

欺诈检测中的自动特征选择

The background features a complex network of nodes and edges. A prominent structure is a cube-like grid of larger blue nodes connected by solid lines, centered in the image. Surrounding this are numerous smaller, fainter nodes and edges, some in light blue and others in a pale pinkish-purple, creating a sense of depth and connectivity. The overall color palette is dark blue and black, with the network elements providing a high-tech, digital aesthetic.

面对海量数据不知如何是好
传统的基于规则的金融模型

设备数据

Mac地址,ip,基站

转账记录

转账时间,身份,金额

活动动轨迹

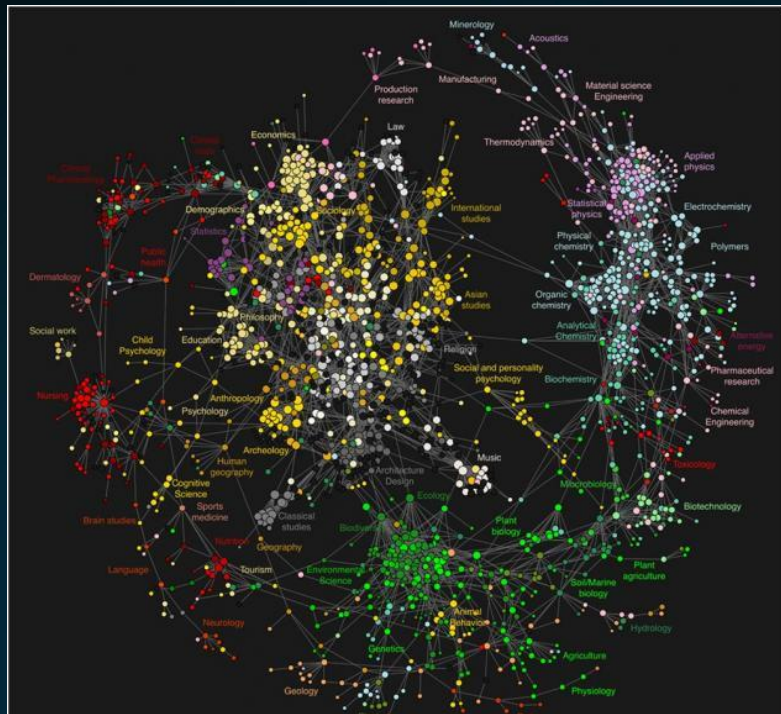
生物特征(输入密码节奏),

电商数据

点击,下单,加入购物车,收藏

关系数据

同人,好友,接近程度...



