机器学习
Machine Learning

鲍军鹏

2024年2月（V1.5.1）

西安交通大学计算机学院
Email: baojp@xjtu.edu.cn

课程简介

★ **终极目标：**

让计算机能够象人一样学习，能够从数据发现规律和有用信息。

★ **通过学习：**

了解主要研究方向、途径及研究内容，掌握基本原理及方法，初步具备解决实际问题的能力

★ **参考书：**

[1] Jiawei Han（美国）、Micheline Kamber（加拿大）、Jian Pei（加拿大）. Data Mining: Concepts and Techniques（英文第3版）. 北京：机械工业出版社，2012.
[2] 周志华. 机器学习. 北京: 清华大学出版社, 2016.
[3] 吴恩达（Andrew Ng）. Machine Learning. https://www.coursera.org/learn/machine-learning
[4] 史忠植. 知识发现（第2版）. 北京：清华大学出版社，2011
[5] Tom Mitchell. Machine Learning. .北京：机械工业出版社，2008

主要内容

第一章：概述

第二章：数据与度量

第三章：分类方法

第四章：聚类方法

第五章：人工神经网络与深度学习

第六章：智能优化方法

实验： 编程实践

考核内容与方式

★平时（60%）

- 实践与实验：40%
  - 平时要调试代码，实验课上主要展示和讨论结果。
- 论文阅读与讲解：8%
  - 从最近2年顶级国际会议论文中选一篇精读，并做成PPT给大家讲解。
  - 会议列表：AAAI、ACL、KDD、IJCAI、NIPS、CVPR、ICML、ICCV、 ICLR 。
- 课堂表现：12%
  - 主动回答问题，主动发言，得12分。
  - 被动回答问题正确者可得10分，被动回答错误者得5分。
  - 沉默不语，从不发言者得0分。

★笔试（40%）

- 各自为战

学习方法

★重在平时，理解原理，代码是王道

- 多练
  - 平时要多写代码，花时间去调试代码，积累解决问题的经验。
  - 善于观察和对比，精益求精，逐步提升自身动手能力。
- 多读
  - 机器学习进展非常快，要通过大量阅读最新的文献来了解学科前沿，发现并聚焦自己的学习兴趣。
  - 提高自身的理解能力，善于总结，善于发现。
- 认真思考，活学活用
  - 虽然算法很多，但是不要刻板地记忆公式，
  - 应该去理解公式的本质涵义和内在思想，
  - 做到举一反三，不断创新。

作业内容

★大作业内容

- 1、分类和预测方法实践
- 2、聚类方法实践
- 3、智能优化方法实践

★基本要求

- 至少完成前2道题目
- 每道题目至少使用2种不同算法或者模型

★加分项

- 完成了全部实践内容，并且对比了3种及以上算法
- 或者完成了其它机器学习相关任务

作业要求

★基本要求：会用现有方法解决问题

- 个人独立完成
- 用Python或者JAVA、C++、C调用库函数完成任务
- 对同一算法的不同参数设置进行对比，寻找优化参数
- 对解决同一问题的不同算法进行对比，包括运行结果（精度、召回率、误差等）和时间（训练时间、执行时间）
- 撰写完整的实验报告（参考LNCS论文模板）
- 提交实验报告（.docx文件）和源代码（只包含.java、.py、.txt文件的zip压缩包），不要包含源数据

作业要求

★较高要求：知其然并知其所以然

- 完成基本要求所有内容
- 不用库中现成算法函数，自行实现基本算法过程
- 对比不同版本的实现算法

★挑战自我：学会突破创新

- 复现当前最新论文中解决该问题的算法，并与现有算法进行对比
- 能够对最新算法进行改进，并完成对比实验
- 以改进算法为基础撰写一篇学术论文

作业要求

★作业评分说明

- 如果作业是调用现成库函数完成挑战内容（核心代码在4行之内完成基本算法过程），那么完成全部内容最多得80分。
- 如果作业是自行实现了算法基本过程（理解了算法内容，自己从头实现了算法过程），那么完成全部内容最多得100分。
- 如果作业是基于实验室工作对现有方法进行了改进，并给出了完整对比实验结果，那么完成全部内容最多得120分（参考指导教师的评价）。

