

课程设计报告

**姓 名： 王烁**

**学 号： 19851103**

**专 业：计算机科学与技术**

**系 别：计算机与信息技术学院**

**指导教师： 王志海**

**2022年6月**

# 机器学习应用系统与大数据分析

Weka的全名是怀卡托智能分析环境（Waikato Environment for Knowledge Analysis），是一款[免费](https://baike.baidu.com/item/%E5%85%8D%E8%B4%B9/131326" \t "https://baike.baidu.com/item/weka/_blank)的，非商业化（与之对应的是SPSS公司商业数据挖掘产品--Clementine）的，基于Java环境下开源的[机器学习](https://baike.baidu.com/item/%E6%9C%BA%E5%99%A8%E5%AD%A6%E4%B9%A0/217599" \t "https://baike.baidu.com/item/weka/_blank)（machine learning）以及[数据挖掘](https://baike.baidu.com/item/%E6%95%B0%E6%8D%AE%E6%8C%96%E6%8E%98/216477" \t "https://baike.baidu.com/item/weka/_blank)（data mining）软件。它和它的[源代码](https://baike.baidu.com/item/%E6%BA%90%E4%BB%A3%E7%A0%81/3969" \t "https://baike.baidu.com/item/weka/_blank)可在其官方网站下载。有趣的是，该软件的缩写WEKA也是新西兰独有的一种鸟名（[新西兰秧鸡](https://baike.baidu.com/item/%E6%96%B0%E8%A5%BF%E5%85%B0%E7%A7%A7%E9%B8%A1/5991395" \t "https://baike.baidu.com/item/weka/_blank)），而Weka的主要开发者同时恰好来自新西兰的[怀卡托大学](https://baike.baidu.com/item/%E6%80%80%E5%8D%A1%E6%89%98%E5%A4%A7%E5%AD%A6/7082322" \t "https://baike.baidu.com/item/weka/_blank)（The University of Waikato）。

Weka的全名是怀卡托智能分析环境（Waikato Environment for Knowledge Analysis），同时Weka也是新西兰的一种鸟名，而Weka的主要开发者来自新西兰。Weak作为一个公开的数据挖掘工作平台，集合了大量能承担数据挖掘任务的[机器学习算法](https://baike.baidu.com/item/%E6%9C%BA%E5%99%A8%E5%AD%A6%E4%B9%A0%E7%AE%97%E6%B3%95/18635836" \t "https://baike.baidu.com/item/weka/_blank)，包括对数据进行预处理，分类，回归、聚类、[关联规则](https://baike.baidu.com/item/%E5%85%B3%E8%81%94%E8%A7%84%E5%88%99" \t "https://baike.baidu.com/item/weka/_blank)以及在新的交互式界面上的可视化[1]。

## 1 网站简介

Weka的官方地址是http://www.cs.waikato.ac.nz/ml/weka/。点开左侧download栏，可以进入下载页面，里面有windows，mac os，linux等平台下的版本，以windows系统作为示例。目前稳定的版本是3.6[2]。

[Weka](https://so.csdn.net/so/search?q=Weka&spm=1001.2101.3001.7020" \t "https://blog.csdn.net/weixin_30381317/article/details/_blank)系统汇集了最前沿的机器学习算法和数据预处理工具，以便用户能够快速灵活地将已有的据处理方法应用于新的数据集。它为数据挖掘的整个过程提供全面的支持，包括准备输入数据、统计评估学习方案、输入数据和学习效果的可视化。Weka除了提供大量学习算法之外，还提供了适应范围很广的预处理工具，用户通过一个统一界面操作各种组件，比较不同的学习算法，找出能够解决问题的最有效的方法。

Weka系统包括处理标准[数据挖掘](https://so.csdn.net/so/search?q=%E6%95%B0%E6%8D%AE%E6%8C%96%E6%8E%98&spm=1001.2101.3001.7020" \t "https://blog.csdn.net/weixin_30381317/article/details/_blank)问题的所有方法:回归、分类、聚类、关联规则以及属性选择。分析要进行处理的数据是一个重要的环节，Weka提供了很多用于数据可视化和预处理的工具。输入数据可以有两种形式，第一种是以ARFF格式为代表的文件；另一种是直接读取数据库表。