量大

50亿个点,8000亿条边

异构

每个点有不同的属性

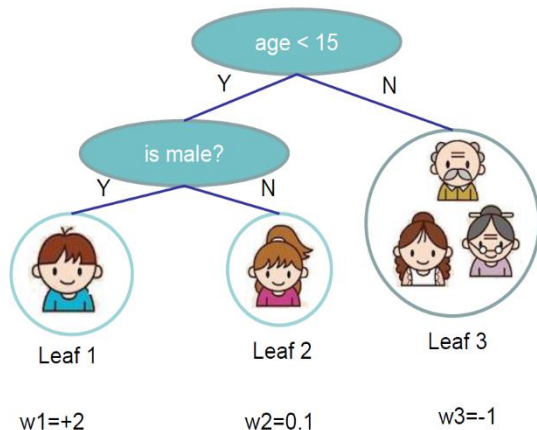
样本少

负样本占比一般在1%以下

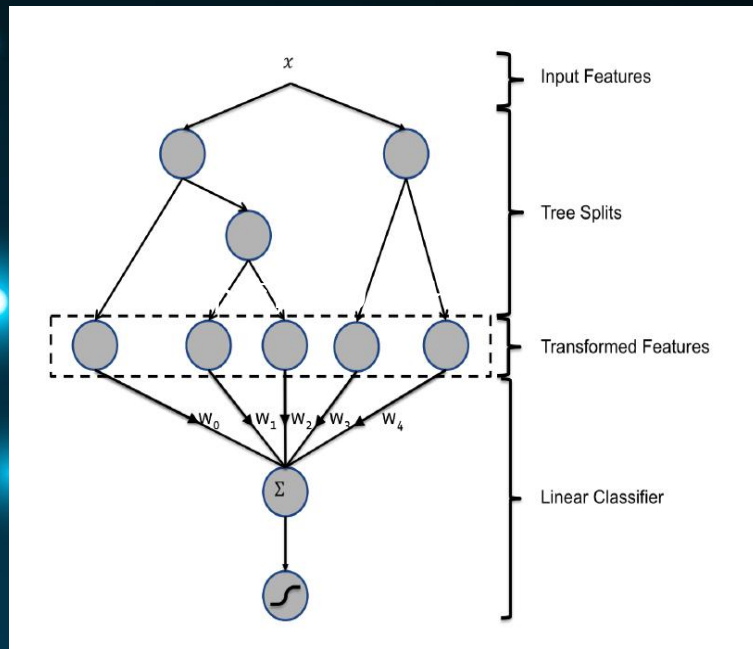
- Define complexity as (this is not the only possible definition)

$$\Omega(f_t) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T w_j^2$$

Number of leaves L2 norm of leaf scores



$$\Omega = \gamma 3 + \frac{1}{2} \lambda (4 + 0.01 + 1)$$



第一阶段:借鉴广告,gbdt数结构作为维度

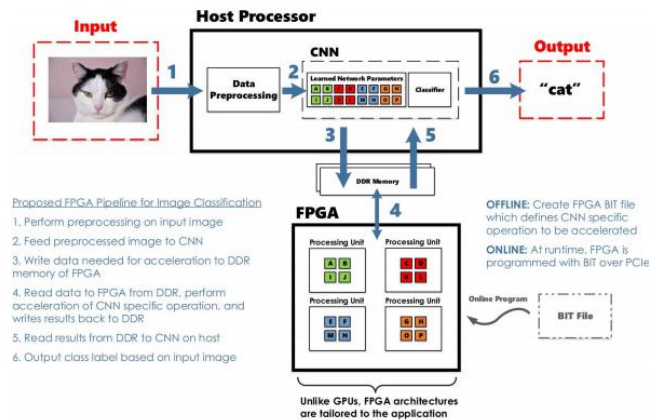
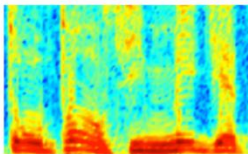


Figure 2: Proposed deployment flow for image classification using FPGA for acceleration.

DL应用在表征学习

AUDIO



Audio Spectrogram

DENSE

IMAGES

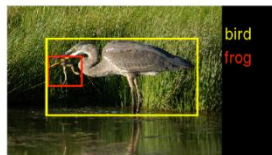


Image pixels

DENSE

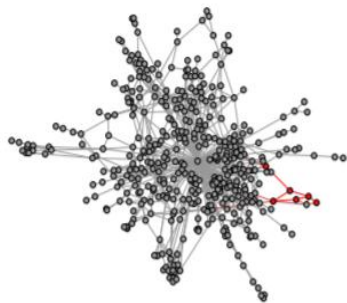
TEXT

0	0	0	0.2	0	0.7	0	0	0
---	---	---	-----	---	-----	---	---	---	-----	-----

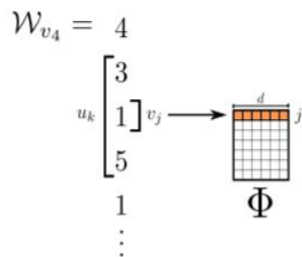
Word, context, or document vectors

SPARSE

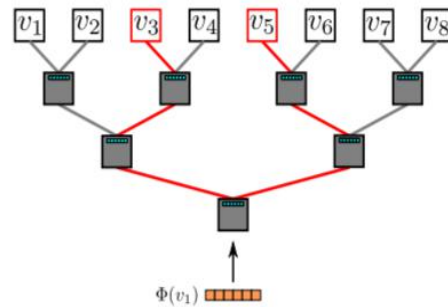
w2v



(a) Random walk generation.

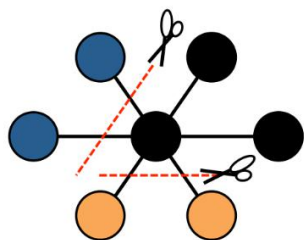


(b) Representation mapping.

