海外研修计划答辩

机器学习在病菌抗性检测中的应用

—— 项目答辩报告 ——

罗晶 大四致远工科CS

提纲

1 实习概况

2 研究成果

3 项目总结

提纲

1 实习概况

2 研究成果

3 项目总结

实习概况

- 研修单位：牛津大学计算健康信息实验室
- 研修时间：2017.06.26-2017.09.02
- 指导教授：David Clifton & Yang Yang

Prof. David Clifton – Group Leader

Prof. Clifton is an Associate Professor in the Department of Engineering Science of the University of Oxford, and a Governing Body fellow of Balliol College, Oxford. He is a Research Fellow of the Royal Academy of Engineering.

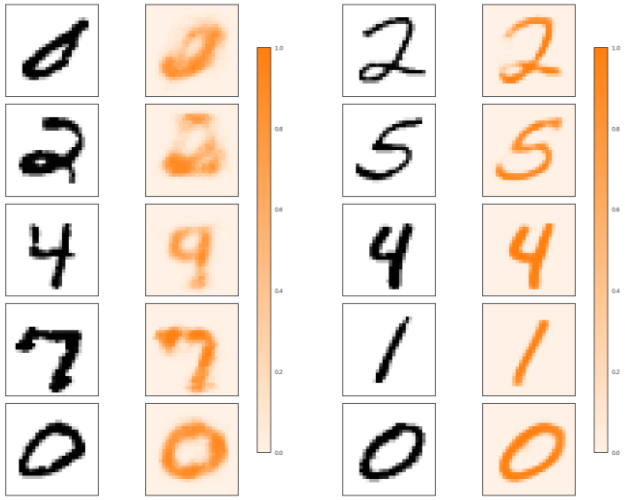
His research focuses on the development of "**big data**" machine learning for tracking the health of complex systems.

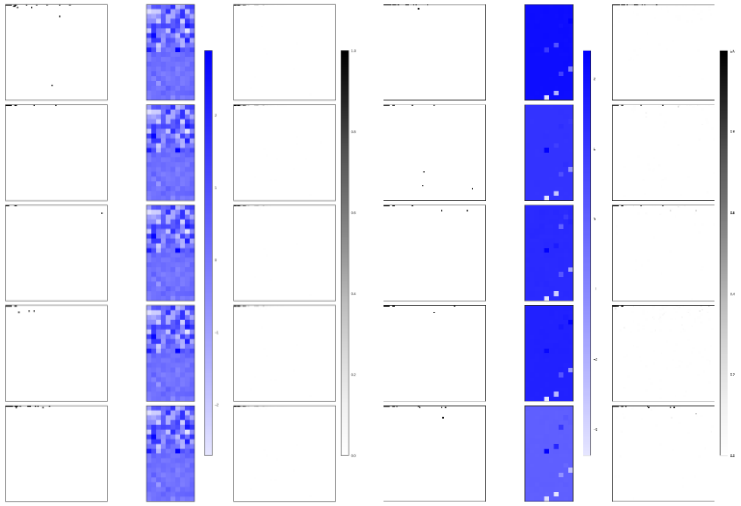
Dr. Yang Yang – Research Fellow

Yang comes from SJTU and joins the CHI Lab as Oxford University's second K.C. Wong Fellow. Her research interests include signal processing and machine condition monitoring.

实习概况

- 研修任务
  - 在David Clifton教授的指导下，使用机器学习的相关知识（分类和降维算法），研究肺结核杆菌基因序列和其对抗生素抗药性的关系，对已知基因序列的肺结核杆菌做出抗性预测。
  - 编写的程序主要包括两部分，一是对肺结核杆菌基因序列的预处理和降维，其次是在搭建好机器学习环境后，基于Python对各类主流分类算法的实现。





