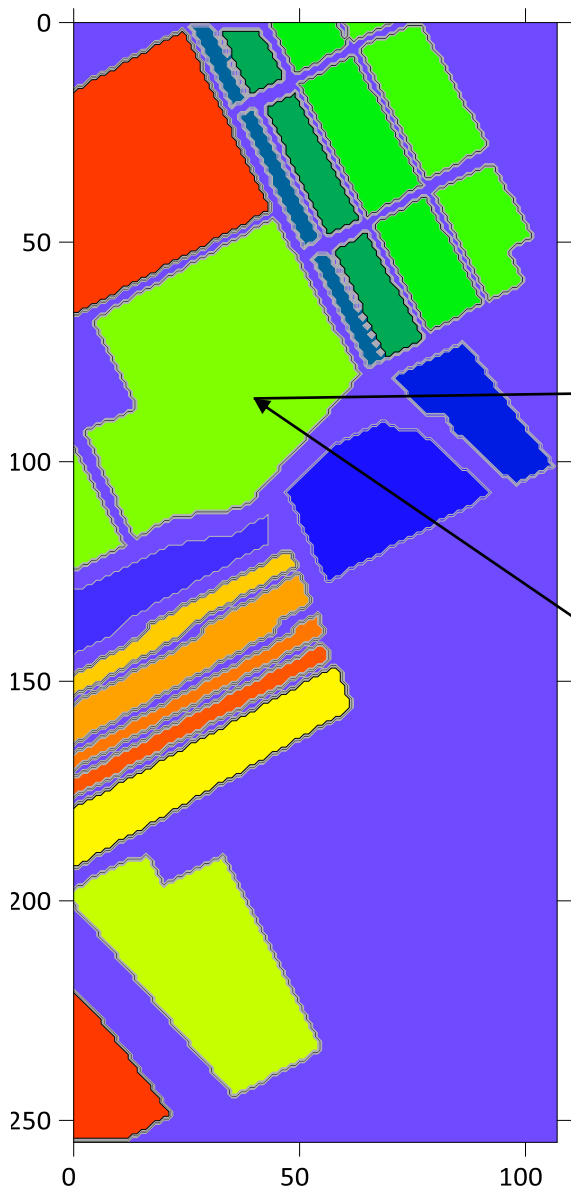


卷积特征

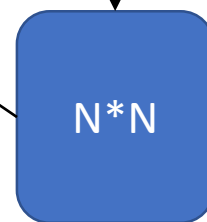


提升了6个百分点

3-14 / 邻域条件概率



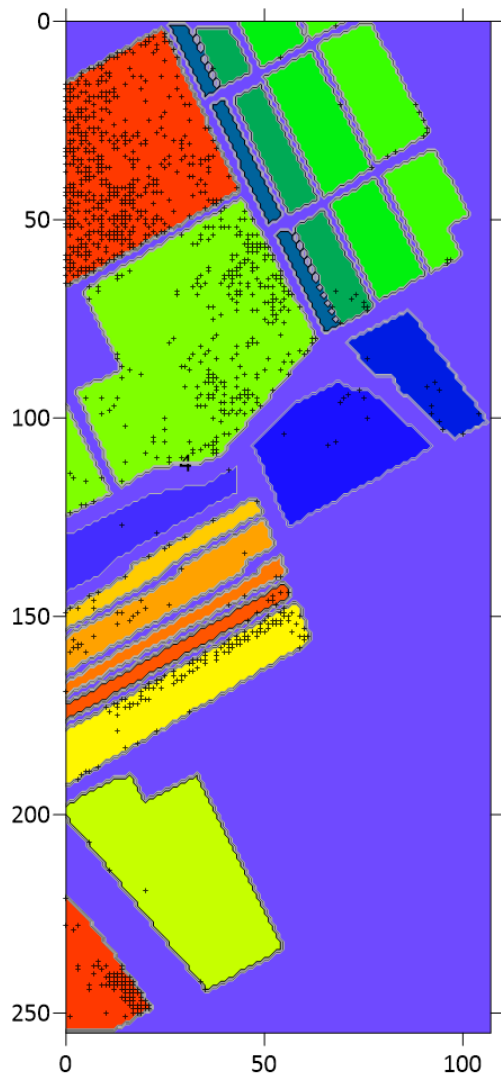
用邻域的概率分布来代替均值和方差的统计信息来进行卷积



用 $N*N$ 的矩阵去卷积每个波段

3-15 / 邻域条件概率效果

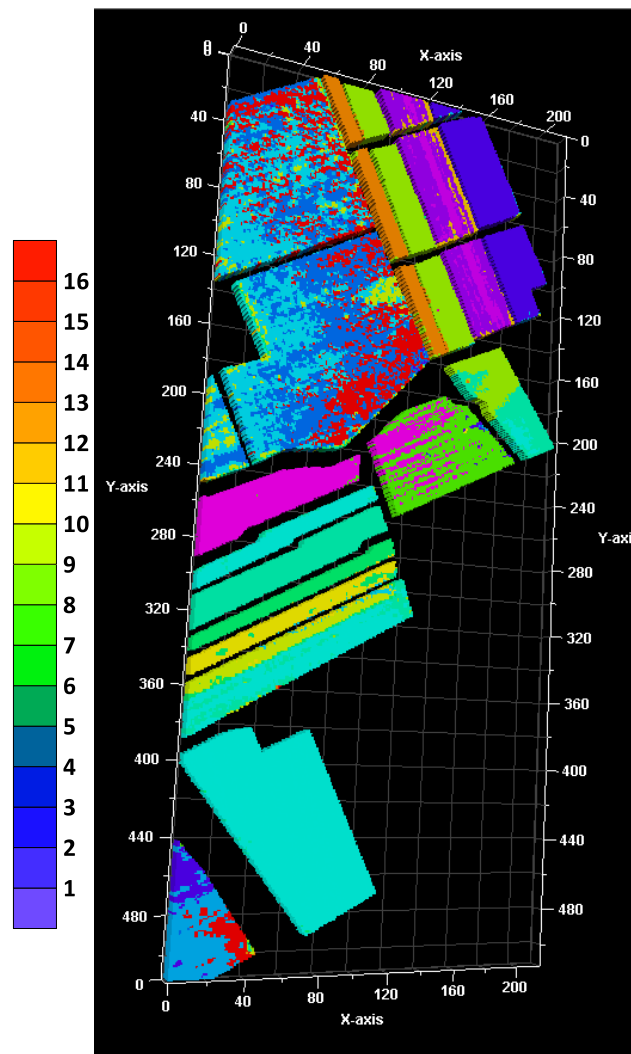
原始特征效果



邻域概率特征效果

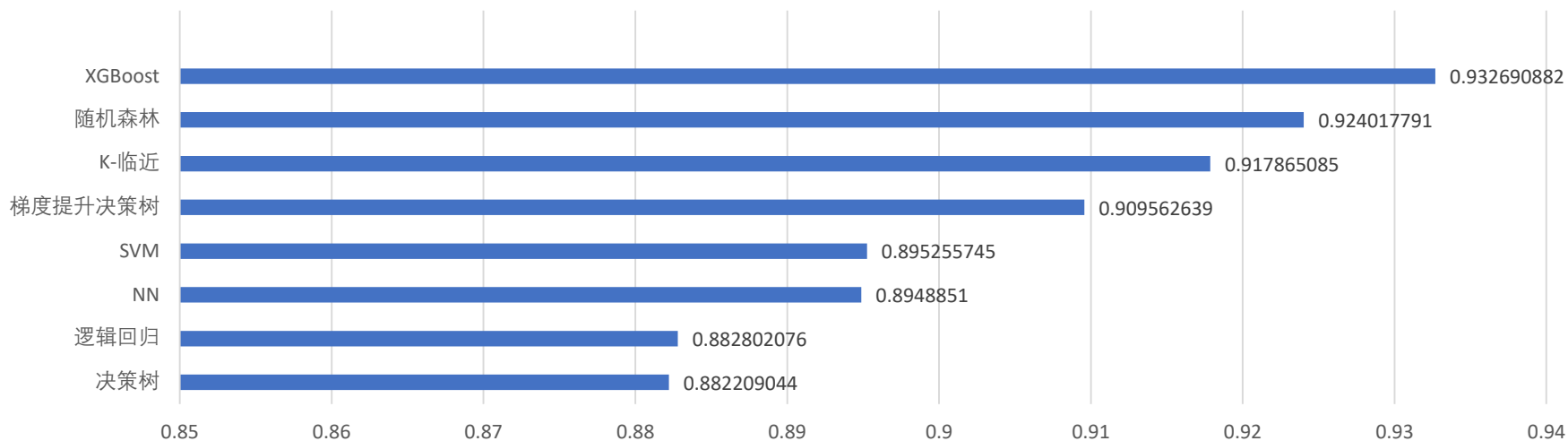