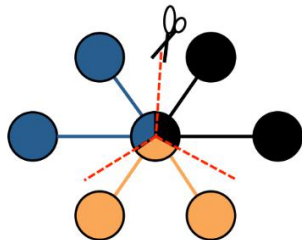


(c) Hierarchical Softmax.

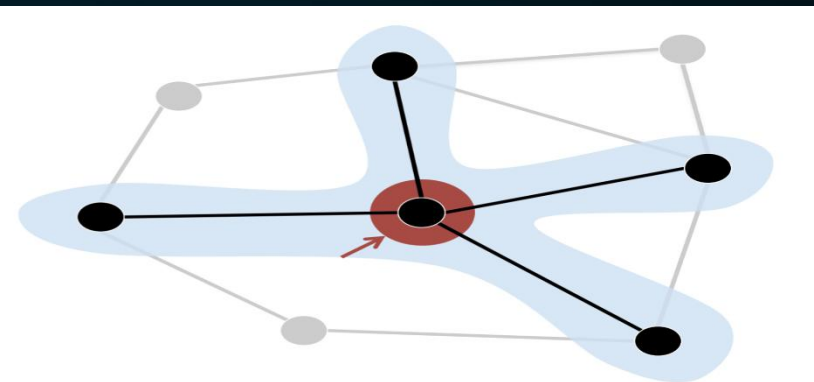
G2V



Edge Cut



Vertex Cut



使用工具:提取维度**graphx,titan**

Machine
1

Machine
2

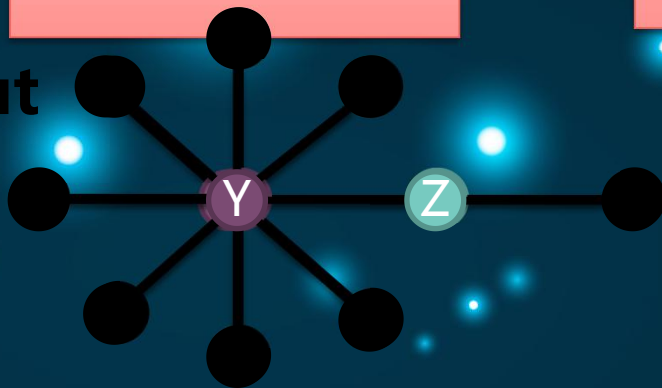