使用Weka的方式主要有三种：第一种是将学习方案应用于某个[数据集](https://so.csdn.net/so/search?q=%E6%95%B0%E6%8D%AE%E9%9B%86&spm=1001.2101.3001.7020" \t "https://blog.csdn.net/weixin_30381317/article/details/_blank)，然后分析其输出，从而更多地了解这些数据；第二种是使用已经学习到的模型对新实例进行预测。第三种是使用多种学习器，然后根据其性能表现选择其中的一种来进行预测。用户使用交互式界面菜单中选择一种学习方法，大部分学习方案都带有可调节的参数，用户可以通过属性列表或对象编辑器修改参数，然后通过同一个评估模块对学习方案的性能进行评估。

Weka主界面称为Weka GUI选择器，它通过右边的四个按钮提供四种主要的应用程序供用户选择。

其中，Weka系统提供的最容易使用的图形用户接口称为探索者。通过选择菜单和填写表单，可以调用Weka所有功能。

### 1.1 数据挖掘和机器学习

数据挖掘和机器学习这两项技术的关系非常密切。机器学习方法构成数据挖掘的核心，绝大多数数据挖掘技术都来自机器学习领域，数据挖掘又向机器学习提出新的要求和任务。

数据挖掘就是在数据中寻找模式的过程。这个寻找过程必须是自动的或半自动的，并且数据总量应该是具有相当大的规模，从中发现的模式必须有意义并能产生一定的效益。通常，数据挖掘需要分析数据库中的数据来解决问题，如客户忠实度分析、市场购物篮分析等。

当今已进入海量数据时代。例如，全世界已经有约1000 000 o00 000个网页;沃尔玛仅一个小时就有一百万的交易量，其数据库里数据已有2.5拍(即2.5×1015)字节的信息，等等。

这些海量数据不可能采用手工方式进行处理，因此，迫切要求能进行数据分析的自动化方法，这些都由机器学习提供。

机器学习分为两种主要类型。第一种称为有监督学习，或称为预测学习，其目标是在给定一系列输入输出实例构成的数据集的条件下，学习输入x到输出y的映射关系。这里的数据集被称为训练集，实例的个数称为训练样本数。第二种机器学习类型称为无监督学习，或称为描述学习，在给定一系列仅由输入实例构成的数据集的条件下，其目标是发现数据中的有趣模式。无监督学习有时候也称为知识发现，这类问题并没有明确定义，因为我们不知道需要寻找什么样的模式，也没有明显的误差度量可供使用。而对于给定的x，有监督学习可以对所观察的值和预测的值进行比较[3]。

### 1.2 预处理

数据挖掘是在大量的、潜在有用的数据中挖掘出有用模式的过程。因此，源数据的质量直接影响到挖掘的效果，高质量的数据是进行有效挖掘的前提。但是，由于数据挖掘所使用的数据往往不是专门为挖掘准备的，期望数据质量完美并不现实，人的错误、测量设备的限制以及数据收集过程的漏洞都可能导致一些问题，如缺失值和离群值。

由于无法在数据的源头控制质量，数据挖掘只能通过以下两个方面设法避免数据质量问题:①数据质量问题的检测与纠正;②使用能容忍低质量数据的算法。第一种方式在数据挖掘前检测并纠正一些质量问题，这个过程称为数据预处理;第二种方式需要提高算法的健壮性。

数据预处理是数据挖掘的重要步骤，数据挖掘者的大部分时间和精力都要花在预处理阶段。Weka专门提供若干过滤器进行预处理，还在探索者界面中提供选择属性标签页专门处理属性的自动选择问题。数据预处理涉及的策略和技术非常广泛，主要包括如下技术。

1)聚集

聚集(Aggregation)就是将两个或多个对象合并为单个对象。一般来说，定量数据通常通过求和或求平均值的方式进行聚集，定性数据通常通过汇总进行聚集。聚集通过数据归约来减少数据量，所导致的较小数据集只需要较少内存和处理时间的开销，因此，可以使用开销更大的数据挖掘算法。另外，聚集使用高层数据视图，起到了范围或度量转换的作用。虽然站在很高的角度去检视问题容易避免只见树木不见森林的问题，但也可能导致有趣细节的丢失。

2)抽样

如果处理全部数据的开销太大，数据预处理可以使用抽样，只选择数据对象的子集进行分析。使用抽样可以压缩数据量，因此，能够使用效果更好但开销较大的数据挖掘算法。由于抽样是一个统计过程，好的抽样方案就是确保以很高的概率得到有代表性的样本，即:样本近似地具有原数据相同的性质。