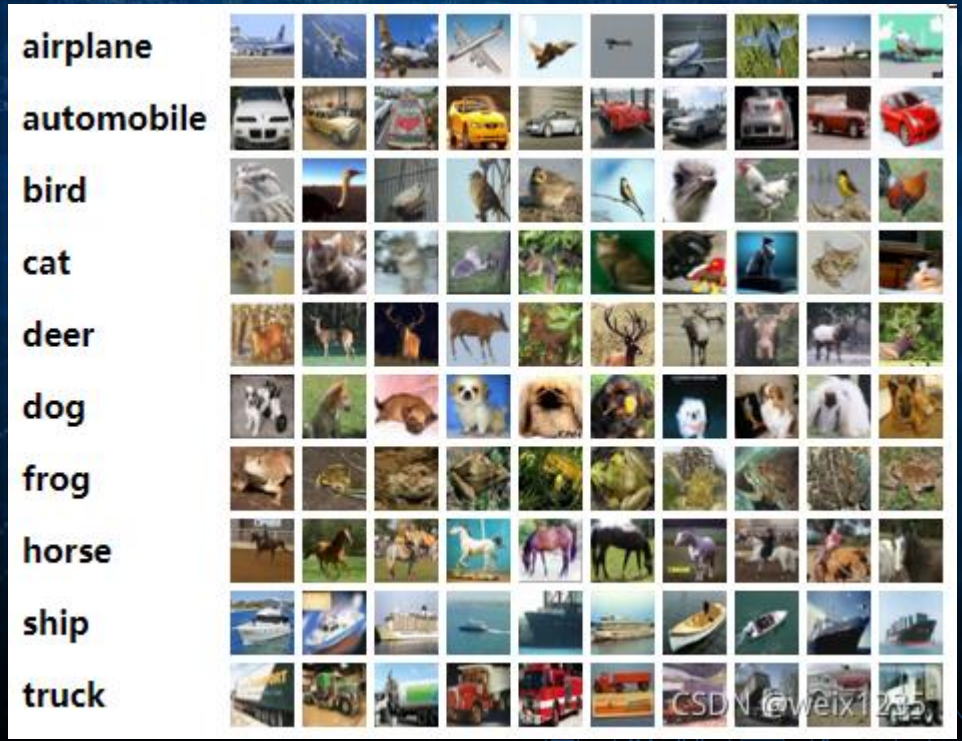
实验与实践

- ★ 分类与预测算法实践
- ★ 聚类算法实践
- ★ 优化算法实践

分类与预测算法实践

★挑战1：图像识别

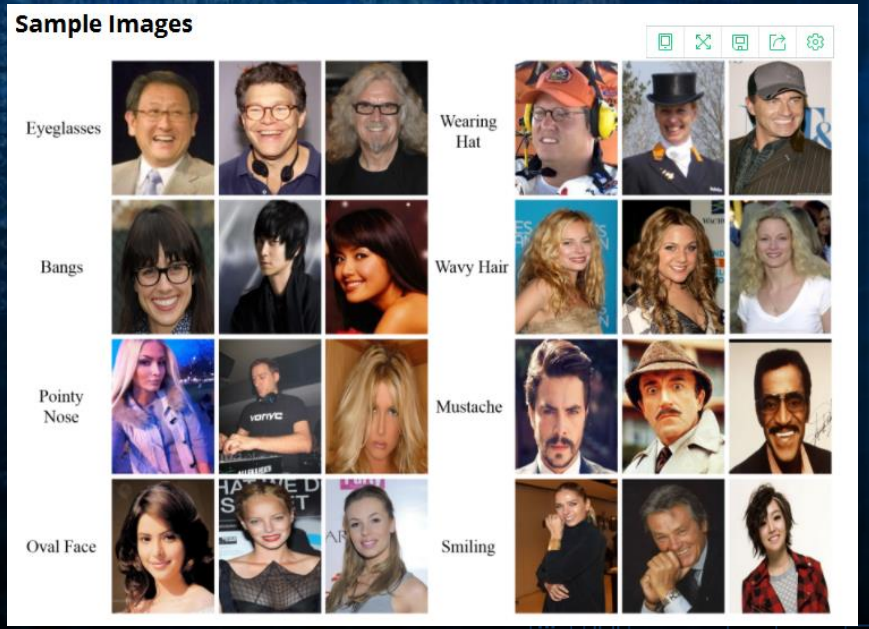
- 在CIFAR-10 数据集上进行学习，然后测试是否能正确识别输入图像的类别。
- 可使用SVM、BP、LR、KNN、CNN(VGG、Resnet)等算法实现。
- 请比较分析传统机器学习方法和深度学习算法的识别效果、学习效率和运行速度。