提纲

1 实习概况

2 研究成果

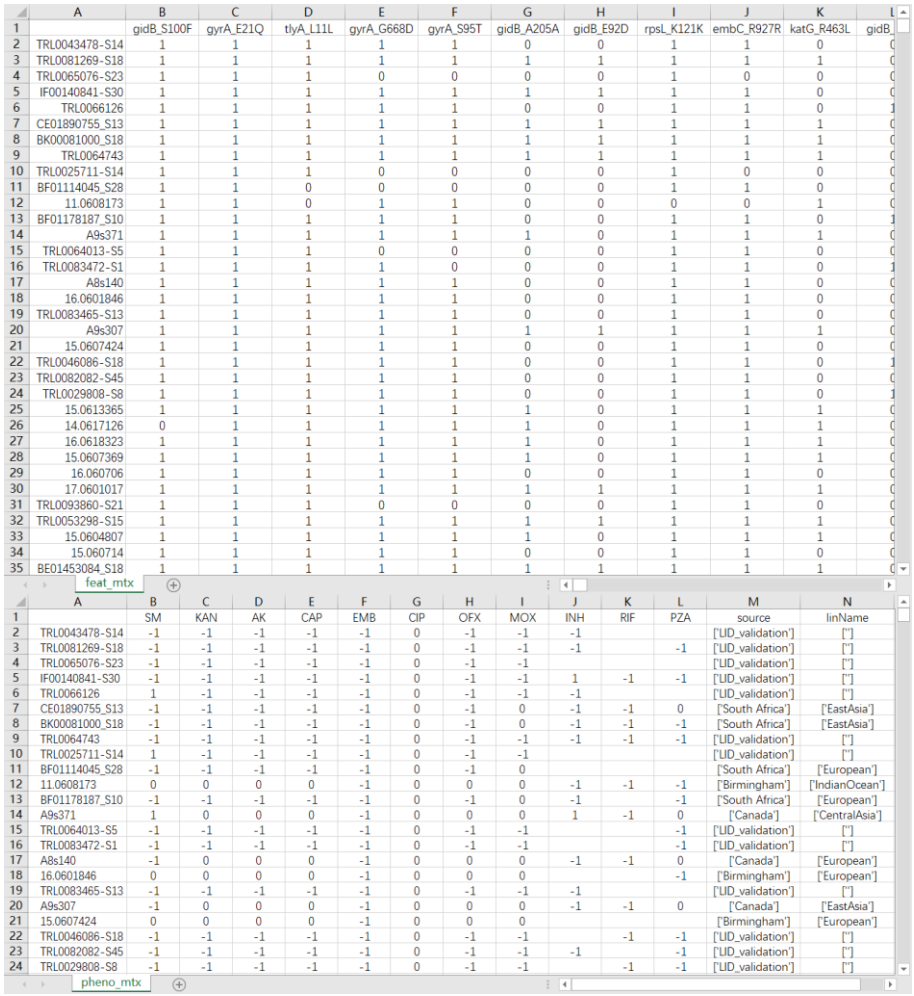
3 项目总结

研究背景

- 抗生素的滥用与“超级细菌”的出现
  - 近几年来，人类医疗对抗生素长期大量的使用过滤了一些对抗生素敏感的病菌，但一些耐药的致病菌却活跃了起来，相当于变相培育了一些“超级细菌”。
  - 以肺结核杆菌为例，市场上普遍使用的药是EMB、RIF，但对这两种药的过度依赖致使肺结核杆菌对这两种的抗药率大大提升。
  - 为了防止病菌对抗生素产生普遍的抗药性，对症下药、精准下药是很有必要的。
- 基因测序的进步和提速
  - 传统实验室培养细菌进行抗性检测往往要长达两三个月。
  - 通过基因测序进行抗性预测和药物诊断成为未来对抗“超级细菌”的最有效手段之一。

数据集概况

- 样本数量：13660
- 基因数量：23
- 特征维度（SNP数量）：5219
- 表现型数量：11
- 挑战：
  - 基因数据维度过高
  - 存在未标注的数据
  - 抗性和感性数据量不平衡
  - 对不同药的表现型是关联的



