Mo Tu We Th Fr Sa Su

Memo No.
Date

Transfer_Defect_Learning

矩阵 X　X^T X的转置　$tr(X)$ X的迹
X^{-1} X的逆矩阵　矩阵 $X \in \mathbb{R}^{m \times n}$ 具有 n 个实例 m 列(特征)
λ 表示第几个实例　$x^{(i)}$ 表示第 i 个特征

令　源项目数据集为 $D_S = \{(x_{Si}, y_{Si})\}_{i=1}^{n_1}$
因　$x_{Si} \in \mathbb{R}^{1 \times m}$ 表示输入(例如：一个可能表示向量度量的)
y_{Si} 表示缺陷信息(例如：clean 或 buggy)

令　目标项目数据集为 $D_T = \{x_{Ti}\}_{i=1}^{n_2}$
(注意，x_{Ti} 与 x_{Si} 的度量特征相同)

令 n_1, n_2 分别为源程序和目标程序的文件数量

令 $P(X_S)$ 和 $P(X_T)$ 是 $X_S = \{x_{Si}\}_{i=1}^{n_1}$ 和 $X_T = \{x_{Ti}\}_{i=1}^{n_2}$ 的边缘分布 (X_S X_T 分别来自源程序和目标程序)

通常 $P(X_S)$ 与 $P(X_T)$ 可以不同

TCA 的目的是准确地学习到一个种迁移方法 φ (transformation mapping) 将源程序和目标程序的数据映射到一个潜在的特征空间使得 $P(X_S)$ 和 $P(X_T)$ 的不同变小，从而 直接方法？

Memo No.
Date

Mo Tu We Th Fr Sa Su

在 X_S $p(X_S)$ 和 Y_S 训练的模型可以对 $p(X_T)$ 作出预测

Transfer Component Analysis (TCA)

TCA is to learn a transformation φ to map the original data of source and target domains to a latent space where the difference between domains $Dist(\varphi(X_S), \varphi(X_T))$ is small and the data variance after transformation $Var(\{\varphi(X_S), \varphi(X_T)\})$ is large

很多现实应用中，φ 可以看作是线性的，此时 $\varphi(x) = x \cdot \theta$

$\theta \in \mathbb{R}^{m \times d}$，~~$X \in \mathbb{R}^{n \times m}$~~ 由此将 x 降为 d 维

因而 latent feature $\varphi(X_S) = X_S\theta$ $\varphi(X_T) = X_T\theta$

之后用 $X_S\theta$ 与 Y_S train classifier f

再用 f 预测 $X_T\theta$

不同正则方法对 TCA 在不同情况应用下性能有影响

TCA+

就是选取合适的正则方法的TCA

遍历一定规则

规则设定基于数据集的特征的相似性

Data set characteristic VetVector (DCV)

通过计算实例之间的距离 (欧氏距离得名由来)

一对

每个项目测量项目里每对实例的欧氏距离

$DIST = \{ d_{ij} : \forall i,j, 1 \le n, 1 < j \le n, i < j \}$

此外还将numInstances加入DCV

$DCV = (dist\_mean, dist\_median, dist\_min, dist\_max, dist\_std, numInstances)$

Source Project DCV : C_S Target Project DCV : C_T

为了测量 C_S 与 C_T 的距离，定义了 similarity vector (SV)

并据此为距离分级

$[S[e]]$ 表示 S 中第e个元素的值 $S[e] \in C_T[e]$ 同理

给每对相对应元素分档 通过 $[S[e]] \times a < [C_T[e]] \le [S[e]] \times b$ ($b > a > 0$)

$\forall S, S \Rightarrow T[e]$ ，S与T在e元素距离的等级

Benchmark Sets

① ReLink　论文中表IV 为 概况描述信息
表VI 为 在数据集中被使用的度量

② AEEEM　表V 概况描述信息
表VII 使用的度量

Experiment Design (实验设计)

① within-project defect prediction (同软件缺陷预测)
② 无迁移学习跨软件缺陷预测
③ TCA CDPD
④ TCA+ CDPD

① 一半数据集将随机分成两半，一部分为数据集train set
一部分为test set

[另一种化替方案是使用10次交叉验证(train set: test set = 9:1)]

为了避免片面，将执行100次整个过程，取平均预测结果

Mo Tu We Th Fr Sa Su

Memo No.

Date / /

② 将数据集中的项目两两结对，而每对有两种情况(a=>b或b=>a)，而后例(a=>b) a为训练集，b为测试集

③ 与②类似，但是应该先用TCA训练(加工) source and target data set. 而后再训练预测模型，应用不同正则分别得到多个结果

④ 应用PC TCA+ 规则和方式选则适宜正则方法，再用TCA训练 source and target data set 最后训练预测模型

Machine Learning Classifier (机器学习分类器)

均使用 logistic regression (逻辑回归)

特别地使用 LIBLINEAR

线性执行时，我们选用参数“-S 0”和“-B 1”

⑤ F-measure 作为评估指标