Statistics and Application 统计学与应用, 2021, 10(1), 31-46
Published Online February 2021 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2021.101004

Hans汉斯

红酒品种聚类分析

杨　帆，苏理云

重庆理工大学理学院，重庆
Email: cloudhopping@163.com, 18580427076@163.com

收稿日期：2021年1月10日；录用日期：2021年2月12日；发布日期：2021年2月19日

摘　要

本文根据UCI红酒化学成分数据，指标为：红酒品牌、酒精浓度、苹果酸含量、灰度、灰的碱度、镁含量、总酚类化合物量、类黄酮量、原花青素年、颜色强度、色调、稀释酒、脯氨酸，进行红酒品种的聚类分析。本文基于K-means聚类法、层次聚类法及Dbscan聚类法，对红酒化学成分数据进行聚类分析。首先对数据进行筛选，并选择适合的、有意义的，且适合聚类的指标，然后进行数据预处理，最后通过R语言实现红酒品种的聚类。通过聚类结果进行解释，给出红酒的品种类别与质量的好坏。

关键词

红酒品种，K-均值聚类算法，层次聚类法，Dbscan聚类算法

Cluster Analysis of Red Wine Varieties

Fan Yang, Liyun Su

Chongqing University of Technology, Chongqing
Email: cloudhopping@163.com, 18580427076@163.com

Received: Jan. 10th, 2021; accepted: Feb. 12th, 2021; published: Feb. 19th, 2021

Abstract

In this paper, according to the chemical composition data of UCI red wine, the indicators are: red wine brand, alcohol concentration, malic acid content, gray scale, alkalinity of gray, magnesium content, total phenolic compounds amount, flavonoid amount, procyanidin year, color intensity, hue, diluted wine and proline, to conduct cluster analysis of red wine varieties. Based on K-means clustering method, hierarchical clustering method and Dbscan clustering method, this paper conducts clustering analysis on the chemical composition data of red wine. First, the data were screened and appropriate, meaningful and clustering indexes were selected. Then, data pretreat-

文章引用: 杨帆, 苏理云. 红酒品种聚类分析[J]. 统计学与应用, 2021, 10(1): 31-46.
DOI: 10.12677/sa.2021.101004

ment was carried out. Finally, the clustering of red wine varieties was realized through R language. The classification and quality of red wine were explained by clustering results.

Keywords

Red Wine Varieties, K-Mean Clustering Algorithm, Hierarchical Clustering Method, Dbscan Clustering Algorithm

1. 引言

红酒是经发酵而成的含有多种营养成分的饮料酒，是对人体有益的健康酒精饮品，受到全世界人民的喜爱，红酒的质量也受到高度关注。红酒的原料是新鲜葡萄或葡萄汁，除了原料的好坏，红酒所蕴含的矿物质、营养元素，酒的色泽，酸度，酒精含量都是红酒品质的重要衡量标准[1]。长期以来，红酒质量的评定都是通过高资质的品酒员对红酒各项指标打分并求和来评定的。但在以品酒员为主导的评定方式中，品酒员的感官指标具有很强的主观因素，会导致不同的人员来进行红酒质量的评定，评定结果存在较大差异的不确定性。因此，建立合理、客观的红酒品种与质量评定方法是十分重要的。

红酒所选取的原料与所酿红酒的质量直接相关，把红酒及原料进行化学成分分析后所得到的理化指标数据会在一定程度上客观地反映红酒的品种与质量。所以，可以根据红酒化学成分理化指标数据来进行分析，可得到客观、全面的评定方式。把红酒化学成分理化指标变量处理与简化为红酒质量关系密切的少数指标，考虑采用聚类的方法对处理后的红酒化学成分理化指标数据进行分析，可以客观地得到红酒品种类别，通过不同类别的特征指标就可客观地评定该类别的红酒质量。

聚类是一种无监督模式识别方法，即在没有任何先验信息的指导下，从一个数据集中发现潜在的相似模式，对数据集进行分组，以使得同一类内的相似性尽可能大，同时不同类之间的差异性尽可能大[2]，是数据挖掘和人工智能领域的重要研究内容。目前主要的聚类算法按方法类型划分主要有划分的方法、层次的方法、基于密度的方法、基于网络的方法、基于模型的方法；聚类分析已被广泛应用于统计学、机器学习、空间数据库、生物学以及市场营销等领域[3]。

本文以数据为 UCI (University of California Irvine)的红酒化学成分数据为基础，数据指标为：红酒品牌、酒精浓度、苹果酸含量、灰度、灰的碱度、镁含量、总酚类化合物量、类黄酮量、原花青素年、颜色强度、色调、稀释酒、脯氨酸。首先阐述了聚类分析的概念，并介绍 K-means 聚类方法、层次聚类方法、DBSCAN 聚类算法。接着对红酒化学成分数据变量进行选择与简化筛选出与红酒质量关系密切的少数指标，在变量指标选定后对原始数据进行预处理。通过 R 软件运用 K-means 聚类方法、层次聚类方法、DBSCAN 聚类算法进行聚类分析，得到红酒的品种类别与质量的好坏并评价聚类结果的好坏。

2. 聚类算法

2.1. 聚类算法的基本概念

聚类分析(Clustering Analysis)是数据挖掘中的一个重要研究领域。它是一种无监督的学习方法，它通过一定的规则将数据按照定义的相似性划分为若干个类或簇，这些类或簇是由许多在性质上相似的数据

点构成的[4]。同一个类中的数据彼此相似，而与其它类中的数据相异。聚类分析已被广泛应用于统计学、机器学习、空间数据库、生物学以及市场营销等领域，例如在商务领域中，聚类分析能够帮助市场分析人员来更清楚的认识市场，对一些不同的消费者群体进行有效的划分，进而从中发现不同的购买能力，寻求新的潜在的市场；在生物学领域，聚类分析可以通过生物数据库中的数据对不同的物种进行分类或对基因进行学习，来认知不同种群的内在结构；在地理信息系统中，聚类分析可以通过对遥感器传回的数据进行分析，确定河流、街道的位置和结构，并提取有用的信息加以利用[5]；在因特网上，聚类分析可以通过对数据的聚类来批量处理同种数据行为，修改或修复某种数据文档的有用信息；在医学领域，聚类分析可以通过对某种病毒或细菌的分类来自动确定某种病症[6]。

聚类是一个无监督的学习过程，它同分类的根本区别在于：分类是需要事先知道所依据对象的类别特征，而聚类是要找到这个对象的类别特征，因此，在很多应用中，聚类分析作为一种数据预处理过程，是进一步分析和处理数据的基础[7]。一个能产生高质量聚类的算法必须满足下面两个条件：类内(intra-class)数据或对象的相似性最强，以紧致度描述；类间(inter-class)数据或对象的相似性最弱，以分离度描述。聚类质量的高低通常取决于聚类算法所使用的相似性测量的方法和实现方式，同时也取决于该算法能否发现部分或全部隐藏的数据的模式。

