机器学习算法之聚类算法

继上一篇《机器学习之分类算法》中大致梳理了一遍在机器学习中常用的分类算法，类似的，这一姊妹篇中将会梳理一遍机器学习中的聚类算法，最后也会拓展一些其他无监督学习的方法供了解学习。

1. 机器学习

> 机器学习是近20多年兴起的一门多领域交叉学科，它涉及到概率论、统计学、计算机科学以及软件工程等多门学科。机器学习理论主要是设计和分析一些让计算机可以自动“学习”的算法。机器学习算法是一类能从数据中自动分析获得规律，并利用规律对未知数据进行预测的算法 [1]。

机器学习（Machine Learning）是人工智能（AI）中很重要的一部分，因为在目前的实践过程中，大多数人工智能问题是由机器学习的方式实现的。所以说机器学习是实现人工智能（AI）的一个途径，即以机器学习的手段解决人工智能中的问题。它可以被设计用程序和算法自动学习并进行自我优化，同时，需要一定数量的训练数据集（traing dataset）来构建过往经验“知识”。

目前机器学习已经广泛应用于数据挖掘、计算机视觉、自然语言处理、语音和手写识别、生物特征识别、医学诊断、检测信用卡欺诈、证券市场分析、搜索引擎、DNA序列测序、无人驾驶、机器人等领域。

2.机器学习方法

机器学习算法有很多，有分类、回归、聚类、推荐、图像识别领域等等，具体算法比如线性回归、逻辑回归、朴素贝叶斯、随机森林、支持向量机、神经网络等等。在机器学习算法中，没有最好的算法，只有“更适合”解决当前任务的算法。

机器学习算法的分类方式有很多种，

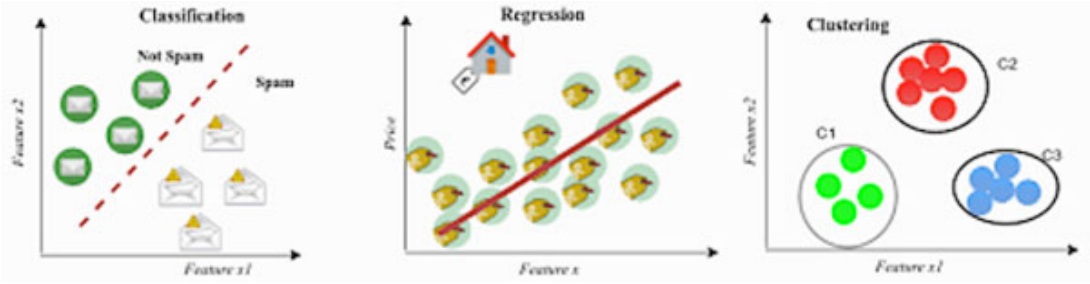
如果按照学习方式分类可分为

学习方式	英文	描述
监督式学习	Supervised Learning	训练集目标:有标注; 如回归分析，统计分类
无监督式学习	Unsupervised Leanring	训练集目标:无标注;如聚类、GAN(生成对抗网络)
半监督式学习	Semi-supervised Leanring	介于监督式与无监督式之间
增强学习	Reinforcement	智能体不断与环境进行交互，通过试错的方式来

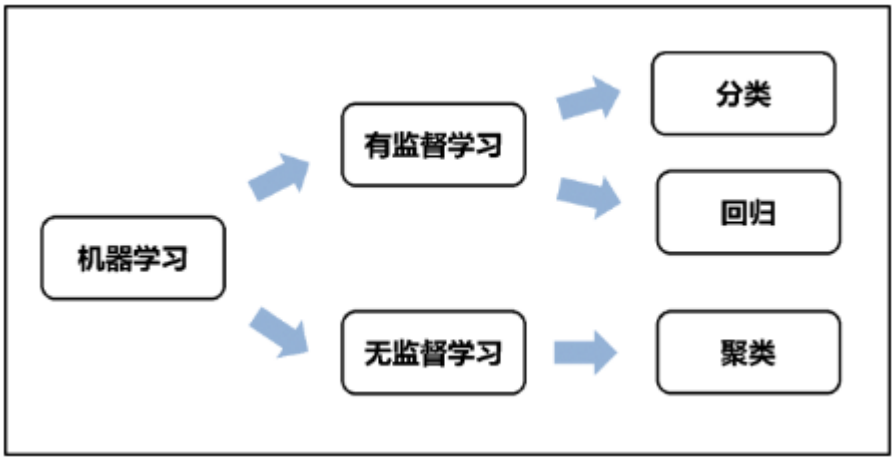
学习方式	英文	描述
	Leanring	获得最佳策略

如果按照学习任务分类可分为以下三类

学习任务	英文	描述	下图例子[2]
分类	Classification	分类是预测一个标签 （是离散的），属于监督学习	垃圾邮件分类
回归	Regression	回归是预测一个数量 （是连续的），属于监督学习	房价走势预测
聚类	Clustering	属于无监督学习	数据分组