feat_mtx

	A	B	C	D	E	F	G	H	I	J	K	L
1		gidB_S100F	gyrA_E21Q	tlyA_L11L	gyrA_G668D	gyrA_S95T	gidB_A205A	gidB_E92D	rpsL_K121K	embC_R927R	katG_R463L	gidB_
2	TRL0043478-S14	1	1	1	1	1	0	0	1	1	0	0
3	TRL0081269-S18	1	1	1	1	1	1	1	1	1	1	0
4	TRL0065076-S23	1	1	1	0	0	0	0	1	0	0	0
5	IF00140841-S30	1	1	1	1	1	1	1	1	1	0	0
6	TRL0066126	1	1	1	1	1	0	0	1	1	0	1
7	CE01890755_S13	1	1	1	1	1	1	1	1	1	1	0
8	BK00081000_S18	1	1	1	1	1	1	1	1	1	1	0
9	TRL0064743	1	1	1	1	1	1	1	1	1	1	0
10	TRL0025711-S14	1	1	1	0	0	0	0	1	0	0	0
11	BF01114045_S28	1	1	0	0	0	0	0	1	1	0	0
12	11.0608173	1	1	0	1	1	0	0	0	0	1	0
13	BF01178187_S10	1	1	1	1	1	0	0	1	1	0	1
14	A9s371	1	1	1	1	1	1	0	1	1	1	0
15	TRL0064013-S5	1	1	1	0	0	0	0	1	1	0	0
16	TRL0083472-S1	1	1	1	1	0	0	0	1	1	0	1
17	A8s140	1	1	1	1	1	0	0	1	1	0	0
18	16.0601846	1	1	1	1	1	0	0	1	1	0	0
19	TRL0083465-S13	1	1	1	1	1	0	0	1	1	0	0
20	A9s307	1	1	1	1	1	1	1	1	1	1	0
21	15.0607424	1	1	1	1	1	0	0	1	1	0	0
22	TRL0046086-S18	1	1	1	1	1	0	0	1	1	0	1
23	TRL0082082-S45	1	1	1	1	1	0	0	1	1	0	0
24	TRL0029808-S8	1	1	1	1	1	0	0	1	1	0	1
25	15.0613365	1	1	1	1	1	1	0	1	1	1	0
26	14.0617126	0	1	1	1	1	1	0	1	1	1	0
27	16.0618323	1	1	1	1	1	1	0	1	1	1	0
28	15.0607369	1	1	1	1	1	1	0	1	1	1	0
29	16.060706	1	1	1	1	1	0	0	1	1	0	0
30	17.0601017	1	1	1	1	1	1	1	1	1	1	0
31	TRL0093860-S21	1	1	1	0	0	0	0	1	1	0	0
32	TRL0053298-S15	1	1	1	1	1	1	1	1	1	1	0
33	15.0604807	1	1	1	1	1	1	0	1	1	1	0
34	15.060714	1	1	1	1	1	0	0	1	1	0	0
35	BE01453084_S18	1	1	1	1	1	1	1	1	1	1	0