分类与预测算法实践

★挑战2：人脸识别

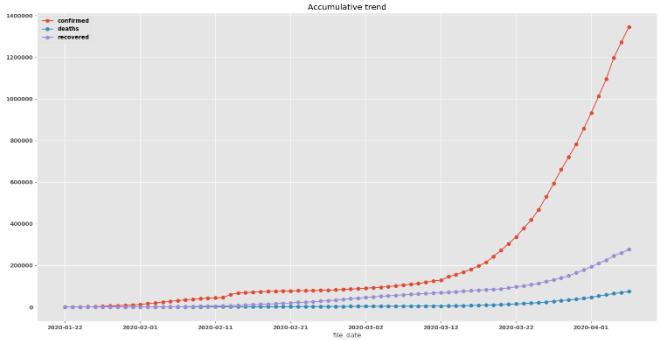
- 在CelebA人脸数据集（香港中文大学的开放数据）上加入本组人员与数据集中标注相近的照片（如带眼镜）进行学习，然后从十个自己照片中识别出自己。
- 可使用CNN (VGG、Resnet)、SVM、BP、LR、KNN等算法实现
- 请比较分析传统机器学习方法和深度学习算法的识别效果、学习效率和运行速度。



分类与预测算法实践

★挑战3：时序数据预测

- 预测美国新型冠状病毒（COVID-19）新增确诊人数，疫情数据来自美国约翰·霍普金斯大学公开数据集 https://github.com/CSSEGISandData/COVID-19
- 可使用LSTM、SVR、BP、LR等算法来分析和预测疫情
- 并比较不同算法的识别效果、学习效率和运行速度。



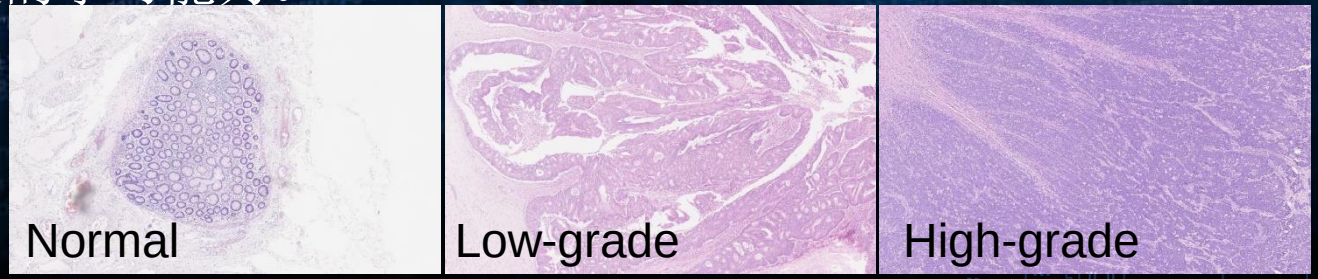
K	DT	DU	DV	DW	DX	DY	DZ	EA	EB	EC	ED	EE	EF
Combined_Key	5/13/20	5/14/20	5/15/20	5/16/20	5/17/20	5/18/20	5/19/20	5/20/20	5/21/20	5/22/20	5/23/20	5/24/20	5/25/20
Autauga, Alabama, US	90	100	100	108	118	124	130	135	148	151	156	160	171
Baldwin, Alabama, US	237	247	253	261	267	267	269	277	277	278	280	281	284
Barbour, Alabama, US	70	74	79	81	85	89	92	96	101	106	108	113	116
Bibb, Alabama, US	46	47	51	52	52	53	53	54	54	56	59	60	63
Blount, Alabama, US	48	48	48	49	49	50	50	50	51	52	52	52	52
Bullock, Alabama, US	27	27	32	35	39	50	58	64	71	95	103	109	138
Butler, Alabama, US	231	244	254	265	278	289	297	306	315	325	329	338	360
Calhoun, Alabama, US	133	134	135	136	139	139	142	142	143	145	145	148	153
Chambers, Alabama, US	333	334	335	337	337	338	339	342	343	343	343	349	350
Cherokee, Alabama, US	24	24	25	26	27	28	30	30	31	32	32	32	33
Chilton, Alabama, US	76	78	79	81	83	84	85	87	88	89	90	91	91
Choctaw, Alabama, US	73	77	81	84	85	89	124	127	132	135	139	140	143
Clarke, Alabama, US	65	67	70	74	75	77	86	89	91	92	97	102	113
Clay, Alabama, US	24	24	24	24	24	24	24	24	24	24	24	24	24
Cleburne, Alabama, US	13	13	13	13	13	13	13	13	13	13	13	13	13
Coffee, Alabama, US	156	156	159	161	167	171	175	180	182	187	193	197	205
Colbert, Alabama, US	79	82	87	93	96	100	106	112	113	119	126	137	144
Conecuh, Alabama, US	17	18	19	19	20	21	21	22	23	23	24	24	28
Coosa, Alabama, US	32	32	32	32	32	32	32	32	32	32	32	33	34
Covington, Alabama, US	59	59	60	61	62	62	63	63	66	67	67	69	72
Crenshaw, Alabama, US	50	51	55	56	56	56	57	57	57	59	64	64	66