2.2. K-means 聚类算法

K-均值聚类算法是聚类分析中的一种基本划分式方法。在一九六七年由麦克奎因(MacQueen)首次提出[8]。由于其算法简便易懂，且在计算速度上具有无可比拟的优势，通常被作为大样本聚类分析的首选方案。因此成为了最大众化的聚类方法之一而被广泛应用。K-Means 聚类算法是一种常用的基于划分的聚类分析方法，该聚类算法的最终目标就是根据输入参数 k (这里的 k 表示需要将数据对象聚成几簇)，然后把数据对象分成 k 个簇。该算法的基本思想：首先指定需要划分的簇的个数 k 值；然后随机地选择 k 个初始数据对象点作为初始的聚类中心；第三，计算其余的各个数据对象到这个初始聚类中心的距离(这里一般采用距离作为相似性度量)，把数据对象划归到距离它最近的那个中心所处在的簇类中；最后，调整新类并且重新计算出新类的中心，如果两次计算出来的聚类中心未曾发生任何的变化，那么就可以说明数据对象的调整已经结束，也就是说聚类采用的准则函数是收敛的，表示算法结束(这里采用的是误差平方和的准则函数)。

分析误差平方和准则函数可以看出 E 是样本与聚类中心差异度之和的函数，样本集 X 给定的情况下 E 的值取决于 c 个聚类中心。E 描述 n 个样本聚类成 c 个类时所产生的总的误差平方和。显然，若 E 值越大，说明误差越大，聚类结果越不好。因此，我们应该寻求使 E 值最小的聚类结果，即误差平方和准则的最优结果。这种聚类通常称为最小误差划分。误差平方和准则函数适用于各类样本比较集中而且样本数目悬殊不大的样本分布。当不同类型的样本数目相差较大时，采用误差平方和准则[9]。

K-Means 聚类算法属于一种动态聚类算法，也称作逐步聚类法，该算法的一个比较显著的特点就是迭代过程，每次都要考察对每个样本数据的分类正确与否，如果不正确，就要进行调整。当调整完全部的数据对象之后，再来修改中心，最后进入下一次迭代的过程中。若在一个迭代中，所有的数据对象都已经被正确的分类，那么就不会有调整，聚类中心也不会改变，聚类准则函数也表明已经收敛，那么该算法就成功结束。

2.3. 层次聚类法

层次聚类(Hierarchical Clustering)是聚类算法的一种，通过计算不同类别数据点间的相似度来创建一棵有层次的嵌套聚类树[10]。在聚类树中，不同类别的原始数据点是树的最低层，树的顶层是一个聚

类的根节点。创建聚类树有自下而上合并和自上而下分裂两种方法，根据自下而上的方法称为凝聚性的聚类算法；而自上而下的方法称为分裂型聚类算法。一个完全层次聚类的质量由于无法对已经做的合并或分解进行调整而受到影响。但是层次聚类算法没有使用准则函数，它所需要对数据结构的假设更少，所以适用性更强。

1) 层次聚类的合并算法通过计算两类数据点间的相似性，对所有数据点中最为相似的两个数据点进行组合，并反复迭代这一过程。简单的说层次聚类的合并算法是通过计算每一个类别的数据点与所有数据点之间的距离来确定它们之间的相似性，距离越小，相似度越高。并将距离最近的两个数据点或类别进行组合，生成聚类树。

2) 层次聚类的分裂方法是首先将所有对象放在一个族中，然后慢慢的细分为越来越小的族，直到每个对象自行形成一族，或者直达满足其他的一个终结条件，例如，满足了期望的族数目或者两个最近族之间的距离达到了所给定的阈值以下。

在层次聚类法中，都需要用户提供所希望得到的聚类的单个数量和阈值作为聚类分析的终止条件，但对于复杂数据来说，要事先给定出限制是十分困难的。尽管层次聚类的方法实现很简单，但会遇见合并或分裂点的选择的困难。因合并与分裂在聚类算法中是不可逆的，无法进行修正，只能进行下一步的新分裂或新合并，所以分裂点与合并点的选择是至关重要，会直接影响聚类结果的质量的高低。

2.4. DBSCAN 聚类算法

DBSCAN(Density-Based Spatial Clustering of Applications with Noise)聚类算法，它是一种基于高密度连通区域的、基于密度的聚类算法，能够将具有足够高密度的区域划分为簇，并在具有噪声的数据中发现任意形状的簇[11]。为了更好的解释 DBSCAN 方法需要解释下列两个定义：1) 对象的 ε 邻域：给定对象在半径 ε 内的区域。2) 核心对象：对于给定的数据 m，如果一个对象的 ε 邻域至少包含有 m 个对象，则成为该对象的核心对象。

DBSCAN 通过检查数据集中的每个对象的 ε 邻域来寻找聚类，如果一个点 p 的 ε 邻域包含对于 m 个对象，则创建一个 p 作为核心对象的新簇。然后，DBSCAN 反复地寻找这些核心对象直接密度可达的对象，这个过程可能涉及密度可达簇的合并。当没有新的点可以被添加到任何簇时，该过程结束。算法的中 ε 和 m 是根据先验知识来给出的[12]，所以 ε 与 m 参数的选择在 DBSCAN 算法中是十分重要的。

3. 红酒化学成分数据聚类分析

葡萄酒是通过发酵而成的含多种营养成分的饮料酒，其原料为新鲜葡萄或葡萄汁，是对人体有益的酒精饮品，其品种与质量都受到广泛的关注。长期以来，葡萄酒的品类和质量的评定都是通过葡萄酒各项指标得分的求和来确定，而葡萄酒指标的得分则是根据高资质的品酒员的感官指标来衡量[13]。然而这种评判模式，主观因素的成分过大，评判的结果存在较大的不确定性，所以利用葡萄酒本身的化学成分，进行聚类分析，客观、合理的评定葡萄酒类别及质量是十分重要的。本文对意大利的红酒的化学成分进行聚类分析，通过聚类得到的结果，划分出红酒的品种，根据品种所得到的特征评价品种质量的好坏，并与实际的红酒品种进行比对，评价聚类结果的好坏。

3.1. 数据来源

本文所采用的数据来自于 UCI (University of California Irvine)，选用 2007 年的意大利同一地区的葡萄

酒化学成分分析数据，其中包含 178 个样本与 13 个指标，所有的指标均已转化为数值变量来进行分析。13 个指标分别为：红酒品牌、酒精浓度、苹果酸含量、灰度、灰的碱度、镁含量、总酚类化合物量、类黄酮量、原花青素年、颜色强度、色调、稀释酒、脯氨酸。所有的指标均为数值变量。原始数据前 26 个样本如图 1 所示。