Machine
3

Balanced Vertex-Cut

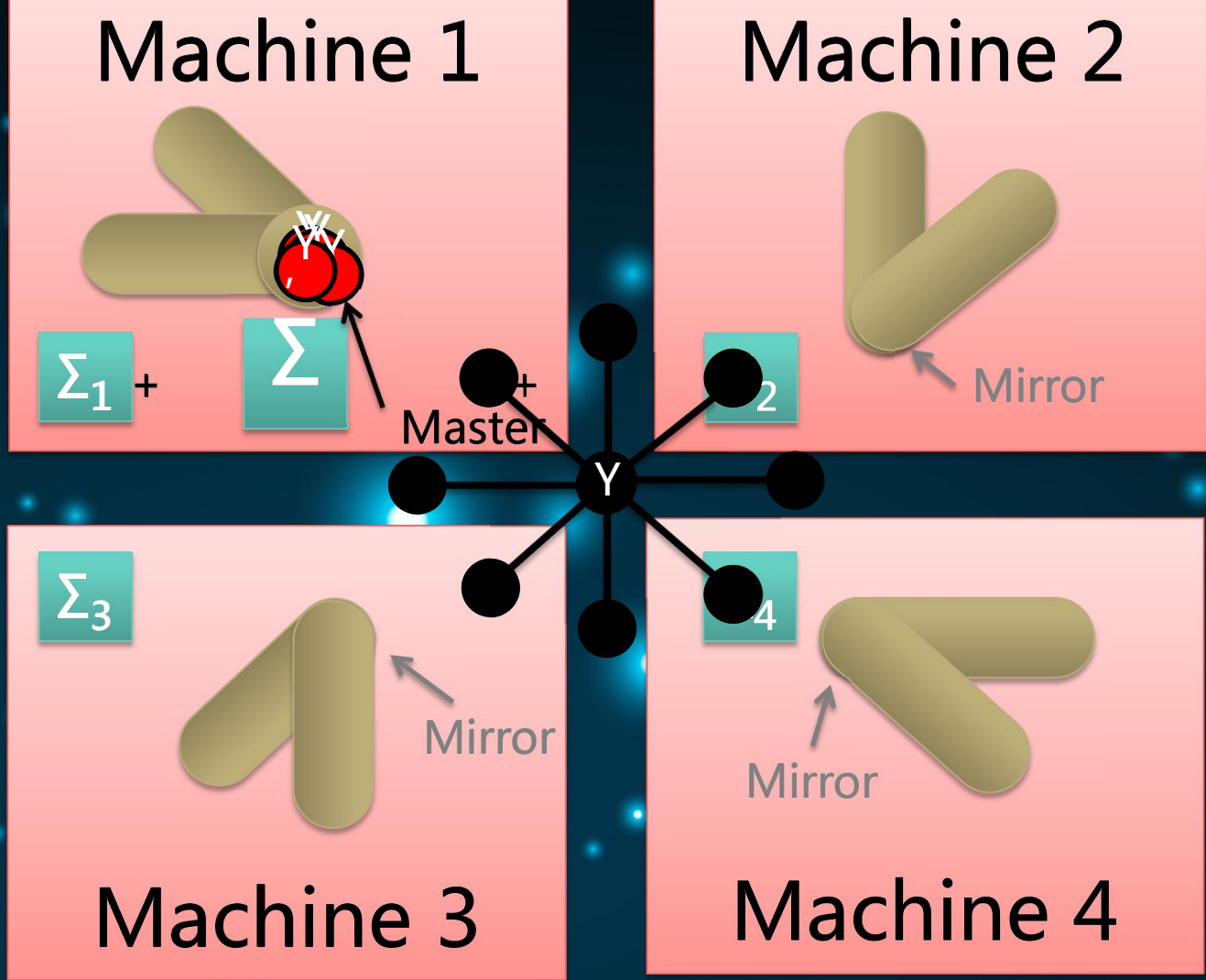
Y Spans 3 Machines

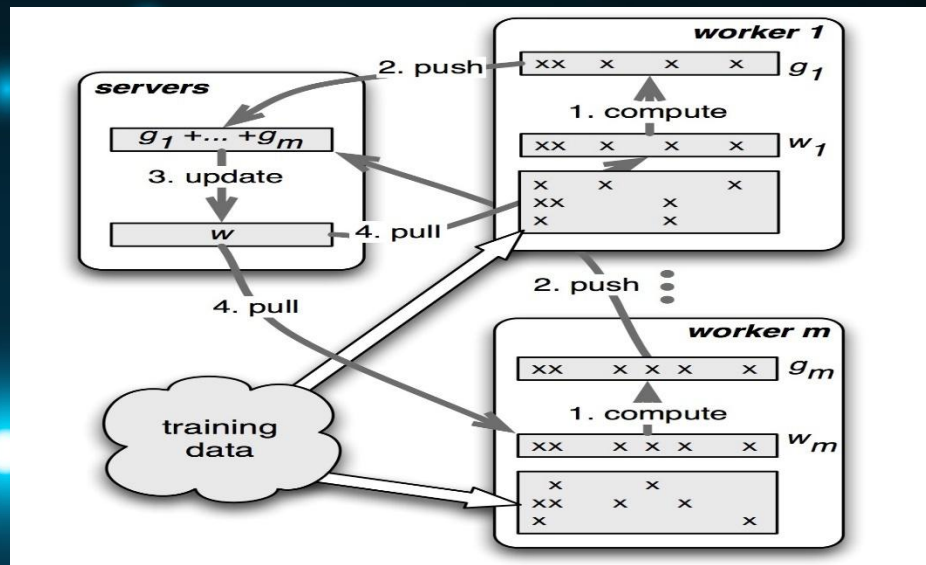
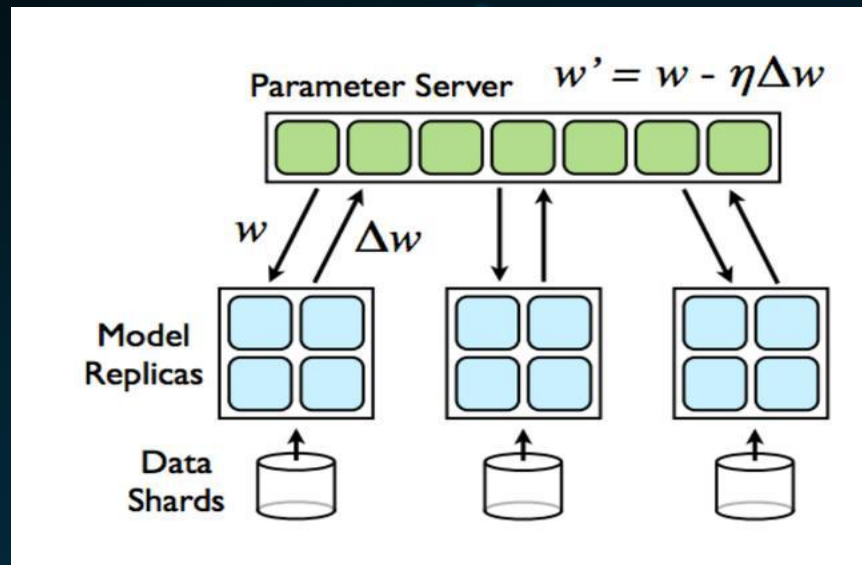
Z Spans 2 Machines