BP聚类效果

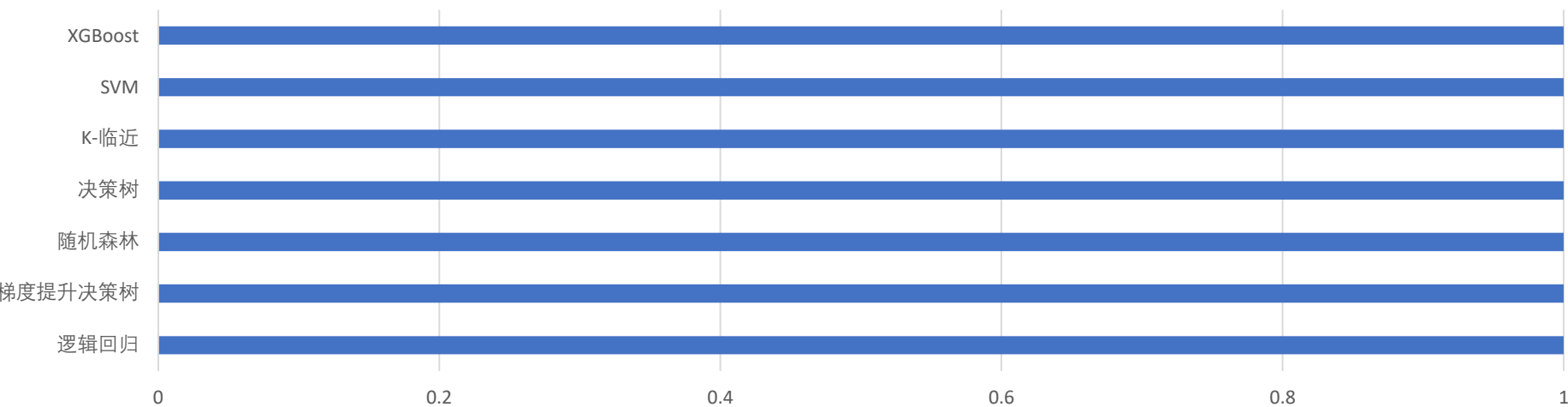


3-16 / 邻域条件概率效果

原始特征



条件概率特征



实现了数据的完全分离

4-1 / 进阶研究工作



算法方面研究

空间的划分、优化方法的选择、效率方面的测试、传统机器学习算法与深度学习技术的结合



数据方面的研究

算法在拟合数据的过程中岩性数据边缘跨度较大、梯度下降不平稳的问题，
利用对抗生成网络进行梯度搭建的问题

5-2 / 预期目标和成果



预期目标

本项目基于空间结构对地质类数据和遥感类数据进行特征挖掘和分析,应用计算机技术、模式识别技术。通过空间邻域的分析、可以对特征进行有效的挖掘;通过基于集成学习的提升树方法的极限拟合可以将现阶段无法完成的识别工作进行进一步的提升。

6-1 / 时间安排

2018.3月10日-11月20日：确定论文题目，查阅、整理参考文献资料，确定研究背景，制定研究方案，安排论文进度，制定开题报告，送交指导教师审核。

2018.11月20日-2018.12月31日：阅读和整理参考文献。

2019.1月1日-2019.4月27日：对各种机器学习算法进行分析和研究，完成论文初稿。

2019.4月28日-2019.5月9日：在老师的指导下，对论文进行多次修改。

2019.5月10日-2019.5月20日：论文最终稿确定，由指导教师审核通过并打印。

THANKS

请各位老师批评指正！