	X	酒精	苹果酸	灰	灰的碱度	镁	总酚类化合物	类黄酮	原花青素	颜色强度	色调	稀释酒	脯氨酸
1	1	14.23	1.71	2.43	15.6	127	2.8	3.06	2.29	5.64	1.04	3.92	1065
2	1	13.2	1.78	2.14	11.2	100	2.65	2.76	1.28	4.38	1.05	3.4	1050
3	1	13.16	2.36	2.67	18.6	101	2.8	3.24	2.81	5.68	1.03	3.17	1185
4	1	14.37	1.95	2.5	16.8	113	3.85	3.49	2.18	7.8	0.86	3.45	1480
5	1	13.24	2.59	2.87	21	118	2.8	2.69	1.82	4.32	1.04	2.93	735
6	1	14.2	1.76	2.45	15.2	112	3.27	3.39	1.97	6.75	1.05	2.85	1450
7	1	14.39	1.87	2.45	14.6	96	2.5	2.52	1.98	5.25	1.02	3.58	1290
8	1	14.06	2.15	2.61	17.6	121	2.6	2.51	1.25	5.05	1.06	3.58	1295
9	1	14.83	1.64	2.17	14	97	2.8	2.98	1.98	5.2	1.08	2.85	1045
10	1	13.86	1.35	2.27	16	98	2.98	3.15	1.85	7.22	1.01	3.55	1045
11	1	14.1	2.16	2.3	18	105	2.95	3.32	2.38	5.75	1.25	3.17	1510
12	1	14.12	1.48	2.32	16.8	95	2.2	2.43	1.57	5	1.17	2.82	1280
13	1	13.75	1.73	2.41	16	89	2.6	2.76	1.81	5.6	1.15	2.9	1320
14	1	14.75	1.73	2.39	11.4	91	3.1	3.69	2.81	5.4	1.25	2.73	1150
15	1	14.38	1.87	2.38	12	102	3.3	3.64	2.96	7.5	1.2	3	1547
16	1	13.63	1.81	2.7	17.2	112	2.85	2.91	1.46	7.3	1.28	2.88	1310
17	1	14.3	1.92	2.72	20	120	2.8	3.14	1.97	6.2	1.07	2.65	1280
18	1	13.83	1.57	2.62	20	115	2.95	3.4	1.72	6.6	1.13	2.57	1130
19	1	14.19	1.59	2.48	16.5	108	3.3	3.93	1.86	8.7	1.23	2.82	1680
20	1	13.64	3.1	2.56	15.2	116	2.7	3.03	1.66	5.1	0.96	3.36	845
21	1	14.06	1.63	2.28	16	126	3	3.17	2.1	5.65	1.09	3.71	780
22	1	12.93	3.8	2.65	18.6	102	2.41	2.41	1.98	4.5	1.03	3.52	770
23	1	13.71	1.86	2.36	16.6	101	2.61	2.88	1.69	3.8	1.11	4	1035
24	1	12.85	1.6	2.52	17.8	95	2.48	2.37	1.46	3.93	1.09	3.63	1015
25	1	13.5	1.81	2.61	20	96	2.53	2.61	1.66	3.52	1.12	3.82	845
26	1	13.05	2.05	3.22	25	124	2.63	2.68	1.92	3.58	1.13	3.2	830

Figure 1. The first 26 samples of the original data
图 1. 原始数据前 26 个样本

3.2. 指标筛选与数据预处理

在进行聚类分析前我们要对指标进行筛选，我们做聚类不需要红酒品种这一指标，所以去掉这一指标；苹果酸，是葡萄中主要的有机酸之一。葡萄酒中，苹果酸软化成乳酸进行发酵，影响葡萄酒的酸度与香味是十分重要的指标进行保留；而灰度和灰的碱度两个指标都是对葡萄酒碱灰的度量，所有去掉灰的指标留下灰的碱度的指标；对于不同的酒厂生产的葡萄酒用的水源是不同的，所以水中矿物质镁的含量也是值得被保留的指标；总酚类化合物与类黄酮都是对发酵方式的度量，只考虑其中一个指标即可，我们去掉总分类化合物的指标；而花青素、颜色强度与色调都是对选用葡萄的度量，色调与颜色强度指标想重复去掉色调的指标；脯氨酸是植物蛋白质的组分之一，并可以游离状态广泛存在于植物体中。在干旱、盐渍等胁迫条件下，许多植物体内脯氨酸大量积累，是对葡萄酒酿造过程中选取其它植物作为配料的度量保留。所以选择后的指标为：酒精浓度、苹果酸含量、灰的碱度、镁含量、类黄酮量、原花青素年、颜色强度、稀释酒、脯氨酸九个指标。这九个所选取的指标能使我们很好的进行聚类，且评价聚类后的红酒类别的质量。所对应处理的前 21 个数据样本如图 2 所示。

我们对数据观察发现，数据本身比较完整并无有缺失值的情况，不需要对数据做缺失值填补。我们接下来看一下数据的均值，最大值、最小值、中位数等，如图 3 所示。

	酒精	苹果酸	灰的碱度	镁	类黄酮	原花青素	颜色强度	稀释酒	脯氨酸
1	14.23	1.71	15.6	127	3.06	2.29	5.64	3.92	1065
2	13.2	1.78	11.2	100	2.76	1.28	4.38	3.4	1050
3	13.16	2.36	18.6	101	3.24	2.81	5.68	3.17	1185
4	14.37	1.95	16.8	113	3.49	2.18	7.8	3.45	1480
5	13.24	2.59	21	118	2.69	1.82	4.32	2.93	735
6	14.2	1.76	15.2	112	3.39	1.97	6.75	2.85	1450
7	14.39	1.87	14.6	96	2.52	1.98	5.25	3.58	1290
8	14.06	2.15	17.6	121	2.51	1.25	5.05	3.58	1295
9	14.83	1.64	14	97	2.98	1.98	5.2	2.85	1045
10	13.86	1.35	16	98	3.15	1.85	7.22	3.55	1045
11	14.1	2.16	18	105	3.32	2.38	5.75	3.17	1510
12	14.12	1.48	16.8	95	2.43	1.57	5	2.82	1280
13	13.75	1.73	16	89	2.76	1.81	5.6	2.9	1320
14	14.75	1.73	11.4	91	3.69	2.81	5.4	2.73	1150
15	14.38	1.87	12	102	3.64	2.96	7.5	3	1547
16	13.63	1.81	17.2	112	2.91	1.46	7.3	2.88	1310
17	14.3	1.92	20	120	3.14	1.97	6.2	2.65	1280
18	13.83	1.57	20	115	3.4	1.72	6.6	2.57	1130
19	14.19	1.59	16.5	108	3.93	1.86	8.7	2.82	1680
20	13.64	3.1	15.2	116	3.03	1.66	5.1	3.36	845
21	14.06	1.63	16	126	3.17	2.1	5.65	3.71	780

Figure 2. The first 21 data samples after indicator screening
图 2. 指标筛选后的前 21 个数据样本

酒精	苹果酸	灰的碱度
Min. :11.03	Min. :0.740	Min. :10.60
1st Qu.:12.36	1st Qu.:1.603	1st Qu.:17.20
Median :13.05	Median :1.865	Median :19.50
Mean :13.00	Mean :2.336	Mean :19.49
3rd Qu.:13.68	3rd Qu.:3.083	3rd Qu.:21.50
Max. :14.83	Max. :5.800	Max. :30.00
镁	类黄酮	原花青素
Min. : 70.00	Min. :0.340	Min. :0.410
1st Qu.: 88.00	1st Qu.:1.205	1st Qu.:1.250
Median : 98.00	Median :2.135	Median :1.555
Mean : 99.74	Mean :2.029	Mean :1.591
3rd Qu.:107.00	3rd Qu.:2.875	3rd Qu.:1.950
Max. :162.00	Max. :5.080	Max. :3.580
颜色强度	稀释酒	脯氨酸
Min. : 1.280	Min. :1.270	Min. : 278.0
1st Qu.: 3.220	1st Qu.:1.938	1st Qu.: 500.5
Median : 4.690	Median :2.780	Median : 673.5
Mean : 5.058	Mean :2.612	Mean : 746.9
3rd Qu.: 6.200	3rd Qu.:3.170	3rd Qu.: 985.0
Max. :13.000	Max. :4.000	Max. :1680.0