抽样方式有多种，最简单的抽样是选取每一个数据行作为样本的概率都相同，这称为简单随机抽样，又分为有放回抽样和无放回抽样两种形式，前者是从N个数据行中以概率1/N分别随机抽取出n个数据行，构成样本子集;后者与有放回抽样的过程相似，但每次都要删除原数据集中已经抽取出来的数据行。显然，有放回抽样得到的样本子集有可能重复抽取到相同的数据行。

当整个数据集由差异较大的数据行构成时，简单随机抽样可能无法抽取到不太频繁出现的数据行，这会导致得到的样本不具代表性。分层抽样(Stratified Sampling)尽量利用事先掌握的信息，充分考虑保持样本结构和总体结构的一致性以提高样本的代表性。其步骤是，先将数据集按某种特征分为若干不相交的“层”，然后再从每一层中进行简单随机抽样，从而得到具有代表性的抽样数据子集。

3)维度归约

维度是指数据集中属性的数目。维度归约(Dimension Reduction)是指创建新属性，通过数据编码或数据变换，将一些旧属性合并在一起以降低数据集的维度。

维度归约可以删除不相关的属性并降低噪声，维度降低会使许多数据挖掘的算法变得更好，还能消除了维灾难带来的负面影响。维灾难是指，随着维度的增加，数据在它所占的空间越来越稀疏，对于分类问题，这意味着可能没有足够的数据对象来创建模型;对于聚类问题，点之间的密度和距离的定义失去意义。因此，对于高维数据，许多分类和聚类等学习算法的效果都不理想。维度归约使模型的属性更少，因而可以产生更容易理解的模型。

4)属性选择

除维度归约外，降低维度的另一种方法是仅只使用属性的一个子集。表面看来似乎这种方法可能丢失信息，但很多情况下，数据集存在冗余或不相关的属性。其中，冗余属性是指某个属性包含了其他属性中的部分或全部信息，不相关属性是指对于手头数据挖掘任务几乎完全没有用处的信息。属性选择是指从数据集中选择最具代表性的属性子集，删除冗余或不相关的属性，从而提高数据处理的效率，使模型更容易理解。

最简单的属性选择方法是使用常识或领域知识，以消除一些不相关或冗余属性，但是，选择最佳的属性子集通常需要系统的方法。理想的属性选择方法是:将全部可能的属性子集作为数据挖掘学习算法的输入，然后选取能产生最好结果的子集。这种方法反映了对最终使用的数据挖掘算法的目的和偏爱。但是，由于n个属性的子集的数量多达2n个，大部分情况下行不通。因此，需要考虑三种标准的属性选择方法:嵌入、过滤和包装。

嵌入方法(Embedded Approach)将属性选择作为数据挖掘算法的一部分。在挖掘算法运行期间，算法本身决定使用哪些属性以及忽略哪些属性。决策树算法通常使用这种方法。

过滤方法(Filter Approach)在运行数据挖掘算法之前，使用独立于数据挖掘任务的方法进行属性选择，即:先过滤数据集产生一个属性子集。

包装方法(Wrapper Approach)将学习算法的结果作为评价准则的一部分，使用类似于前文介绍的理想算法，但通常无法枚举出全部可能的子集以找出最佳属性子集。

根据属性选择过程是否需要使用类别信息，属性选择可分为有监督属性选择和无监督属性选择。前者通过度量类别信息与属性之间的相互关系来确定属性子集，后者不使用类别信息，使用聚类方法评估属性的贡献度，根据贡献度来确定属性子集。

5)属性创建

属性创建就是通过对数据集中旧的属性进行处理，创建新的数据集，这样能更有效的获取重要的信息。由于通常新数据集的维度比原数据集少，因此，可以获得维度归约带来的好处。属性创建有三种方法:属性提取、映射数据到新空间和属性构造。

属性提取是指由原始数据创建新的属性集。例如，对照片数据进行处理，提取一些较高层次的特征，诸如与人脸高度相关的边和区域等，就可以使用更多的分类技术。

映射数据到新空间，是指使用一种完全不同的视角挖掘数据可能揭示重要而有趣的特征。例如，对时间序列实施傅立叶变换，转换为频率信息，可能检测到其中的周期模式。

当原始数据集的属性含有必要信息，但其形式不适合数据挖掘算法的时候，可以使用属性构造，将一个或多个原来的属性构造出新的属性。

6)离散化和二元化

有的数据挖掘算法，尤其是某些分类算法，要求数据是分类属性的形式。发现关联模式的算法要求数据是二元属性的形式。因此，需要进行属性变换，将连续属性转换为分类属性称为离散化(Discretization)，将连续和离散属性转换为一个或多个二元属性称为二元化(Binarization)。