csse_covid_19_data_csse_covid_1

分类与预测算法实践

★挑战4：医学图像识别

- 癌变组织图像识别，其中有正常组织，也有不同级别肿瘤的组织染色图像。数据来自英国Tissue Image Analytics (TIA) Centre (https://warwick.ac.uk/fac/cross_fac/tia/data/extended_crc_grading/)
- 可使用CNN(VGG、Resnet)、SVM、BP、LR、KNN等算法来实现
- 请比较分析传统机器学习方法和深度学习方法对于小数据集的学习能力。



分类与预测算法实践

★挑战5：文本分类

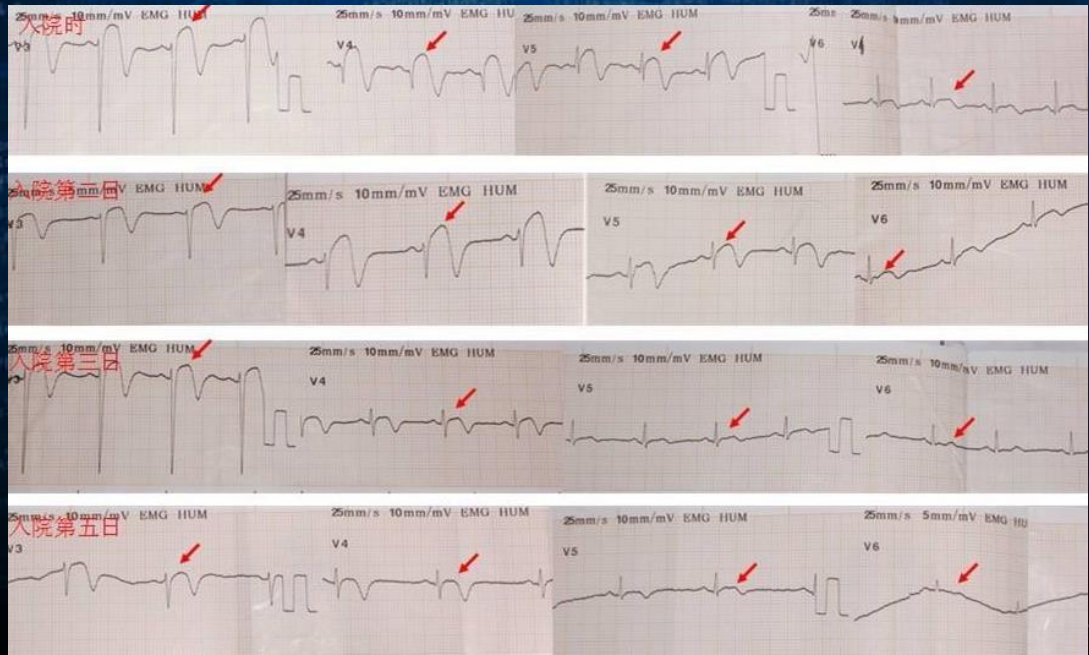
- 新闻分类。今日头条中文新闻（短文本）分类数据集：(https://github.com/fateleak/toutiao-text-classfication-dataset)数据规模：共38万条，分布于15个分类中。
- 可使用LSTM、SVM、BP、LR、KNN、朴素贝叶斯、随机森林等算法实现。
- 请比较分析传统机器学习方法和深度学习算法的识别效果、学习效率和运行速度。

民生	故事	news_story
文化	文化	news_culture
娱乐	娱乐	news_entertainment
体育	体育	news_sports
财经	财经	news_finance
房产	房产	news_house
汽车	汽车	news_car
教育	教育	news_edu
科技	科技	news_tech
军事	军事	news_military
旅游	旅游	news_travel
国际	国际	news_world
证券	股票	stock
农业	三农	news_agriculture
电竞	游戏	news_game