Figure 3. The statistical value of each index
图 3. 数据各指标的统计值

从图 3 我们可以看到，脯氨酸、镁含量的均值都比较高，远远高于其它几个指标，苹果酸、类黄酮、原青花素、稀释酒的差别并不大，为了避免在聚类时因部分指标的值更大而在聚类中产生主导作用，所以我们需要对原始数据进行标准化，使得每个指标都具有相同的尺度。标准化后的部分数据样本如图 4 所示。

我们可以看到标准化后的数据如图 5 所示，脯氨酸、镁含量的均值已经显著下降，与各样本的指标尺度已经相同，至此我们可以开始聚类分析。

	酒精	苹果酸	灰的碱度	镁	类黄酮
1	1.514341	−0.56067	−1.1663	1.908522	1.031908
2	0.245597	−0.49801	−2.48384	0.018094	0.731565
3	0.196325	0.021172	−0.26798	0.08811	1.212114
4	1.686791	−0.34584	−0.80697	0.9283	1.462399
5	0.294868	0.227053	0.450674	1.278379	0.661485
6	1.477387	−0.51591	−1.28608	0.858284	1.362285
7	1.711427	−0.41745	−1.46574	−0.26197	0.491291
8	1.304936	−0.16681	−0.56742	1.488427	0.48128
9	2.253415	−0.62333	−1.64541	−0.19195	0.951817
10	1.058578	−0.88292	−1.04653	−0.12194	1.122011
11	1.354208	−0.15786	−0.44765	0.368173	1.292205
12	1.378844	−0.76655	−0.80697	−0.33199	0.401188
13	0.923081	−0.54277	−1.04653	−0.75208	0.731565
14	2.154872	−0.54277	−2.42395	−0.61205	1.662628
15	1.699109	−0.41745	−2.24429	0.158126	1.612571
16	0.775267	−0.47115	−0.6872	0.858284	0.881737
17	1.600566	−0.37269	0.151234	1.418411	1.111999
18	1.021625	−0.68599	0.151234	1.068331	1.372297
19	1.465069	−0.66808	−0.89681	0.578221	1.902902
20	0.787585	0.683574	−1.28608	1.138347	1.001874

Figure 4. Partial data samples after standardization
图 4. 标准化后的部分数据样本

酒精	苹果酸	灰的碱度
Min. :−2.42739	Min. :−1.4290	Min. :−2.663505
1st Qu.:−0.78603	1st Qu.:−0.6569	1st Qu.:−0.687199
Median : 0.06083	Median :−0.4219	Median : 0.001514
Mean : 0.00000	Mean : 0.0000	Mean : 0.000000
3rd Qu.: 0.83378	3rd Qu.: 0.6679	3rd Qu.: 0.600395
Max. : 2.25341	Max. : 3.1004	Max. : 3.145637
镁	类黄酮	原花青素
Min. :−2.0824	Min. :−1.6912	Min. :−2.06321
1st Qu.:−0.8221	1st Qu.:−0.8252	1st Qu.:−0.59560
Median :−0.1219	Median : 0.1059	Median :−0.06272
Mean : 0.0000	Mean : 0.0000	Mean : 0.00000
3rd Qu.: 0.5082	3rd Qu.: 0.8467	3rd Qu.: 0.62741
Max. : 4.3591	Max. : 3.0542	Max. : 3.47527
颜色强度	稀释酒	脯氨酸
Min. :−1.6297	Min. :−1.8897	Min. :−1.4890
1st Qu.:−0.7929	1st Qu.:−0.9496	1st Qu.:−0.7824
Median :−0.1588	Median : 0.2371	Median :−0.2331
Mean : 0.0000	Mean : 0.0000	Mean : 0.0000
3rd Qu.: 0.4926	3rd Qu.: 0.7864	3rd Qu.: 0.7561
Max. : 3.4258	Max. : 1.9554	Max. : 2.9631

Figure 5. The statistical value of each index after standardization
图 5. 标准化后数据各指标的统计值

3.3. 聚类结果及分析

3.3.1. K-means 聚类

我们选用 R 软件，对处理后的 178 个样本及 9 个指标采用 K-means 聚类方法进行聚类，再结合样本原有的分类指标，考察聚类的准确性与现实的解释意义。在进行聚类时，类的选择十分的重要，这里我们采用计算不同 k 值下的 Gap 统计值来确定最优的 K，如图 6，我们可以明显的看到，GAP 统计值在从 1 到 3 有明显的上升趋势，而在 3 以后统计值一直在 0.65 附近进行波动，所以从 GAP 统计值的图中我们可以明显的看到选择 K = 3 是最合理的，所以在接下来的聚类中我们选择 K = 3 聚为三类。

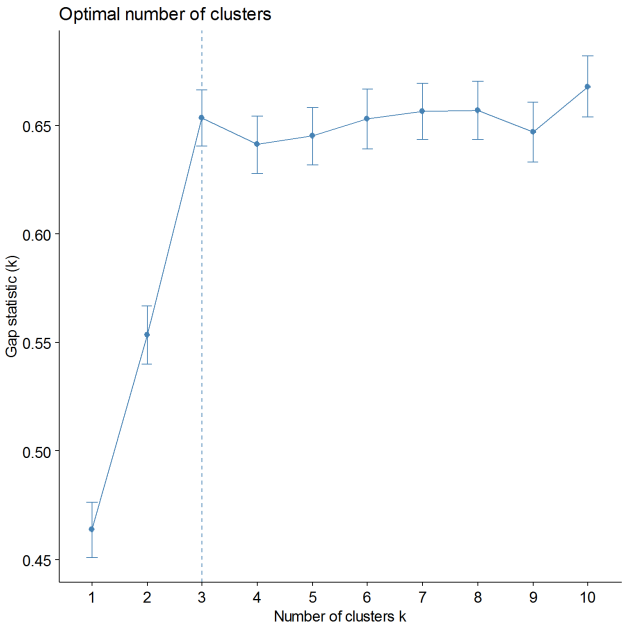

Figure 6. GAP statistics trend chart with K value
图 6. GAP 统计值随 k 值趋势图

聚类结果如图 7 所示，从图中我们可以明确地看到，所有的样本被聚为了三类，组间距离与总体距离的比值在 48.7%，代表组与组之间的距离不是很大，聚类组与组之间的差异并不算很显著。

```
Clustering vector:
  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [37] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 3 3 2 3 3 3 3 3 2 1 3 3
 [73] 3 1 3 3 3 3 1 3 3 3 3 2 3 3 3 3 3 3 3 3 3 3 3 1 3 3 3 3 3 3 3 3 3 3 3
[109] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 2 2 2 2 2 2 2 2 2 2 2 2 2
[145] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2

Within cluster sum of squares by cluster:
[1] 287.1625 224.8907 304.3654
 (between_SS / total_SS =  48.7 %)
```