连续属性离散化为分类属性分为两个子任务:决定需要多少个分类值，以及如何确定将连续属性值映射到这些分类值中。因此，离散化问题就是决定选择多少个分割点，以及确定分割点的位置。利用少数分类值标签替换连续属性的值，从而减少和简化原来的数据。

根据是否使用类别信息，可以将离散化技术分为两类:使用类别信息的称为有监督的离散化，反之则称为无监督的离散化。

等宽和等频离散化是两种常用的无监督的离散化方法。等宽(Equal Width)离散化将属性的值域划分为相同宽度的区间，区间的数目由用户指定。这种方式常常会造成实例分布不均匀。等频(Equal Frequency)离散化也称为等深(Equal Depth)离散化，它试图将相同数量的对象放进每个区间，区间的数目由用户指定。

7)变量变换

变量变换(Variable Transformation)也称为属性变换，是指用于变量的所有值的变换。下面讨论两种重要的变量变换:简单函数变换和规范化。

简单函数变换是使用一个简单数学函数分别作用于每一个值。在统计学中，使用平方根、对数变换和倒数变换等变量变换常用于将不具有高斯分布的数据变换为具有高斯分布的数据。

变量的标准化(Standardization)是使整个值的集合具有特定的性质。

## 2 软件系统功能

Weka系统汇集了前沿的机器学习算法和数据预处理工具，以便用户能够快速灵活地将已有的成熟处理方法应用于新的数据集。它为数据挖掘的整个过程提供全面的支持，包括准备输入数据、统计评估学习方案、输入数据和学习效果的可视化。Weka除了提供大量学习算法之外，还提供了适应范围很广的预处理工具，用户通过一个统一界面操作各种组件，比较不同的学习算法，找出能够解决问题的最有效的方法。

1、处理方法

包括处理标准数据挖掘问题的所有方法：回归、分类、聚类、关联规则和属性选择。

2、输入数据

通过以ARFF格式为代表的文件进行输入

直接读取数据库表

3、Weka主界面Weka GUI Chooser

（1）探索者（Explorer）

Weka系统提供的最容易使用的图形用户接口称为探索者(Explorer)。虽过选择菜单和填写表单，可以调用Weka的所有功能。例如，用户用鼠标仅仅单击几个按钮，就可以完成从ARFF文件中读取数据集，然后建立决策树的工作。Weka界面十分友好，能适时地将不宜用的功能选项设置为不可选;将用户选项设计为表格方式以方便填写;当鼠标移动到界面工具上短暂停留时，会给出用法提示;对算法都给出较为合理的默认值，这样，帮助用户尽量少花精力进行配置就可取得较好的效果等。

虽然探索者界面使用很方便，但它也存在一个缺陷，要求它将所需数据全部一次读进内存，一旦用户打开某个数据集，就会读取全部数据。因此，这种批量方式仅适合处理中小规模的问题。知识流刚好能够弥补这一缺陷。

（2）实验者（Experimenter）

实验者(Experimenter)界面用于帮助用户解答实际应用分类和回归技术中遇到的一个基本问题——对于一个已知问题，哪种方法及参数值能够取得最佳效果？通过Weka提供的实验者工作环境，用户可以比较不同的学习方案。尽管探索者界面也能通过交互完成这样的功能，但通过实验者界面，用户可以让处理过程实现自动化。实验者界面更加容易使用不同参数去设置分类器和过滤器，使之运行在一组数据集中，收集性能统计数据，实现重要的测试实验。

（3）知识流（KnowledgeFlow）

知识流(KnowledgeFlow)界面可以使用增量方式的算法来处理大型数据集,用户可以定制处理数据流的方式和顺序。知识流界面允许用户在屏幕上任意拖曳代表学习算法和数据源的图形构件，并以一定的方式和顺序组合在一起。也就是，按照一定顺序将代表数据源、预处理工具、学习算法、评估手段和可视化模块的各构件组合在一起，形成数据流。如果用户选取的过滤器和学习算法具有增量学习功能，那就可以实现大型数据集的增量分批读取和处理。