pheno_mtx

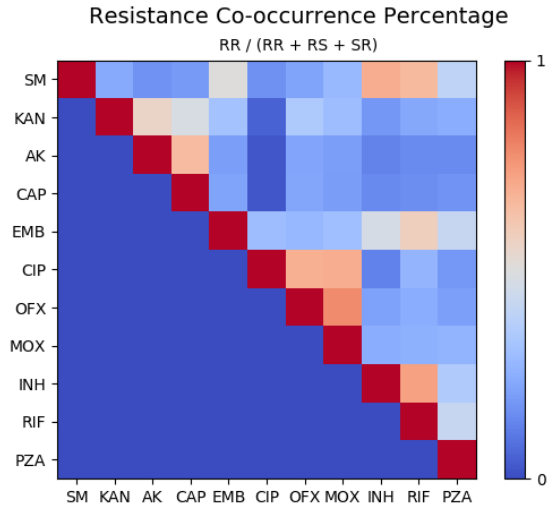
	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1		SM	KAN	AK	CAP	EMB	CIP	OFX	MOX	INH	RIF	PZA	source	linName
2	TRL0043478-S14	-1	-1	-1	-1	-1	0	-1	-1	-1			['UD_validation']	['']
3	TRL0081269-S18	-1	-1	-1	-1	-1	0	-1	-1	-1		-1	['UD_validation']	['']
4	TRL0065076-S23	-1	-1	-1	-1	-1	0	-1	-1				['UD_validation']	['']
5	IF00140841-S30	-1	-1	-1	-1	-1	0	-1	-1	1	-1	-1	['UD_validation']	['']
6	TRL0066126	1	-1	-1	-1	-1	0	-1	-1	-1			['UD_validation']	['']
7	CE01890755_S13	-1	-1	-1	-1	-1	0	-1	0	-1	-1	0	['South Africa']	['EastAsia']
8	BK00081000_S18	-1	-1	-1	-1	-1	0	-1	0	-1	-1	-1	['South Africa']	['EastAsia']
9	TRL0064743	-1	-1	-1	-1	-1	0	-1	-1	-1	-1	-1	['UD_validation']	['']
10	TRL0025711-S14	1	-1	-1	-1	-1	0	-1	-1				['UD_validation']	['']
11	BF01114045_S28	-1	-1	-1	-1	-1	0	-1	0				['South Africa']	['European']
12	11.0608173	0	0	0	0	-1	0	0	0	-1	-1	-1	['Birmingham']	['IndianOcean']
13	BF01178187_S10	-1	-1	-1	-1	-1	0	-1	0	-1		-1	['South Africa']	['European']
14	A9s371	1	0	0	0	-1	0	0	0	1	-1	0	['Canada']	['CentralAsia']
15	TRL0064013-S5	-1	-1	-1	-1	-1	0	-1	-1			-1	['UD_validation']	['']
16	TRL0083472-S1	-1	-1	-1	-1	-1	0	-1	-1			-1	['UD_validation']	['']
17	A8s140	-1	0	0	0	-1	0	0	0	-1	-1	0	['Canada']	['European']
18	16.0601846	0	0	0	0	-1	0	0	0			-1	['Birmingham']	['European']
19	TRL0083465-S13	-1	-1	-1	-1	-1	0	-1	-1	-1			['UD_validation']	['']
20	A9s307	-1	0	0	0	-1	0	0	0	-1	-1	0	['Canada']	['EastAsia']
21	15.0607424	0	0	0	0	-1	0	0	0				['Birmingham']	['European']
22	TRL0046086-S18	-1	-1	-1	-1	-1	0	-1	-1		-1	-1	['UD_validation']	['']
23	TRL0082082-S45	-1	-1	-1	-1	-1	0	-1	-1	-1		-1	['UD_validation']	['']
24	TRL0029808-S8	-1	-1	-1	-1	-1	0	-1	-1		-1	-1	['UD_validation']	['']

数据集分析

- 数据特征
  - S>R
  - 后6种表现型Missing Label的情况严重

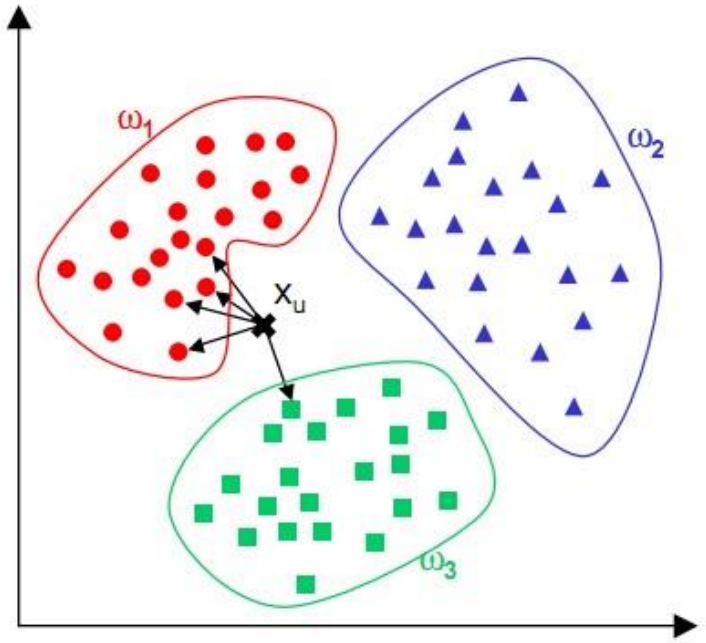
Antibiotic	Susceptible Samples	Resistant Samples	Unknown Samples
EMB	10933	1670	1057
RIF	9660	2915	1085
INH	8137	3592	1931
PZA	9267	1147	3246
SM	5231	1860	6569
OFX	2618	458	10584
CAP	2741	315	10604
AK	2690	273	10697
KAN	1925	242	11493
MOX	1249	262	12149
CIP	529	77	13054