Figure 7. Cluster result diagram
图 7. 聚类结果图

接下来我们来看，聚类所得到的每一个类别的大小如图 8 所示，我们可以看到每个类的样本数都比较的均衡差距不远，整体样本被均衡的分为了三个类。

```
> wine_r$size
[1] 63 51 64
```

Figure 8. Sample size of each category
图 8. 各类的样本数

为了更加清晰的展示聚类的结果，对聚类进行可视化，做出聚类结果图如图 9 所示。

我们可以明显的看到，聚类的效果还是十分理想的，类与类间的划分明确，可以看到前面的 50~60 个样本基本分为一类，中间的样本基本聚为一类，后部分的样本基本聚为一类。而且每个类的类中心比较明确。下面我们需要对类的聚集做出解释，为何会形成这样的聚类，所以我们需要对每一类的特征值进行研究，如图 10 所示。

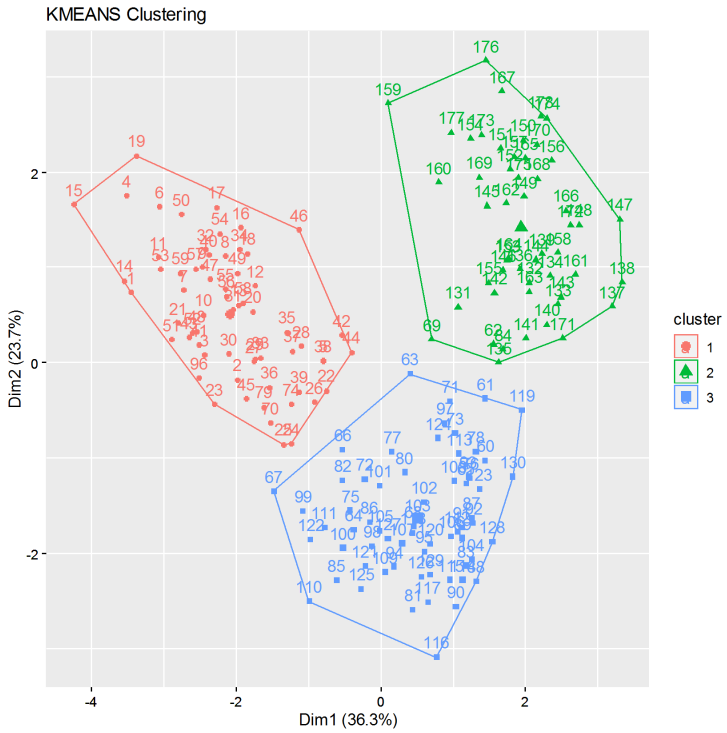

Figure 9. Visualization of clustering results
图 9. 聚类结果可视化图

	酒精	苹果酸	灰的碱度	镁	类黄酮
cluster_1	0.819259	−0.3295	−0.67674	0.64268	0.896357
cluster_2	0.178211	0.825392	0.528764	−0.03407	−1.21085
cluster_3	−0.94847	−0.33339	0.244809	−0.60548	0.082543
	原花青素	颜色强度	稀释酒	脯氨酸	
cluster_1	0.619363	0.134954	0.744115	1.11712	
cluster_2	−0.79156	0.936945	−1.29402	−0.38241	
cluster_3	0.021088	−0.87947	0.298687	−0.79493	

Figure 10. Mean values of all clusters
图 10. 各类特征值均值

从图 10 我们可以分析出，三个类在酒精度的差异上是很大的，第一种、第三种红酒度数较高，而第二种红酒度数较低，从苹果酸指标上来看，第一种和第三种红酒的酸味是比较淡的，而第二种红酒酸味值较高属于比较酸的红酒，灰的碱度来看第三种红酒明显小于第一第二种红酒，说明第三种酒的发酵时间是小于第一二种酒的；从镁含量来看第一种红酒与第三种红酒是相近的有可能采用同一水源，而第二种红酒几乎不含有镁矿物质，猜测可能采用纯净水进行红酒的酿造；从类黄酮指标看来第三种红酒的含量非常小，所以发酵的时间会比较短，正好印证了前面灰的碱度指标所反映出来的发酵时间的推断；原花青素与颜色强度指标应该联合起来一起考察，因为它们都是对于葡萄酒生产时所选用葡萄的度量指标，综合来看三个类所选用的葡萄都不相同，第一种酒、第二种酒都选用青葡萄进行酿造，第三种酒选用红葡萄进行酿造，但在酿造出来的结果上第一种红酒颜色较浅，第二三种酒的颜色都比较深；稀释酒指标与酒精指标相对应也反映出了三种酒酒精浓度的问题；脯氨酸作为度量配方的指标也看出三种酒之间的差异，第一、二种酒都在酒中添加了其余的植物才会有高比例的脯氨酸含量，而第三种酒还是以葡萄为主进行的酿造。

综合上述的分析来看，聚类产生的三种红酒，第一种属于酒精含量较高，酸度较低，用青葡萄酿造，发酵时间较长，颜色较深且纯粹，含有的营养最高，属于优质红酒；第二种酒酒精含量低，酸度较高，用青葡萄酿造，发酵时间长，颜色很深偏红中带黑，所含营养较低，属于品质较差的红酒；第三种酒酒精含量高，酸度适中，采用红葡萄酿造，发酵时间短，颜色较浅，所含营养适中，属于中等质量红酒。

因原始数据中带有样本本身的分类招标，在聚类的过程中我们把指标剔除，现用聚类后的结果与分类指标进行比对，做出可视化图像如图 11。

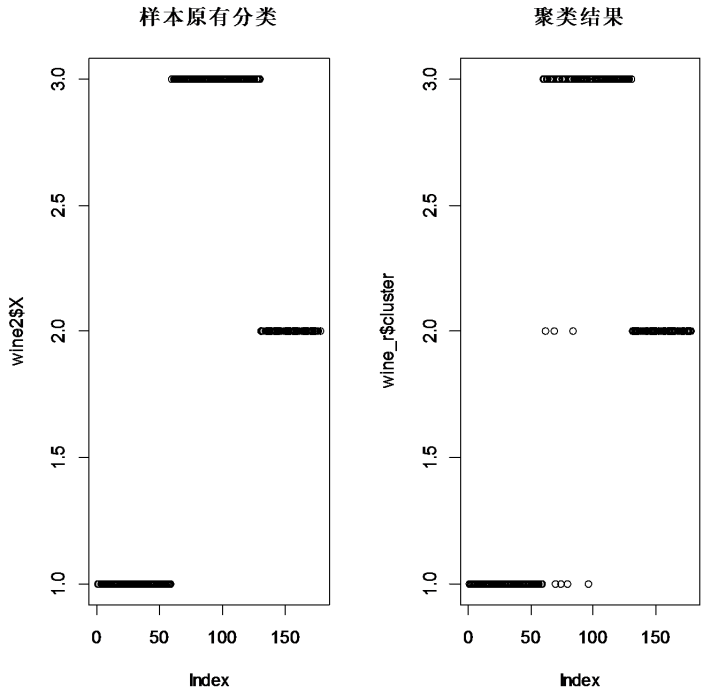

Figure 11. Comparison between the original classification of samples and k-means clustering
图 11. 样本原分类与 K-means 聚类的对比