分类与预测算法实践

★挑战6：时序数据识别

- 心电图（时间序列）识别。数据集可采用UCR时间序列分类文档中的ECG5000数据（http://www.cs.ucr.edu/~eamonn/time_series_data/）。
- 可使用LSTM、SVM、BP、LR等算法实现，
- 并比较不同算法的识别效果、学习效率和运行速度。



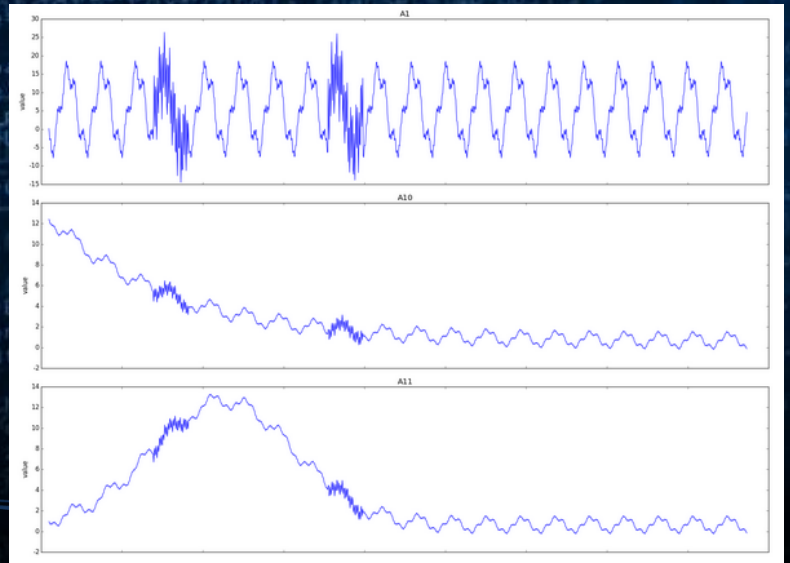
实验与实践

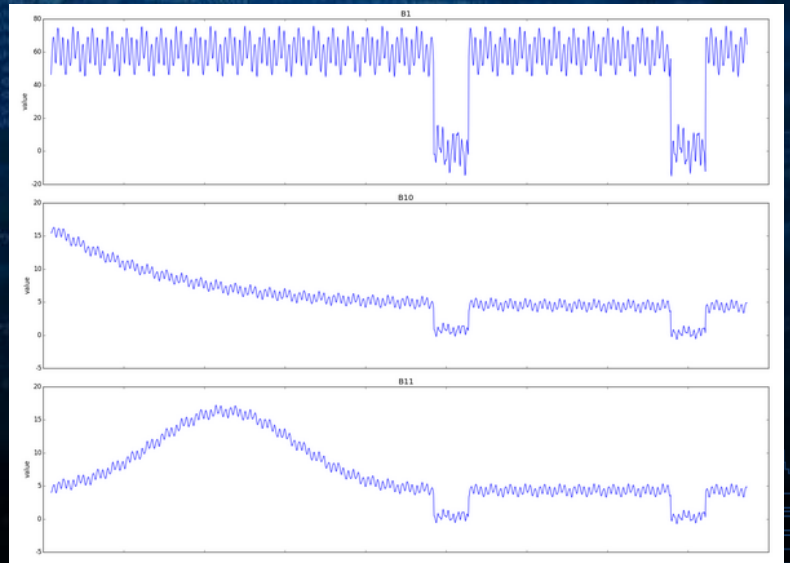
- ★ 分类与预测算法实践
- ★ 聚类算法实践
- ★ 优化算法实践

聚类算法实践

★挑战1:时序数据聚类

- 对时序数据（timeData3）进行聚类，要求能够获得最正确的聚类结果。
- 可使用K-Means、SLink、DBScan等算法实现，
- 并比较不同算法的聚类效果和运行速度。