- 它们之间的关系如下图所示 [2]



3.无监督学习

卷积神经网络之父Yann LeCun在2016年的CMU大学的演讲上曾提出***AI研究的下一站是无监督学习，因为它代表了AI技术的未来***。人类和动物通过观察世界、星宇和理解自然规律来获得尝试，要想机器也学会这么做，无监督学习是赋予机器常识的关键。所以，无监督学习的重要性不言而喻。

在“无监督学习中”，它能够自己进行学习，不需要被显示的告知他们所做的一切是否正确。因为其算法所用的训练样本的标记信息是未知的（即没有标注的过的），样本中只给出了输入变量（自变量X）而没有给出输出变量（因变量y）。所以无监督学习将通过无标记训练样本的学习来揭示数据的内在性质及规律，为进一步的数据分析提供基础。在此类学习任务中研究最多、应用最广的是“聚类”（Clustering）[4]。类似样本数据包括视频和文字。

4.聚类

无监督式学习方法包含“聚类”(Clustering)与“降维”（Dimension Reduction）。其中，研究最多、应用最广的是“聚类”，这篇入门教程也主要介绍无监督学习的聚类任务。聚类试图将训练集中的样本数据划分为若干个(k个)通常互不相交的子集，每个子集称为一个“簇”（cluster）。达到“物以类聚”效果，即**簇与簇之间的相似度低(low intra-cluster similarity)，簇内相似度高(high inter-cluster similarity)**。可以按照数据的相似度(similarity)和距离(disctance)来聚类(clustering)划分成不同的簇，这样每个簇可能会对应一些潜在的概念或类别，如“浅色瓜”、“深色瓜”、“无籽瓜"、“有籽瓜”，甚至“本地瓜”、“外地瓜”等。但需要说明的是，这些概念对聚类算法而言事先是未知的，**聚类过程仅能自动形成簇结构，簇所对应的概念语义需要由使用者来把握和命名。**[4]

簇的特征总结：

- 簇不是事先给定的，而是根据数据的相似性和距离来划分
- 簇的数目和结构都没有事先假定

> 聚类既能作为一个单独的过程来**寻找数据中的内在分布结构**，也能**作为分类等其他学习任务的前驱过程。**

比如，在一些商业应用中需要对用户定义“用户类型”，但这对于商家来说通常是不容易的，这时如果先对用户数据使用聚类从而将用户划分成若干个簇，每个簇即为一个类，然后再基于这些类训练分类模型来判别新用户的类型，这样能更加高效有效的进行分类任务。