● Not cut!

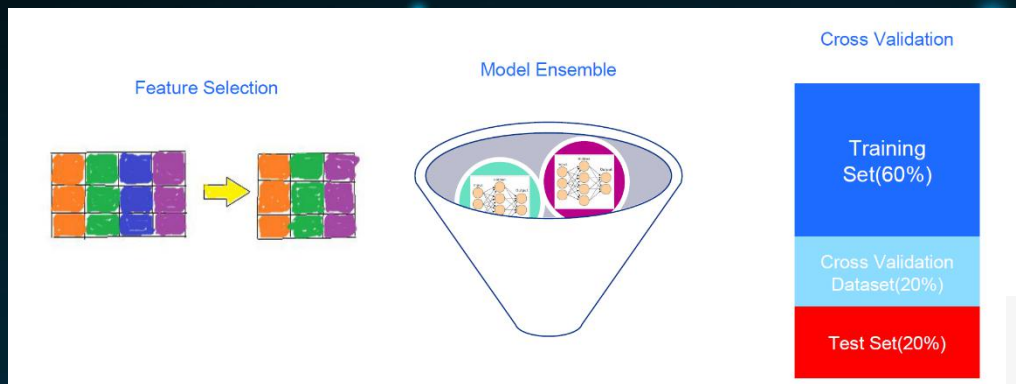


Gather
Apply
Scatter

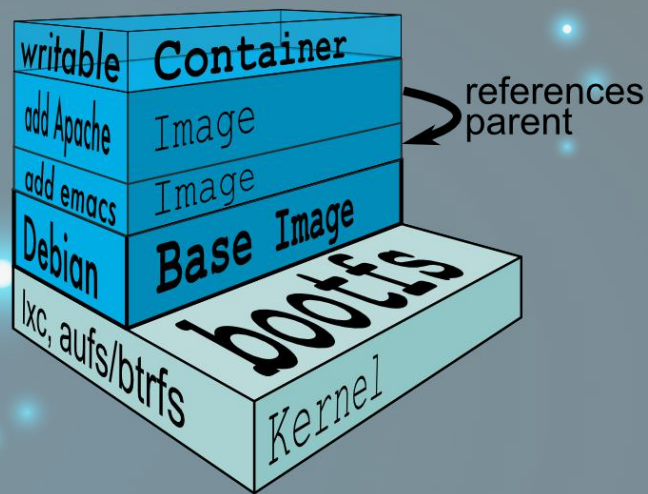




使用工具:模型训练parameter server



数据流:高可用,docker化



效果分析:

1:批量加入与去除

2:训练速度

3:模型维护



Thanks

微博账号：吴炜_机器学习数据挖掘