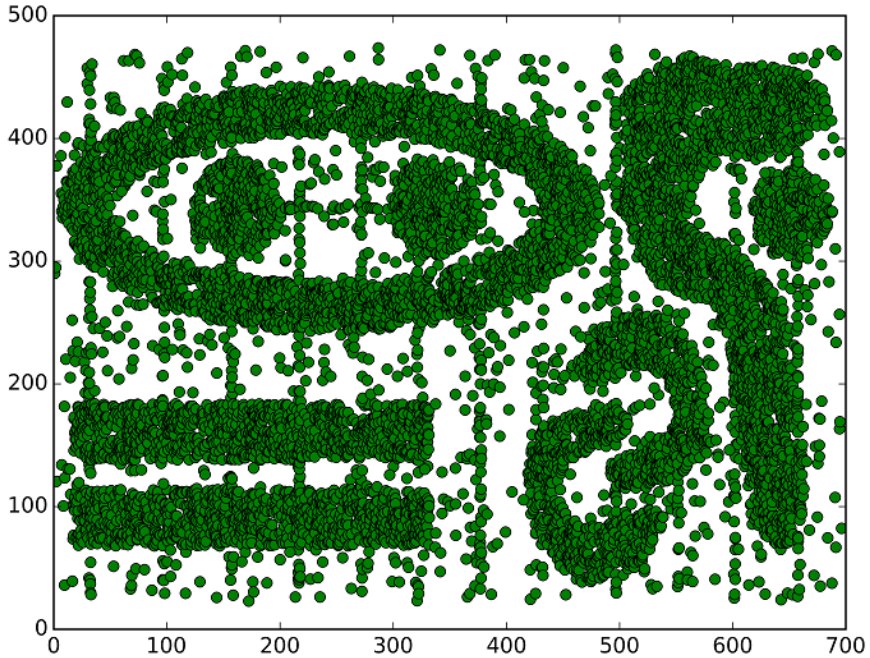




聚类算法实践

★挑战2：二维数据聚类

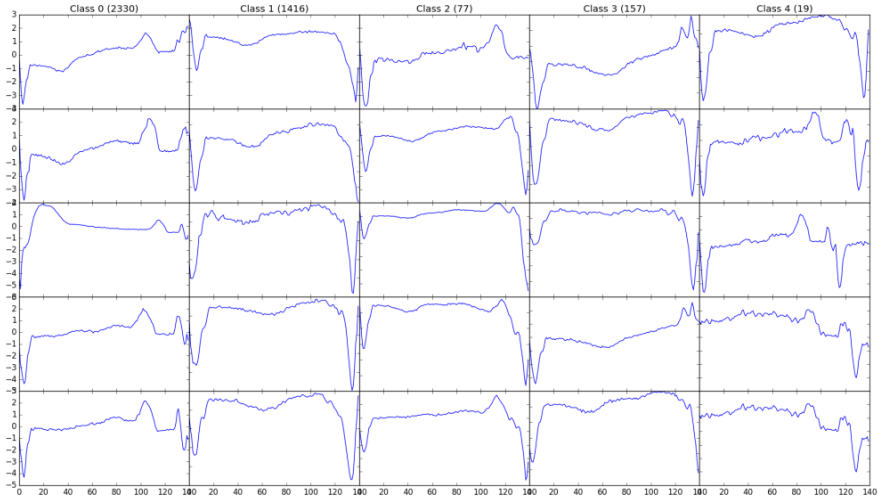
- 对如下二维数据（clusterData1）进行聚类，要求能够获得最正确的聚类结果。
- 可使用K-Means、 SLink、DBScan等算法实现，
- 并比较不同算法的聚类效果和运行速度。



聚类算法实践

★挑战3：时序数据聚类

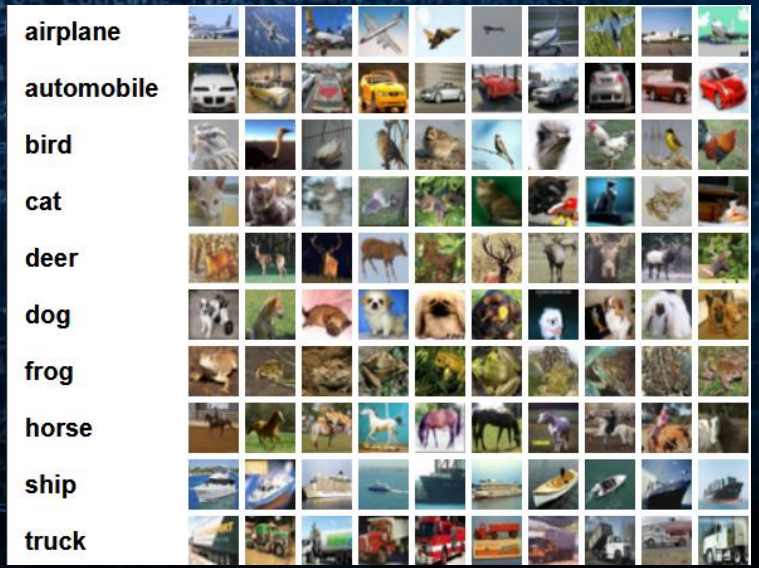
- 对心电图数据（ECG）进行聚类，要求能够获得最正确的聚类结果。数据来源于MIT-BIH Arrhythmia Database (https://www.physionet.org/content/mitdb/1.0.0/)
- 可使用K-Means、SLink、DBScan等算法实现
- 并比较不同算法的聚类效果和运行速度。