同样的聚类分析也广泛应用于一些探索性领域，如统计学与模式分析、决策支持、WEB挖掘、网络安全、地质勘探、心理学、考古学等。

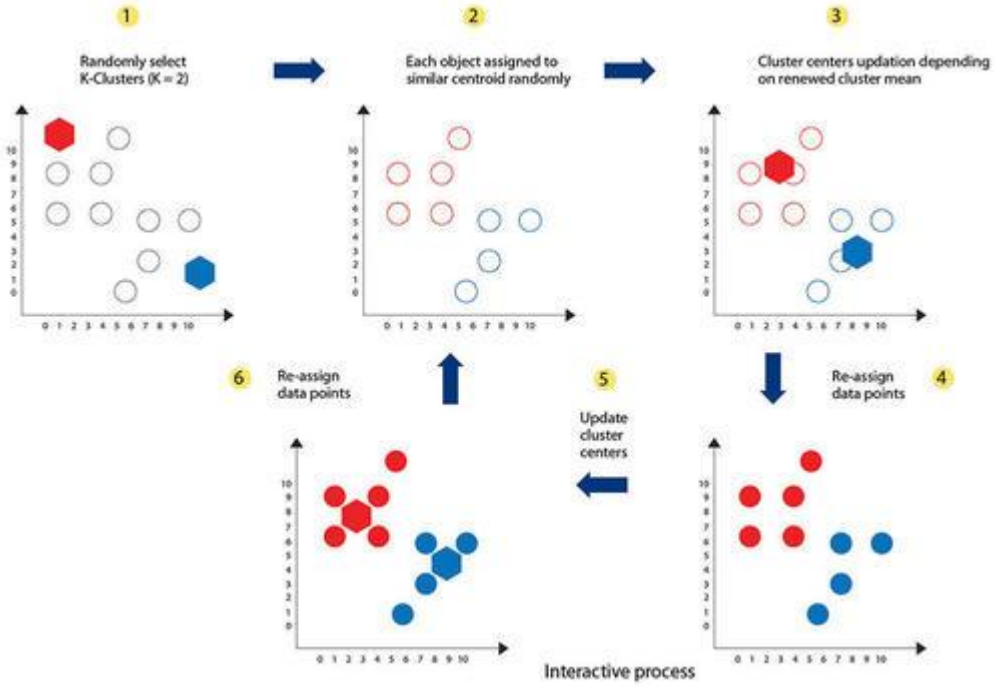
4.1.常见的聚类算法

4.1.1. 原型聚类（prototype-based clustering）

- K均值聚类

K均值聚类是一种“基于原型的聚类”（prototype-based clustering），在现实聚类任务中极为常用。K均值聚类将训练集分成k个簇内相似度高、簇间相似度低的样本聚类。

1. 定义K个质心（centre_id）,这在一开始可以初始化为随机的，也可以从数据集中任选k个对象作为初始簇中心。
2. 将每个训练样本基于其到质心的距离分配到最近的质心所代表的簇 $cluster_i, i \in (1, k)$
3. 重新计算所有簇的质心，将每个质心更新为当前$cluster_i$中所有训练样本点的均值
   不断重复步骤2与3直到收敛（即质心不再发生变化）。



科赛社区中的K-Means实战：NBA控卫聚类——K-Means详解

- 学习向量量化（Learning Vector Quantization）

与K均值算法类似，是找到一组原型向量来聚类，每一个原型向量代表一个簇，将空间划分类若干个簇。不同的是**LVQ假设数据样本带有类别标记**，每个样本X_i，有类别标记y_i，这些类别可以用来辅助聚类。

- 高斯混合聚类 (Mixture of Gaussian)

高斯混合聚类使用了一个很流行的算法：GMM(Gaussian Mixture Model)。高斯混合聚类与k均值聚类类似，但是采用了**概率模型**来表达聚类原型。每个高斯模型（Gaussian Model）就代表了一个簇（类）。GMM是单一高斯概率密度函数的延伸，能够平滑地近似任意形状的密度分布。在高斯混合聚类中，每个GMM会由k个高斯模型分布组成，每个高斯模型被称为一个component，这些component线性加乘在一起就组成了GMM的。

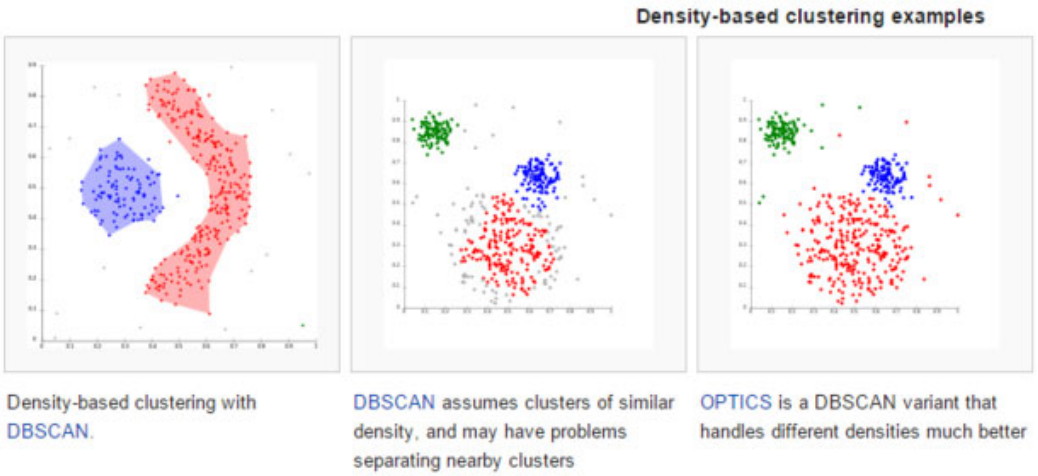
简单地来说，k-Means的结果是每个数据点没分配到其中某一个cluster,而GMM则给出的是这个数据点被分配到每个cluster的概率，又称为soft assignment。