- 通过相关系数分析可以将11种药的表现型分成3组：
  - SM EMB INH RIF (PZA)
  - KAN AK CAP
  - CIP OFX MOX

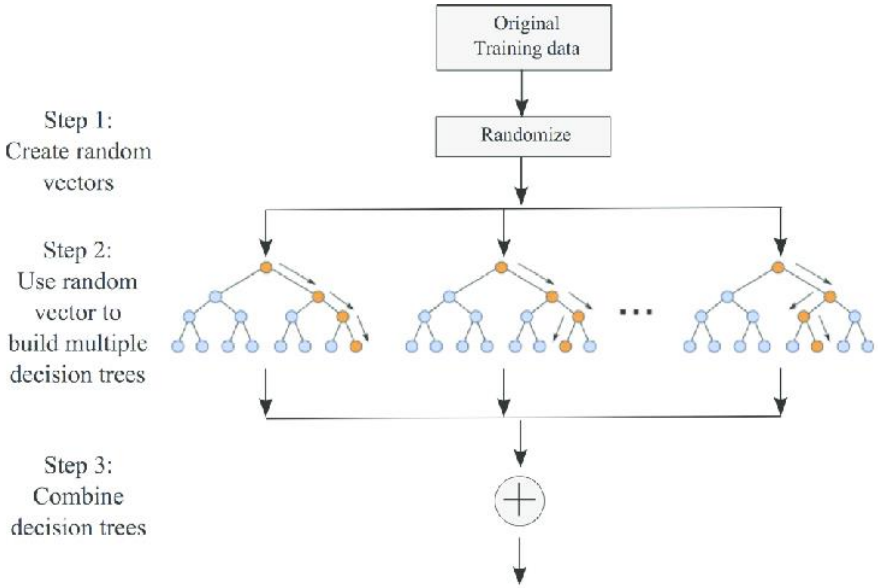


分类器

- K-近邻

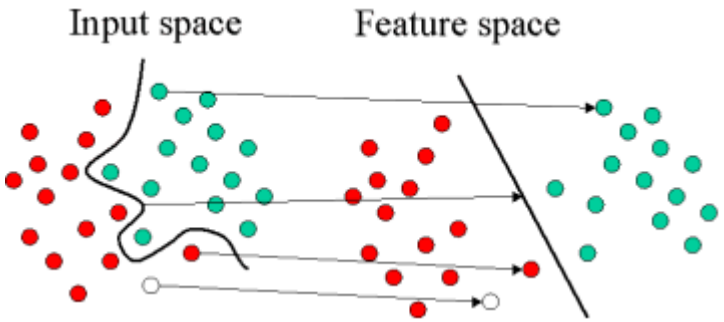

ω1
ω2
ω3
Xu

- 随机森林

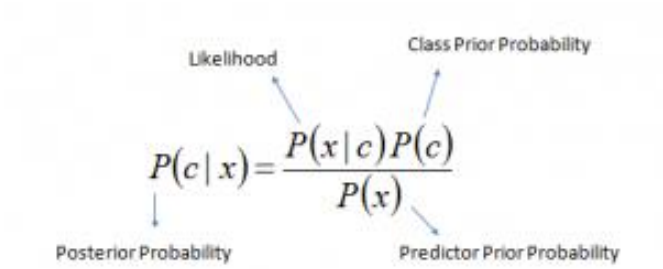

Original Training data
Randomize
Step 1: Create random vectors
Step 2: Use random vector to build multiple decision trees
Step 3: Combine decision trees