（4）简单命令行（Simple CLI）

简单命令行(Simple CLl)界面是为不提供自己的命令行界面的操作系统提供的，该简单命令行界面用于和用户进行交互，可以直接执行Weka命令[3]。

## 3 数据格式

根据应用的不同，数据挖掘对象可以是各种各样的数据，这些数据可以以各种形式存储，如数据库、数据仓库、数据文件、流数据、多媒体、网页，等等。即可集中存储在数据存储库中，也可以分布在世界各地的网络服务器上。

通常将数据集视为待处理的数据对象的集合。由于历史原因，数据对象有多个别名，如记录、点、行、向量、案例、样本、观测等。数据对象也是对象，因此，可以用刻画对象基本特征属性来进行描述。属性也有多个别名，如变量、特征、字段、维、列，等等。

数据集可以类似于一个二维电子表格或数据库表。在最简单的情形下，每个训练输入Xi也是一个N维的数值向量，表示特定事物的一些特征，如人的身高、体重。这些特征也可以称为属性，有时Xi也可以是复杂结构的对象，如图像、电子邮件、时间序列、语句等。

属性可以分为四种类型:标称、序数、区间和比率，其中，标称属性的值仅仅是不同的名称，即，标称值提供区分对象的足够信息，如性别、衣服颜色、天气等；序数属性的值可以提供给确定对象的顺序的足够信息，如成绩等级、职称、学生等；区间属性的值之间的差是有意义的，即存在测量单位，如温度、日历日期等；比率属性的值之间的差和比值都是有意义的，如绝对温度、年龄、长度、成绩分数等。

标称属性和序数属性统称为分类的或定性的属性，它们的取值为集合，即使使用数值来表示，也不具备的大部分性质，因此，应该像对待符号一样对待；区间属性和比率属性统称为定量的或数值的属性，定量属性采用数值来表示，具备数的大部分性质，可以使用整数值或连续值来表示。

大部分数据集都以数据库表和数据文件的形式存在，Weka支持读取数据库表和多种格式的数据文件，其中，使用最多的是一种称为ARFF格式的文件。

### 3.1 ARFF格式

ARFF是一种Weka专用的文件格式，由Andrew Donkin创立，有传言说ARFF代表Andrew's Ridiculous File Format(安德鲁的荒唐文件格式)，但在Weka的正式文档中明确说明ARFF代表Attribute-Relation File Format(属性—关系文件格式)。该文件是ASCII文本文件，描述共享一组属性结构的实例列表，由独立且无序的实例组成，是Weka表示数据集的标准方法，ARFF不涉及实例之间的关系。

在Weka安装目录下的data子目录中，可以找到名称为weather.numeric.arff的天气数据文件，其内容如程序清单1.1所示。数据集是实例的集合，每个实例包含一定的属性，属性的数据类型包括如下几类:标称型(nominal)只能取预定义值列表中的一个;数字型(numeric)，只能是实数或整数;字符串(string)，这是一个由双引号引用的任意长度的字符列表;另外还有日期型(date)和关系型(relational)。ARFF文件就是实例类型的外部表示，其中包括一个标题头(header)，以描述属性的类型，还包含一个用逗号分隔的列表所表示的数据部分(data)。

天气数据的ARFF文件：

@relation weather.symbolic

@attribute outlook {sunny, overcast, rainy}

@attribute temperature {hot, mild, cool}

@attribute humidity {high, normal}

@attribute windy {TRUE, FALSE}

@attribute play {yes, no}

@data

sunny,hot,high,FALSE,no

sunny,hot,high,TRUE,no

overcast,hot,high,FALSE,yes

rainy,mild,high,FALSE,yes

rainy,cool,normal,FALSE,yes

rainy,cool,normal,TRUE,no

overcast,cool,normal,TRUE,yes

sunny,mild,high,FALSE,no

sunny,cool,normal,FALSE,yes

rainy,mild,normal,FALSE,yes

sunny,mild,normal,TRUE,yes

overcast,mild,high,TRUE,yes

overcast,hot,normal,FALSE,yes

rainy,mild,high,TRUE,no

表1-1 weather数据集合

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | outlook | temperature | humidity | windy | play |
| 1 | sunny | hot | high | FALSE | no |
| 2 | sunny | hot | high | TRUE | no |
| 3 | overcast | hot | high | FALSE | yes |
| 4 | rainy | mild | high | FALSE | yes |
| 5 | rainy | cool | normal | FALSE | yes |
| 6 | rainy | cool | normal | TRUE | no |
| 7 | overcast | cool | normal | TRUE | yes |
| 8 | sunny | mild | high | FALSE | no |
| 9 | sunny | cool | normal | FALSE | yes |
| 10 | rainy | mild | normal | FALSE | yes |
| 11 | sunny | mild | normal | TRUE | yes |
| 12 | overcast | mild | high | TRUE | yes |
| 13 | overcast | hot | normal | FALSE | yes |
| 14 | rainy | mild | high | TRUE | no |