4.1.2. 密度聚类（density-based clustering）之DBSCAN聚类

> 基于密度聚类的方法通常将簇看做是数据空间中被低密度区域（代表噪音）分隔开的稠密对象区域。

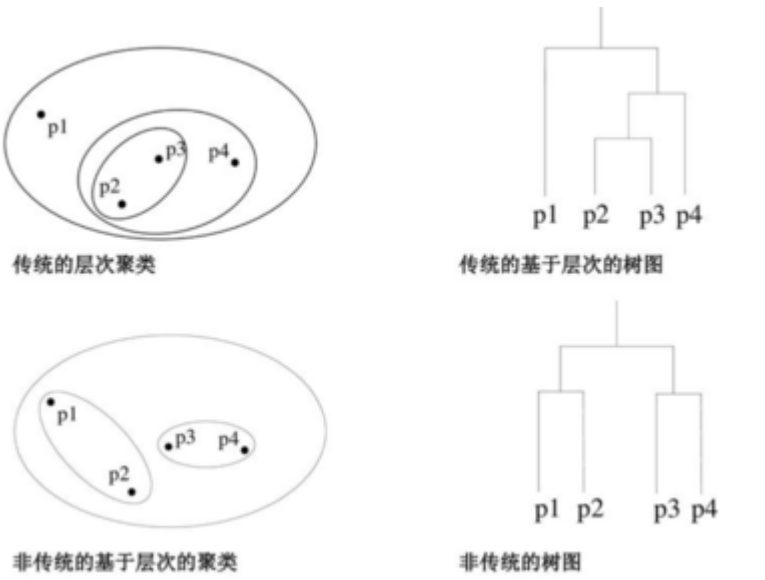
DBSCAN(Density-Based Spatial Clustering of Applications with Noise)聚类是一种著名的密度聚类算法，它基于一组“邻域”参数来刻画样本分布的紧密程度。该算法将具有足够高密度的区域划分为簇，并在具有噪音的空间数据库中发现任意形状的簇，它将簇定义为密度相连的点的最大的集合[5]。它将点的密度将点分为3类：a.稠密区域内部的点（核心点），b.稠密区域边缘上的点（边界点），c.洗漱去与众的点（噪音或背景点）。

OPTICS(Ordering Points to Identify the Clustering Structure)也是一种典型的基于密度的聚类方法，是DBSCAN的变种，对于不同密度能够更好地处理。



4.1.3. 层次聚类（hierarchical clustering）

层次聚类，顾名思义，是一种能够构建有层次的簇的算法。层次聚类视图在不同层次对数据集进行划分，从而形成**树形**的聚类结构。数据集的划分可采用**“自底向上”的聚合策略（或凝聚层次聚类）**，也可以采用**“自顶向下”的分拆策略（或分裂层次聚类）**。

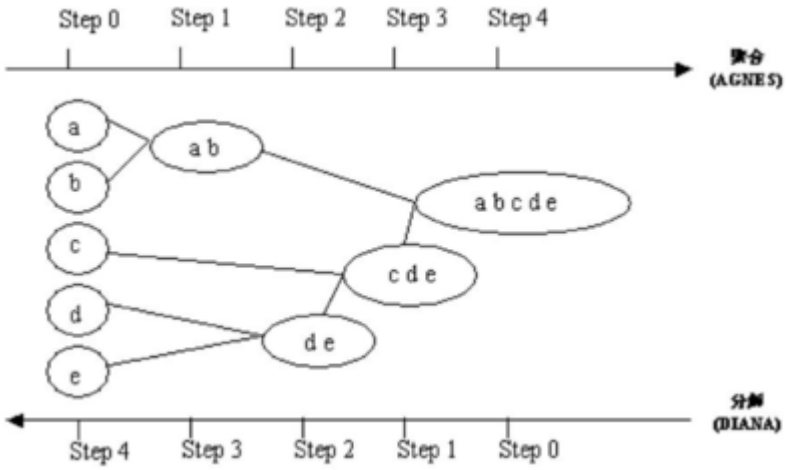


自底向上的聚合层次聚类方法（或凝聚层次聚类）

简单来说，在这个算法的起始阶段，每个数据点都是一个簇。接着两个接近的簇合二为一，构造成越来越大的簇。最终当所有点都被合并到一个簇中时，或者满足一定条件时算法停止。绝大多数的层次聚类方法属于这一类，知识簇间的相似度定义有所不同。算法包括CURE、ROCK。

自顶向下的分解层次聚类方法（或分裂层次聚类）

这种策略作法与自底向上的策略作法相反。它在起始阶段现将所有数据点看成一个簇，然后将其不断分解至数目越来越多的**小簇**，知道所有数据点独自构成一个簇，或者满足一定条件时算法停止。算法包括BIRCH。