分类器

- 支持向量机

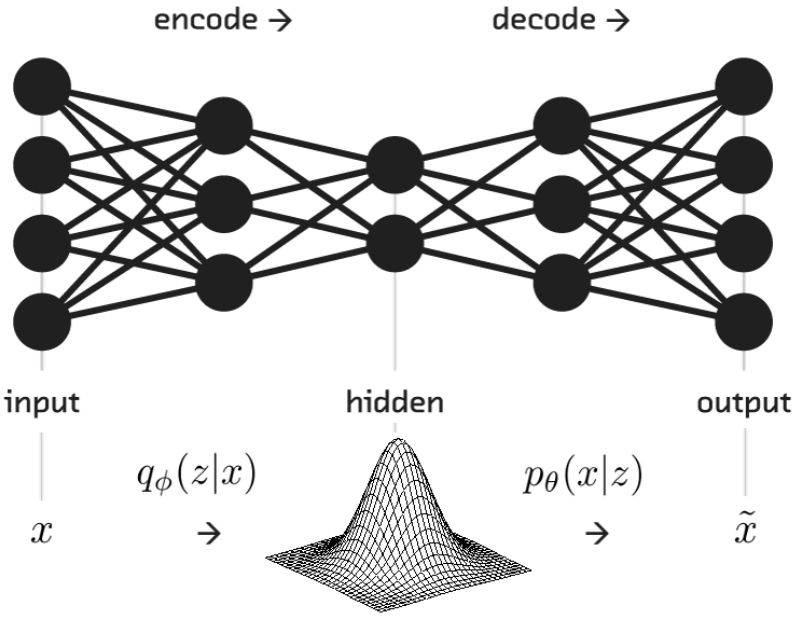


- 朴素贝叶斯分类



$$P(c|X) = P(x_1|c)\times P(x_2|c)\times\cdots\times P(x_n|c)\times P(c)$$

- SemiVAE



分类器

- 实验结果

Drug	SVM-linear		Naive Bayes Classifier		Random Forest		KNN		SVM + PCA		Semisupervised VAE	
	Sen (std)	Spec (std)	Sen (std)	Spec (std)	Sen (std)	Spec (std)	Sen (std)	Spec (std)	Sen (std)	Spec (std)	Sen (std)	Spec (std)
INH	94.2 (6.2)	96.9 (1.9)	93.9 (4.3)	97.4 (1.4)	90.3 (11.9)	96.7 (1.8)	91.9 (10.4)	96.5 (3.3)	93.7 (7.6)	90.1 (6.0)	94.7 (6.2)	90.2 (7.3)
EMB	89.3 (7.1)	92.1 (9.3)	79.3 (12.4)	95.4 (5.9)	74.0 (10.5)	95.0 (6.8)	76.8 (14.4)	95.6 (6.7)	86.9 (7.7)	92.4 (8.9)	89.4 (6.5)	92.2 (9.2)
RIF	95.5 (4.7)	95.6 (5.7)	92.6 (5.3)	96.4 (4.3)	89.0 (12.1)	96.5 (5.0)	92.4 (10.7)	98.6 (1.0)	95.6 (3.5)	92.7 (6.6)	95.9 (3.8)	92.5 (7.3)
PZA	85.4 (17.3)	89.6 (15.1)	60.9 (23.3)	91.2 (13.2)	60.7 (19.5)	92.5 (10.6)	66.7 (18.4)	92.2 (16.6)	76.7 (24.1)	88.5 (16.0)	88.4 (23.8)	89.3 (16.3)