上述代码中，以百分号“%”开始的行称为注释行。与计算机编程语言类似，最前面的注释行应该写明数据集的来源、用途和含义。

@relation一行定义内部数据集的名称——weather，名称应简洁明了，尽可能容易理解。Relation也称为关系。

attribute outlook {sunny, overcast, rainy}行定义名称为outlook的标称型属性，有三个取值: sunny、overcast和rainy。按照同样的方式，@attribute windy {TRUE,FALSE行和@attribute play {yes, no}行分别定义windy和play两个标称型属性。要注意的是，最后一个属性缺省为用于预测的类别变量。本例中，类别变量为标称型属性play，它只能取两个值之一，使得天气问题成为二元(binary)的分类问题。

@attribute temperature real定义名称为temperature的数值型属性，@attribute humidity real定义名称为humidity的数值型属性。这两个属性都是实数型。

@data标志后的各行构成数据集。每行为一个实例样本，由采用逗号分隔的值组成，顺序与由@attribute所定义属性的顺序一致。

### 3.2 CSV格式

逗号分隔值(Comma-Separated Values，CSV，有时也称为字符分隔值，因为分隔字符也可以不是逗号)：其文件以纯文本形式存储表格数据(数字和文本)。纯文本意味着该文件是一个字符序列，不含必须象二进制数字那样被解读的数据。CSV文件由任意数目的记录组成，记录间以某种换行符分隔；每条记录由字段组成，字段间的分隔符是其它字符或字符串，最常见的是逗号或制表符。通常，所有记录都有完全相同的字段序列。

CSV文件格式的通用标准并不存在，但是在RFC 4180中有基础性的描述。使用的字符编码同样没有被指定，但是7-bit ASCII是最基本的通用编码。

CSV是一种通用的、相对简单的文件格式，被用户、商业和科学广泛应用。最广泛的应用是在程序之间转移表格数据，而这些程序本身是在不兼容的格式上进行操作的(往往是私有的和/或无规范的格式)。因为大量程序都支持某种CSV变体，至少是作为一种可选择的输入/输出格式。

“CSV”并不是一种单一的、定义明确的格式(尽管RFC 4180有一个被通常使用的定义)。因此在实践中，术语”CSV”泛指具有以下特征的任何文件：

(1)、纯文本，使用某个字符集，比如ASCII、Unicode、EBCDIC或GB2312(简体中文)等；

(2)、由记录组成(典型的是每行一条记录)；

(3)、每条记录被分隔符分隔为字段(典型分隔符有逗号、分号或制表符；有时分隔符可以包括可选的空格)；

(4)、每条记录都有同样的字段序列。

在这些常规的约束条件下，存在着许多CSV变体，故CSV文件并不完全互通。然而，这些变异非常小，并且有许多应用程序允许用户预览文件(这是可行的，因为它是纯文本)，然后指定分隔符、转义规则等。如果一个特定CSV文件的变异过大，超出了特定接收程序的支持范围，那么可行的做法往往是人工检查并编辑文件，或通过简单的程序来修复问题。因此在实践中，CSV文件还是非常方便的。

CSV格式最好被用来表现记录集合或序列，其中的每条记录都有完全相同的字段序列。CSV格式没有被限定于某个特定字符集。不管用Unicode还是用ASCII，都没有问题(尽管特定程序支持的CSV可能会有它们自己的局限性)。甚至从一个字符集翻译到另一个字符集，CSV文件都不会有问题(不象几乎所有的私有数据格式)。然而，CSV不提供任何途径来表明使用的是什么字符集。

“CSV”格式中大量变体的存在说明并没有一个”CSV标准”。在常见用法中，几乎任何定界符分隔的文本数据都可以被统称为”CSV”文件。不同的CSV格式可能不会兼容[5]。

## 参考文献

1. 百度百科, URL: http://baidu.com
2. 简书,URL: http://jianshu.com

[3] CSDN博客, URL: https://blog.csdn.net/

[4] 爱码网, URL: https://likecs.com

[5] 维基百科,URL:http://zh.wikipedia.org