我们可以发现，聚类的精准度是比较不错的，聚类的结果也表明聚类所做出的红酒品类划分与品酒员给出的经验划分高度一致，聚类的效果十分理想。

3.3.2. 层次聚类

同样利用已经处理后的 178 个样本及 9 个指标进行层次聚类的分析，来与 K-menas 聚类来进行对比，延用已处理好的原数据，层次聚类结果，进行可视化后如图 12 所示。

从图 12 我们可以通过颜色明显的看到，被聚为了三类，同时样本的类分配的比较均衡，第二类只略微少于第三类。从数值结果来看，前 64 个样本被划为一类，中间 67 个样本被划为一类，后 47 个样本被聚为一类。如图 13 所示。

从图 13 我们可以得到。层次聚类与 K-means 聚类是有差别的，K-means 聚类的各个类的样本数分别为 63, 51, 64，在第二类与第三类的样本数上有较大的差异。再来对比两个聚类方法的聚类结果。通过对比图 12 与图 7 可以知道，在层次聚类方法中的第二类与 K-means 聚类的第三类的样本是基本一致的，可以看作层次聚类出来的第二类为 K-means 聚类的第三类，层次聚类出来的第三类为 K-means 聚类的第二类。

```
> wine_r2$cluster
  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [37] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 1 2 2
 [73] 2 1 2 2 2 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 2 2 2 2 2 2 2 2 2 2 2
[109] 2 2 2 2 2 2 2 2 2 2 2 2 2 1 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3
[145] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
```

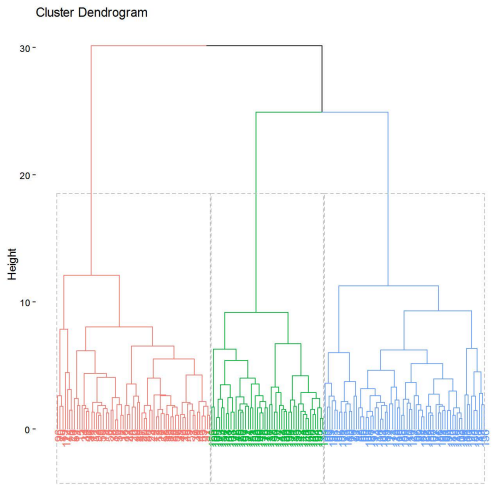

Figure 12. Hierarchical clustering result diagram
图 12. 层次聚类结果图

```
> wine_r2$size
[1] 64 67 47
```

Figure 13. Number of samples
图 13. 各类样本数

下面我们需要对类的聚集做出解释，为何会形成这样的聚类，计算每个类样本所有指标的均值后，如图 14 所示。

	酒精	苹果酸	灰的碱度	镁	类黄酮
cluster_1	1.514341	−0.56067	−1.1663	1.908522	1.031908
cluster_2	−0.3703	1.37283	1.049555	0.08811	0.521325
cluster_3	−0.9246	2.133697	0.450674	−0.75208	−1.56105

	原花青素	颜色强度	稀释酒	脯氨酸
cluster_1	1.221438	0.251009	1.842721	1.010159
cluster_2	0.627406	−1.07325	0.73003	−0.90151
cluster_3	−1.38182	−0.52111	−1.88972	−0.0854

Figure 14. Hierarchical clustering of category eigenvalues
图 14. 层次聚类类别特征值

从图 14 我们可以分析出，三个类在酒精度的差异上是很大的，第一种、第三种红酒度数较高，而第二种红酒度数较低，从苹果酸指标上来看，第二种和第三种红酒的酸味是比较高的果酸含量高的红酒，而第一种红酒酸味值较低，灰的碱度来看第三种红酒明显小于第一第二种红酒，说明第三种酒的发酵时间是小于第一二种酒的；从镁含量来看第一种红酒与第三种红酒镁含量高微量元素多营养价值高，而第二种红酒几乎不含有镁矿物质，营养价值小；从类黄酮指标看来第三种红酒的含量非常差别大，所以发

酵的时间也各不相同；原花青素与颜色强度指标反应出，第一种酒、第三种酒都选用青葡萄进行酿造，第三种酒选用红葡萄进行酿造，但在酿造出来的结果上第二种红酒颜色较浅，第一、三种酒的颜色都比较深；稀释酒指标与酒精指标相对应也反映出了三种酒酒精浓度的问题；脯氨酸作为度量配方的指标也看出三种酒之间的差异，第一、二种酒都在酒中添加了其余的植物才会有高比例的脯氨酸含量，而第三种酒还是以葡萄为主进行的酿造。

综合上述的分析来看，聚类产生的三种红酒，第一种属于酒精含量较高，酸度较低，用青葡萄酿造，发酵时间较长，颜色浅红剔透，含有的营养最高，属于优质红酒；第二种酒酒精含量低，酸度较高，用青葡萄酿造，发酵时间长，颜色很浅，所含营养较低，属于品质较差的红酒；第三种酒酒精含量高，酸度适中，采用红葡萄酿造，发酵时间短，颜色较浅，所含营养适中，属于中等质量红酒。

再来把层次聚类的结果与样本原分类指标进行比对，如图 15 所示。

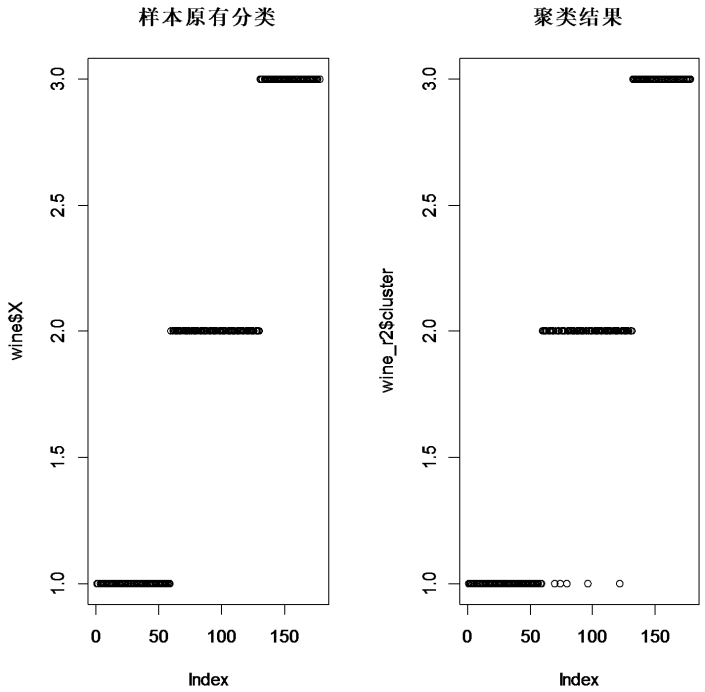

Figure 15. Comparison of original classification and hierarchical clustering of samples
图 15. 样本原分类与层次聚类的对比

可也看到聚类的效果也是十分不错的，比较结果说明，对于在本文使用的数据层次聚类方法的效果十分不错，产生的聚类结果与原本数据的分类指标都高度吻合，说明聚类的红酒品类结果与品酒员的经验评价结果是一致的。