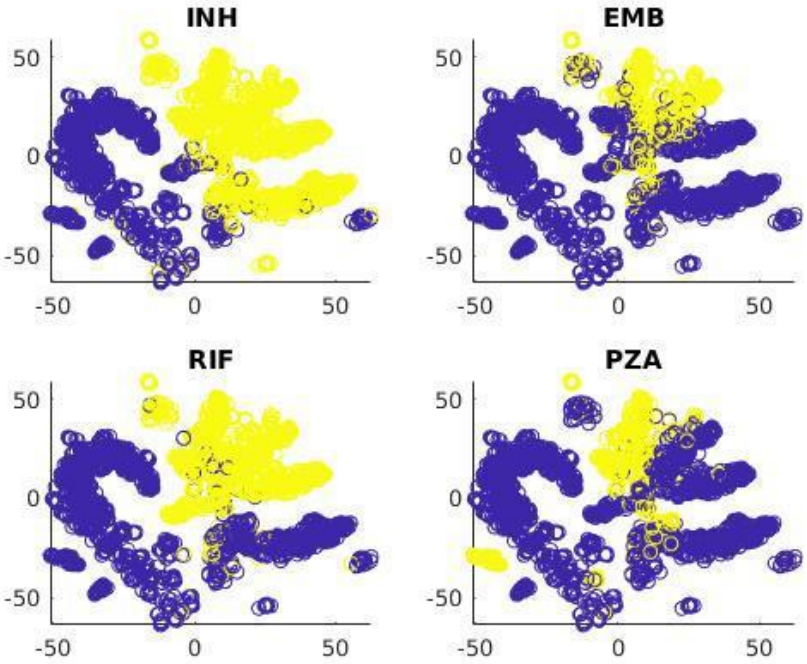
降维算法与多标签学习

- 多标签学习

INH	RIF	EMB	PZA	Count	Tag
1	-1	-1	-1	621	INH
-1	1	-1	-1	93	RIF
-1	-1	-1	1	123	PZA
1	1	-1	-1	472	INH&RIF
1	-1	1	-1	28	INH&EMB
1	-1	-1	1	34	INH&PZA
1	1	1	-1	411	INH&RIF&EMB
1	1	-1	1	223	INH&RIF&PZA
1	1	1	1	629	INH&RIF&EMB&PZA
-1	-1	-1	-1	5312	Susceptible to all four