聚类算法实践

★挑战4：图像聚类

- 对CIFAR-10数据进行聚类，就是在学习过程中不使用图像标签信息。标签信息只用于统计评价聚类结果。
- 可使用K-Means、SLink、DBScan等算法实现。
- 并比较不同算法的聚类效果和运行速度。



Superclass	Classes
aquatic mammals	beaver, dolphin, otter, seal, whale
fish	aquarium fish, flatfish, ray, shark, trout
flowers	orchids, poppies, roses, sunflowers, tulips
food containers	bottles, bowls, cans, cups, plates
fruit and vegetables	apples, mushrooms, oranges, pears, sweet peppers
household electrical devices	clock, computer keyboard, lamp, telephone, television
household furniture	bed, chair, couch, table, wardrobe
insects	bee, beetle, butterfly, caterpillar, cockroach
large carnivores	bear, leopard, lion, tiger, wolf
large man–made outdoor things	bridge, castle, house, road, skyscraper
large natural outdoor scenes	cloud, forest, mountain, plain, sea
large omnivores and herbivores	camel, cattle, chimpanzee, elephant, kangaroo
medium–sized mammals	fox, porcupine, possum, raccoon, skunk
non–insect invertebrates	crab, lobster, snail, spider, worm
people	baby, boy, girl, man, woman
reptiles	crocodile, dinosaur, lizard, snake, turtle
small mammals	hamster, mouse, rabbit, shrew, squirrel
trees	maple, oak, palm, pine, willow
vehicles 1	bicycle, bus, motorcycle, pickup truck, train
vehicles 2	lawn–mower, rocket, streetcar, tank, tractor

聚类算法实践

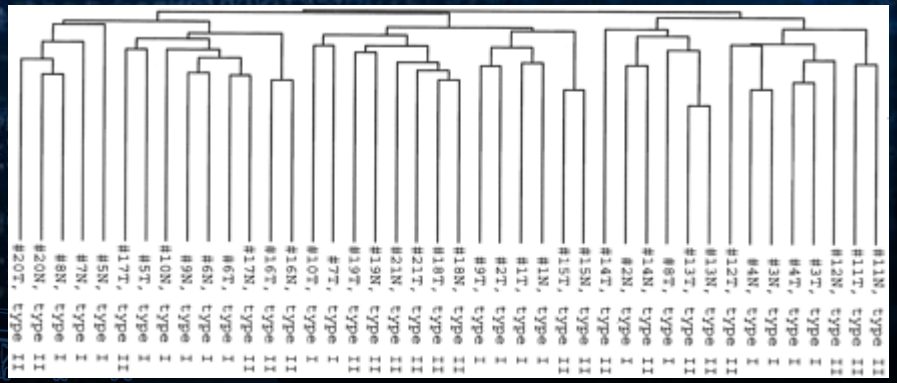
★挑战5：人脸数据聚类

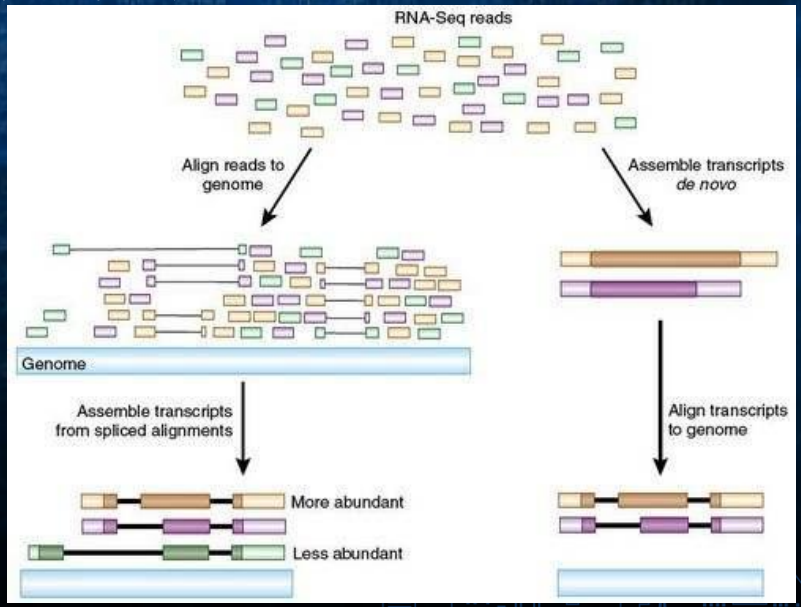
- 人脸聚类。利用LFW dataset数据集 http://viswww.cs.umass.edu/lfw/ 进行聚类。
- 可使用K-Means、SLink、DBScan等算法实现。
- 并比较不同算法的聚类效果和运行速度。

