机器学习作业一

对Adults数据集，做了4个分类器

2015 年 11 月 12 日

小组成员

- 马凌霄
  - 学号：1501111302
  - 院系：信息科学技术学院
  - 邮箱：xysmlx@pku.edu.cn
  - 分工：k-Nearest Neighbors，GPU加速的k-Nearest Neighbors，数据预处理，数据可视化，综合测试对比，实验报告LaTeX排版
- 李奕
  - 学号：1501214394
  - 院系：信息科学技术学院
  - 邮箱：xysmlx@gmail.com
  - 分工：Logistics Regression，Support Vector Machine，Naive Bayes

摘　　要

本文选取

目　　录

§1 背景介绍

1.1 数据介绍：Adult Data Set

本文选取UCI Machine Learning Repository中的Adult数据集[1]。
Adult数据集是根据某人的各种信息预测他的收入是否超过50,000/年。
Adult数据集有48842条记录。Adult数据集的每条记录有以下信息：

- age: continuous.
- workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
- fnlwgt: continuous.
- education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool. education-num: continuous.
- marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
- occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
- relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
- race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
- sex: Female, Male.
- capital-gain: continuous.
- capital-loss: continuous.
- hours-per-week: continuous.
- native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam,

[1] https://archive.ics.uci.edu/ml/datasets/Adult

Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.

1.2 GPU加速机器学习& CUDA介绍

1.3 测试环境

测试使用了两台计算机及其GPU进行测试，硬件配置和编译器版本如表1-5所示。

表 1 计算机1

项目	详细信息
CPU	Core i7-2630QM (2.0GHz, 4 Cores, 6MB L3 Cache)
内存	10GB DDR3 1333MHz
测试所用磁盘	480GB Sandisk Extreme Pro SSD (Read: 550MB/s)
操作系统	Windows 10 Professional x64

表 2 计算机2

项目	详细信息
CPU	Core i5-4460 (3.2GHz, 4 Cores, 6MB L3 Cache)
内存	16GB DDR3 1600MHz
测试所用磁盘	1TB Seagate 7200RPM HDD (Read: 121MB/s)
操作系统	Windows 10 Professional x64

表 3 GPU1

项目	详细信息
型号	nVIDIA GT550M
流处理器	1480 MHz× 96 Cores
显存	2GB DDR3 900MHz
显存位宽	128bit

表 4 GPU2

项目	详细信息
型号	nVIDIA GTX745
流处理器	1033 MHz× 384 Cores
显存	4GB DDR3
显存位宽	128bit

表 5 编译器版本

项目	版本
C/C++	Microsoft Visual Studio 2013
C/C++	GNU C++ 4.8
Python	Python 3.5
GPU	CUDA 7.5

1.4 评价指标

本文选取精确度（Precision）、准确率（Accuracy）、召回率（Recall）、转移性（Specificity）、F-measure这五个指标作为评价指标。

假设原始样本中有两类，其中：

- 总共有P个类别为1的样本，假设类别1为正例。
- 总共有N个类别为0的样本，假设类别0为负例。

经过分类后：

- 有TP个类别为1 的样本被系统正确判定为类别1，FN 个类别为1 的样本被系统误判定为类别0，显然有$P = TP + FN$；
- 有FP 个类别为0 的样本被系统误判断定为类别1，TN 个类别为0 的样本被系统正确判为类别0，显然有$N = FP + TN$；

定义1. 精确度（Precision）：

$$P = \frac{TP}{(TP + FP)}$$

反映了被分类器判定的正例中真正的正例样本的比重。

定义2. 准确率（Accuracy）：

$$A = \frac{(TP + TN)}{(P + N)} = \frac{(TP + TN)}{(TP + FN + FP + TN)}$$

反映了分类器统对整个样本的判定能力——能将正的判定为正，负的判定为负。

定义3. 召回率(Recall)，也称为True Positive Rate:

$$R = \frac{TP}{(TP + FN)} = 1 - \frac{FN}{T}$$

反映了被正确判定的正例占总的正例的比重。

定义4. 转移性（Specificity），也称为True Negative Rate：

$$S = \frac{TN}{(TN + FP)} = 1 - \frac{FP}{N}$$

明显的这个和召回率是对应的指标，只是用它在衡量类别0的判定能力。

定义5. F-measure：

$$F = \frac{2 * \text{召回率} \times \text{准确率}}{(\text{召回率} + \text{准确率})}$$

§ 2 数据处理

2.1 数据预处理

由于仅使用原数据在各个算法中测试的结果均不太好，所以需要对数据进行预处理，根据多次调整数据预处理并且进行实验，得出最佳的数据预处理方法如下所示：

- 删去不完整的记录。
- 删去fnlwgt、education-num、marital-status、capital-gain、capital-loss、native-country。
- 对于hours-per-week
  - 数值≤ 39，标记为0；
  - 数值> 39，标记为1；
- 将age离散化：分为11组分别标记为$0-10$：≤ 20，$21-25$，$26-31$，$32-36$，$37-40$，$41-46$，$47-51$，$52-56$，$57-60$，$61-66$，> 66

经过预处理后，数据集变为：45222条记录，每条记录有8个信息和1个最终标记。

参考文献

[1] Yan, Xifeng, and Jiawei Han. "gspan: Graph-based substructure pattern mining." Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on. IEEE, 2002.

[2] Lin, Wenqing, Xiaokui Xiao, and Gabriel Ghinita. "Large-scale frequent subgraph mining in mapreduce." Data Engineering (ICDE), 2014 IEEE 30th International Conference on. IEEE, 2014.

[3] Kessl, Robert, et al. "Parallel Graph Mining with GPUs." The 3rd International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications. 2014.