3.3.3. Dbscan 聚类

同样利用已经处理后的 178 个样本及 9 个指标进行 DBSCAN 聚类的分析，因 DBSCAN 是基于密度的聚类方法，所以我们优先选定领域内最少点个数为 5，并计算距离参数的取值。计算的结果如图 16 所示。

观察图 16，图 16 是由矩阵中的 k = 5 的最近邻的距离，然后按距离从小到大排序后，绘制而成的图像，可以明显观察到在 y = 2.5 处，出现比较明显的拐点。所以我们的最优距离参数(eps)取值为 2.5。参数设置完毕后进行 DBSCAN 方法的聚类，结果如图 17 展示。

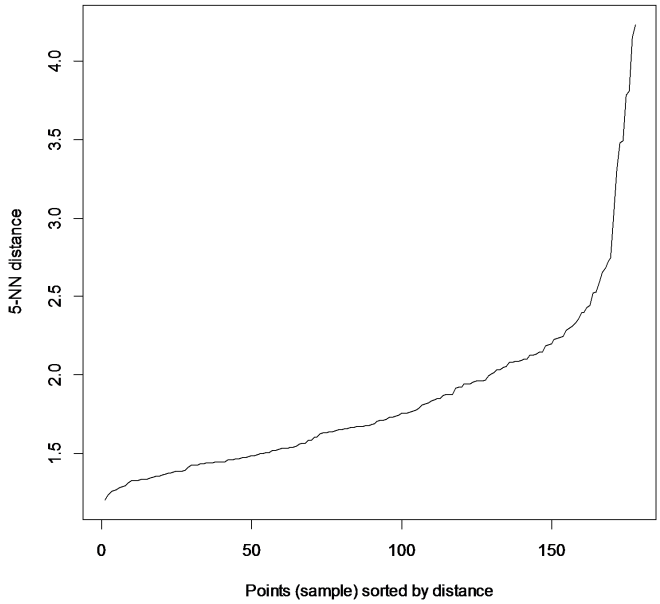

Figure 16. Sample point K-distance curve
图 16. 样本点 K-距离曲线

```
> wine_r3
dbscan Pts=178 MinPts=5 eps=2.5
       0   1
border 4   7
seed   0 167
total  4 174
```

Figure 17. Dbscan algorithm clustering result graph
图 17. Dbscan 算法聚类结果图

把聚类的结果进行可视化绘图展示为图 18。

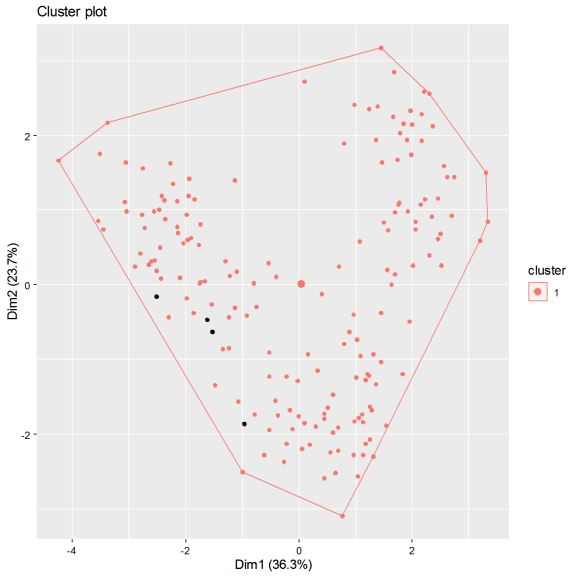

Figure 18. DBSCAN algorithm clustering results visual display figure
图 18. DBSCAN 算法聚类结果可视化展示图

我们可以看到在 DBSCAN 这种基于密度的聚类方法得到聚类结果为一个类，即只有一个红酒品类。参考原本数据中被剔除的红酒品种指标有三种红酒品种，说明聚类的效果十分不理想，出现这种结果的原因是 DBSCAN 聚类方法认为这三种类在密度上的差别不大，所以归为了同一类处理。DBSCAN 聚类的结果与样本原分类指标进行比对，如图 19 所示。

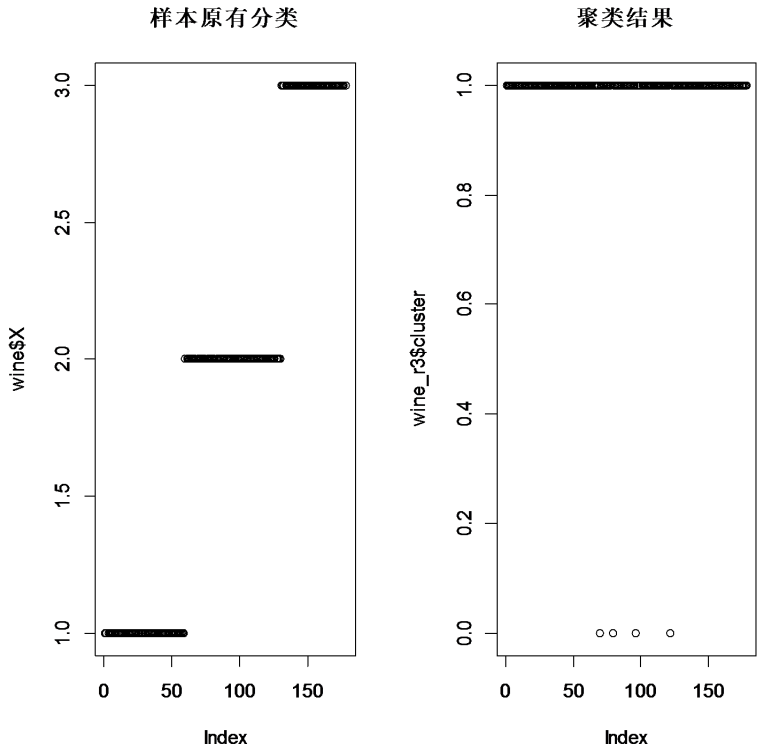

Figure 19. Comparison between the original classification of samples and DBSCAN clustering
图 19. 样本原分类与 DBSCAN 聚类的对比

也可以看到聚类的效果很不理想，DBSCAN 聚类方法并没有明显的区分出原始指标的类别来；只划分出了一个大类，说明现有的红酒化学成分数据并不适用于 DBSCAN 的聚类方法。

3.4. 聚类结果对比分析

把三种聚类方法的结果与原始数据指标分类绘图对比，如图 20 所示。

对本文使用的红酒化学成分数据来说，K-means 聚类与层次聚类得到的效果更好，是最接近原原始样本指标品种分类的；DBSCAN 聚类方法的效果不理想并没有得到预期想要的聚类效果，说明本文的样本数据不适用于 DBSCAN 聚类方法。K-means 聚类的结果为三类，分别对应 63、51、64 个样本，而且得到每个聚类的特征向量，对聚类的结果能有更好的解释。层次聚类得到的结果与 K-means 聚类类似得聚类为三类，分别对应 64，67，47 个样本，在第二类与第三类上与 K-means 聚类有所差异，但差别不大。DBSCAN 聚类结果为一类，对应 174 个样本，余下四个样本当作异常值点，对比原始样本品种分类不具有指导意义。综合来看，本文所研究的数据，聚类效果最好的为 K-means 聚类，且根据聚类效果与特征向量可以划分为三类，第一种属于酒精含量较高，酸度较低，用青葡萄酿造，发酵时间较长，颜色较深且纯粹，含有的营养最高，属优质红酒；第二种酒酒精含量低，酸度较高，用青葡萄酿造，发酵时间长，颜色很深偏红中带黑，所含营养较低，属于品质较差的红酒；第三种酒酒精含量高，酸度适中，采用红