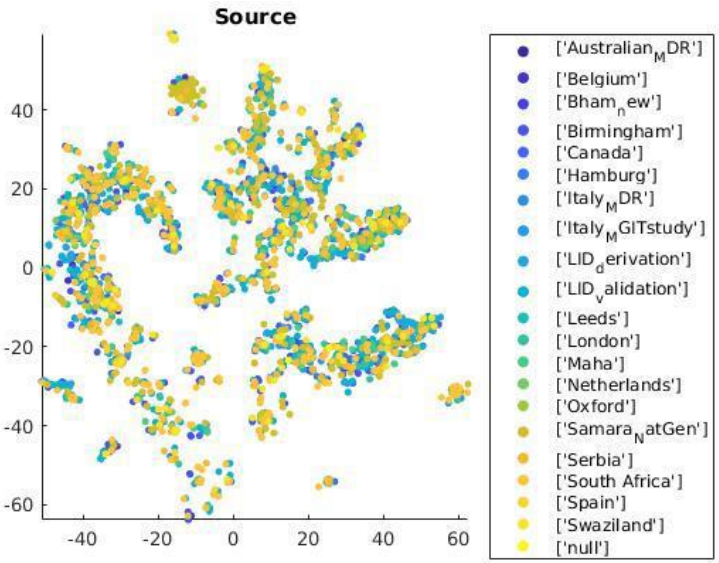
降维算法与多标签学习

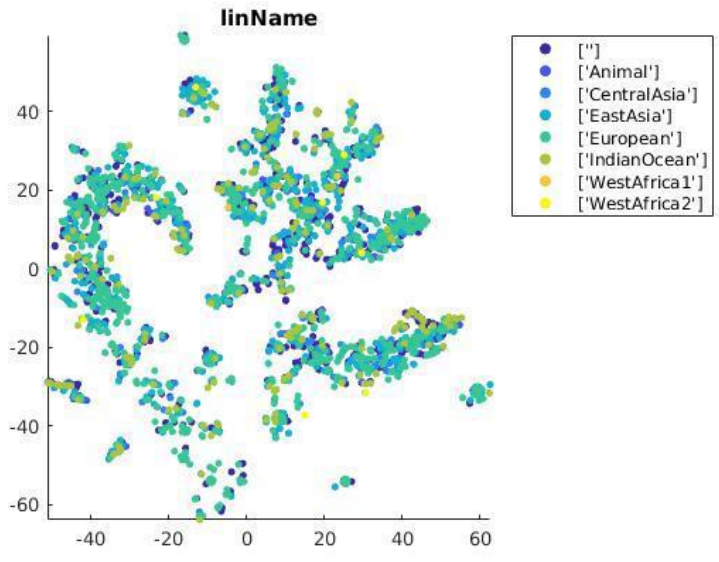
- 基于t-SNE的深度学习
  - 将数据压缩至二维，出现了清晰的聚类效果



降维算法与多标签学习

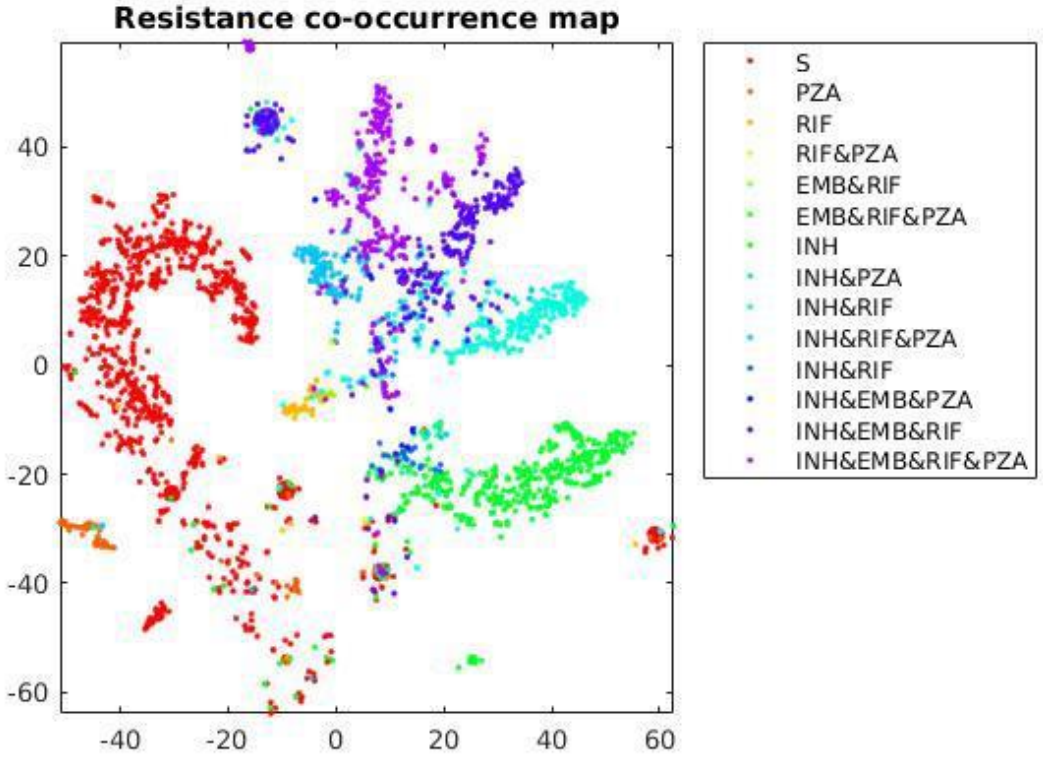
- 基于t-SNE的深度学习
  - 降维后的数据聚类与数据的来源无关





降维算法与多标签学习

- 基于t-SNE的深度学习
  - 多标签学习：聚类效果明显



提纲

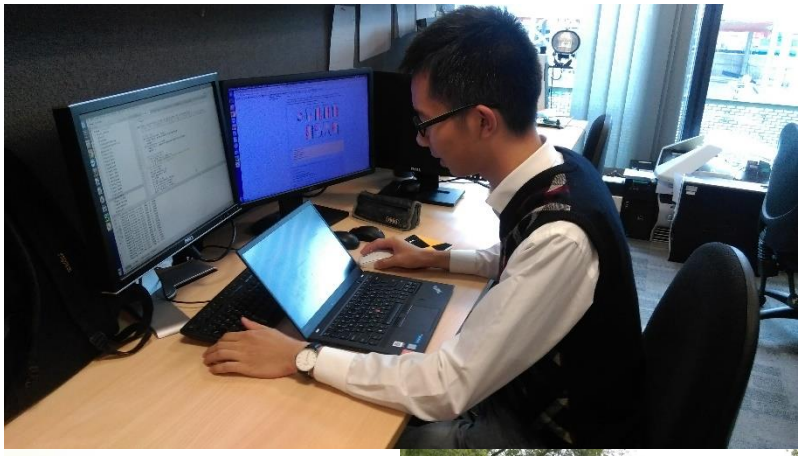
1 实习概况

2 研究成果

3 项目总结

项目总结

- 巩固了此前学习的机器学习的基础知识，练习将学到的算法切合实际地应用到有现实意义的问题中。
- 学到了新的机器学习算法和模型，了解并应用了VAE模型，诸多降维算法等。
- NIPS healthcare workshop
- 英国生活




THANKS