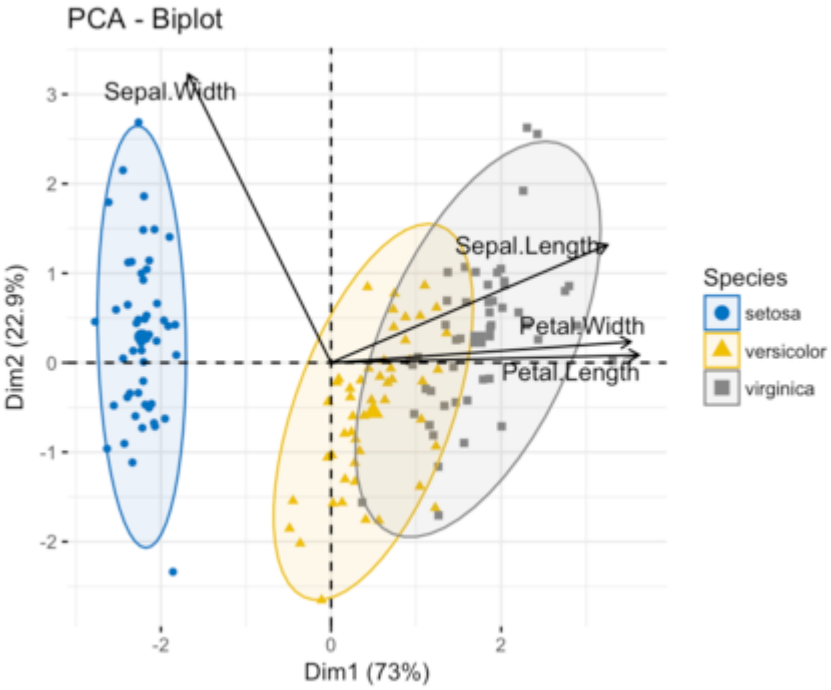


5.其他无监督学习方法

5.1. 主成成分分析（Principal Component Analysis,简称PCA）

降维也是无监督学习的一种，其中最常用的一种方法就是**主成成分分析**。它提供了一种压缩数据的方式，可以将数据变化为元素之间彼此不相关,最大化保持数据的内在信息，并通过衡量在投影方向上的数据方差的大小来衡量该放下的重要性。主要用于提取数据的主要特征分量，常用于高维数据的降维。因为在原始的高维空间中，会存在冗余信息及噪音信息，会在实际应用中造成误差降低准确率（模型过拟合）。通过将原高维空间中的数据点映射到低维度空间中，可以减少冗余信息造成的误差，提高精度。同时也可以加速后续的计算速度。

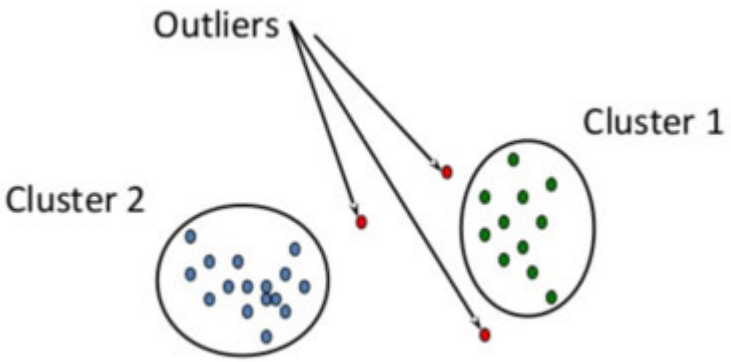
如下图所示在鸢尾花数据集上的特征降维[3]，每个元尾花数据有4个属性，可以通过PCA降维，将4维属性降为2维。第一主成分（PC1）即最大的特征值是Speal.Width,第二主成分Speal.Length,以此类推。



5.2. 异常值检测(Anomaly Detection或Outlier Detection)

> 异常值检测常借助聚类或距离计算进行，如将远离所有簇中心的样本作为异常点，或者将密度极低处的样本作为异常点。最近有研究提出基于“隔离性”（isolation）可快速检测出异常点。[4]

异常检测算法具有少量的异常样本和大量的正常样本，常应用于诈骗是吧、工业零件问题检测等。



6. 参考文献

[1] https://zh.wikipedia.org/wiki/%E6%9C%BA%E5%99%A8%E5%AD%A6%E4%B9%A0

[2] https://www.slideshare.net/SebastianRaschka/nextgen-talk-022015/18-Generative_ClassiersNaive_BayesBayes_Theorem_Pj

[3] http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/112-pca-principal-component-analysis-essentials/

[4] 周志华. 机器学习.清华大学出版社，2016

[5] http://slidesplayer.com/slide/11514145/