葡萄酿造，发酵时间短，颜色较浅，所含营养适中，属于中等质量红酒，对比品酒员给出的葡萄酒各项指标得分加和而得到的经验红酒品种分类也高度吻合。

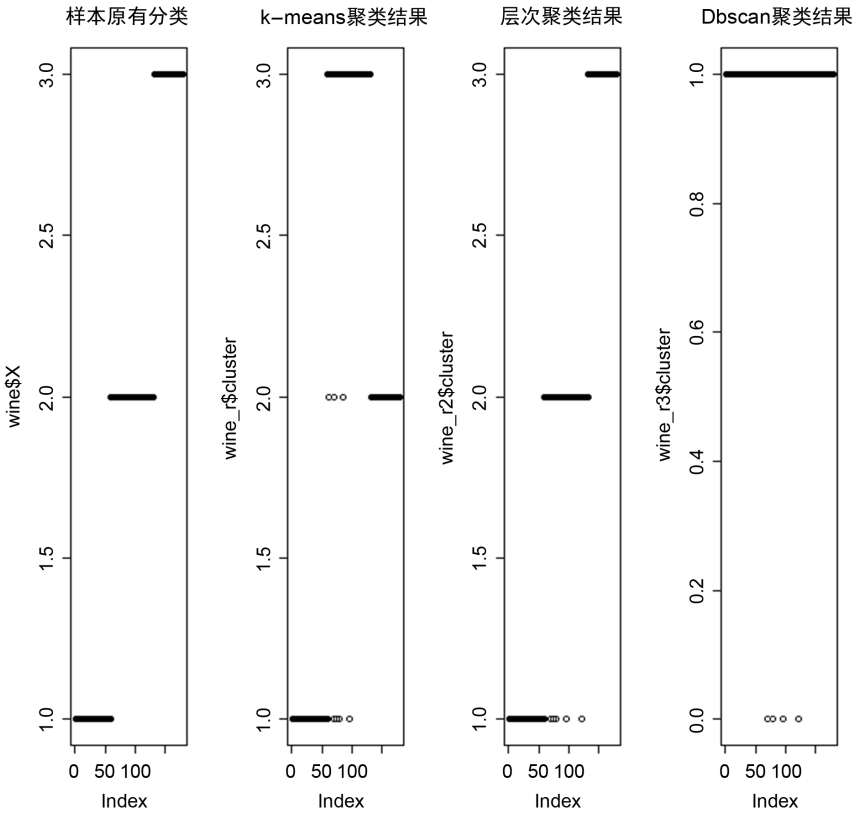

Figure 20. Comparison of original classification and clustering results of samples
图 20. 样本原分类与聚类结果的对比

4. 结论

根据红酒化学成分数据，基于聚类的结果来看，K-means 聚类方法与层次聚类方法的聚类效果十分理想，而 DBSCAN 聚类方法并不适用于该数据。从最优的聚类结果分析，该红酒化学成分数据的红酒品种可划分为三种，第一种属于酒精含量较高，酸度较低，用青葡萄酿造，发酵时间较长，颜色较深且纯粹，含有的营养最高，属于优质红酒；第二种酒酒精含量低，酸度较高，用青葡萄酿造，发酵时间长，颜色很深偏红中带黑，所含营养较低，属于品质较差的红酒；第三种酒酒精含量高，酸度适中，采用红葡萄酿造，发酵时间短，颜色较浅，所含营养适中，属于中等质量红酒；样本量分别为 63、51、64。对比品酒员给出的葡萄酒各项指标得分加和而得到的经验红酒品种分类也高度吻合，所以聚类方法可以给葡萄酒的品种分类及质量评定提供参考及可选择的方法。

基金项目

重庆市自然科学基金(批准号：cstc2018jcyjAX0464)，重庆市研究生教育教学改革研究重大项目(yjg191017)，重庆市高等教育教学改革研究一般项目(193180)，重庆理工大学高等教育教学改革研究重点项目(2018ZD05)，重庆市专业学位研究生教学案列库建设项目(201967，应用统计专业学位研究生教学案例建设)。

参考文献

[1] Hou, G.L., Ge, B., Sun, L.L. and Xing, K.X. (2020) A Study on Wine Sensory Evaluation by the Statistical Analysis Method. *Czech Journal of Food Sciences*, **38**, 1-10. https://doi.org/10.17221/438/2017-CJFS

[2] 周涓, 熊忠阳, 张玉芳, 等. 基于最大最小距离法的多中心聚类算法[J]. 计算机应用, 2006, 26(6): 1425-1427.

[3] 喻彪, 骆雯, 赖朝安. 数据挖掘聚类算法研究[J]. 现代制造工程, 2009(3): 141-145.

[4] 贺玲, 吴玲达, 蔡益朝. 数据挖掘中的聚类算法综述[J]. 计算机应用研究, 2007(1): 10-13.

[5] 周开乐, 杨善林, 丁帅, 罗贺. 聚类有效性研究综述[J]. 系统工程理论与实践, 2014, 34(9): 2417-2431.

[6] Bernhard, F. (1988) Algorithms for Clustering Data. In: Jain, A.K. and Dubes, R.C., Eds., *Prentice Hall Advanced Reference Series in Computer Science*, Prentice Hall, Englewood Cliffs, NJ, Vol. 21, 137-138.

[7] 王莉. 数据挖掘中聚类方法的研究[D]: [博士学位论文]. 天津: 天津大学, 2004.

[8] 吴晓蓉. K-均值聚类算法初始中心[D]: [硕士学位论文]. 长沙: 湖南大学, 2006.

[9] 张科泽, 杨鹤标, 沈项军, 等. 基于节点数据密度的分布式 K-means 聚类算法研究[J]. 辽宁工程技术大学, 2011, 28(10): 3643-3645, 3655.

[10] 段明秀. 层次聚类算法的研究及应用[D]: [硕士学位论文]. 长沙: 中南大学, 2009.

[11] 周水庚, 周傲英, 曹晶. 基于数据分区的 DBSCAN 算法[J]. 计算机研究与发展, 2000, 37(10): 1153-1159.

[12] Zhang, Y.C., Chen, S., Chen, S.Y., Chen, H. and Guo, P. (2020) A Novel Lidar Gradient Cluster Analysis Method of Nocturnal Boundary Layer Detection during Air Pollution Episodes. *Atmospheric Measurement Techniques*, **13**, 6675-6689. https://doi.org/10.5194/amt-13-6675-2020

[13] Blanquet, J., Fur, Y.L. and Ballester, J. (2017) Computerized Delimitation of Odorant Areas in Gas-Chromatography Olfactometry by Kernel Density Estimation. *Data Processing on French White Wines*, **167**, 29-35. https://doi.org/10.1016/j.chemolab.2017.05.015