聚类算法实践

★挑战6：生物信息数据聚类

- 对UCI胰腺癌基因数据(https://archive.ics.uci.edu/ml/datasets/gene+expression+cancer+RNA-Seq)进行聚类，要求能够获得最正确的聚类结果。
- 可使用K-Means、SLink、DBScan等算法实现
- 并比较不同算法的聚类效果和运行速度。

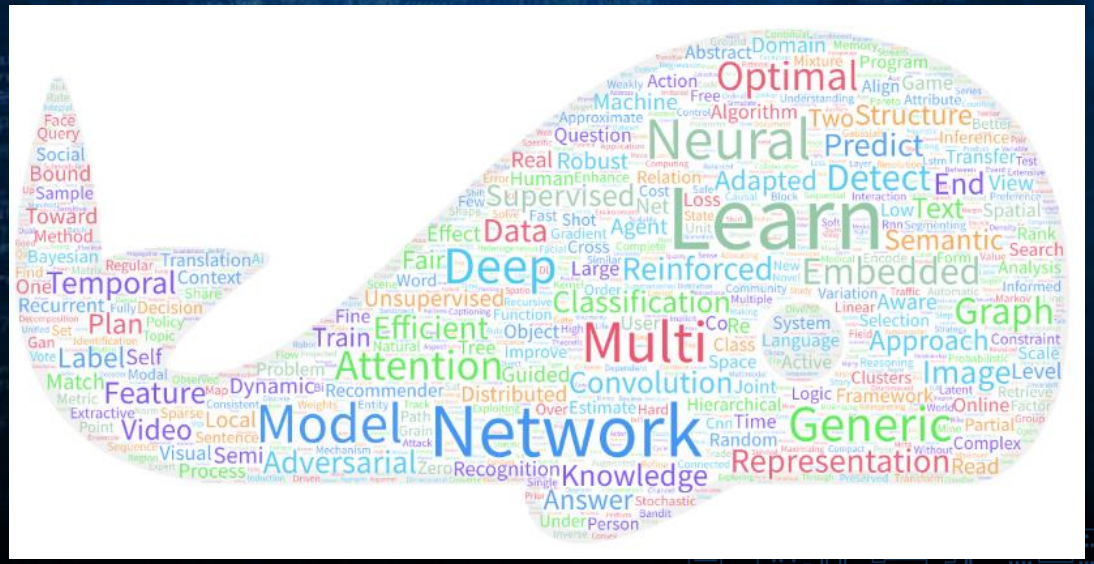




聚类算法实践

★挑战7：文本数据聚类

- 对AAAI2014接收论文和AAAI2019年接收论文(https://archive.ics.uci.edu/ml/datasets/AAAI+2014+Accepted+Papers)分别进行聚类，要求能够获得最正确的聚类结果。
- 可使用K-Means、SLink、DBScan等算法实现，
- 并比较不同算法的聚类效果和运行速度。



实验与实践

- ★ 分类与预测算法实践
- ★ 聚类算法实践
- ★ 优化算法实践

优化算法实践

★挑战：TSP

- 求解货郎担（旅行商）问题，即TSP。要求尽可能快地给出较优结果，最好能优于已知最优解。
- 使用TSPLIB数据。
  - http://comopt.ifi.uni-heidelberg.de/software/TSPLIB95/tsp/
- 可使用**Genetic Algorithm、Ant Colony Optimization 、 Particle Swarm Optimization**等算法实现，
- 并比较不同算法的优化结果和运行速度。

学习资源

- ★请注册华为云账号
- ★华为云官网：www.huaweicloud.com
- ★本课程有教育部-华为智能基座项目支持，为大家提供代金券可在上课期间免费使用华为云资源
- ★华为学习园地
  - https://edu.huaweicloud.com/

感谢聆听
欢迎提问