1. Данные

Использование кластер-анализа для решения данной задачи наиболее эффективно. В общем случае кластер-анализ предназначен для объединения некоторых объектов в классы (кластеры) таким образом, чтобы в один класс попадали максимально схожие, а объекты различных классов максимально отличались друг от друга. Количественный показатель сходства рассчитывается заданным способом на основании данных, характеризующих объекты

Сгруппируем по признакам совокупности данных

соэ	лейкоциты	глюкоза
5	10,8	4,5
5	17,2	6,2
29	6,8	4,3
15	10,8	4,5
3	10,5	4,3
25	18,5	6,5
5	5,6	4,3
9	8,3	6,2
3	7,2	1,3
15	16	5,6
4	16,7	5
3	12,3	3,1
3	10,3	4,5
10	14	4,3
5	15,8	3,5

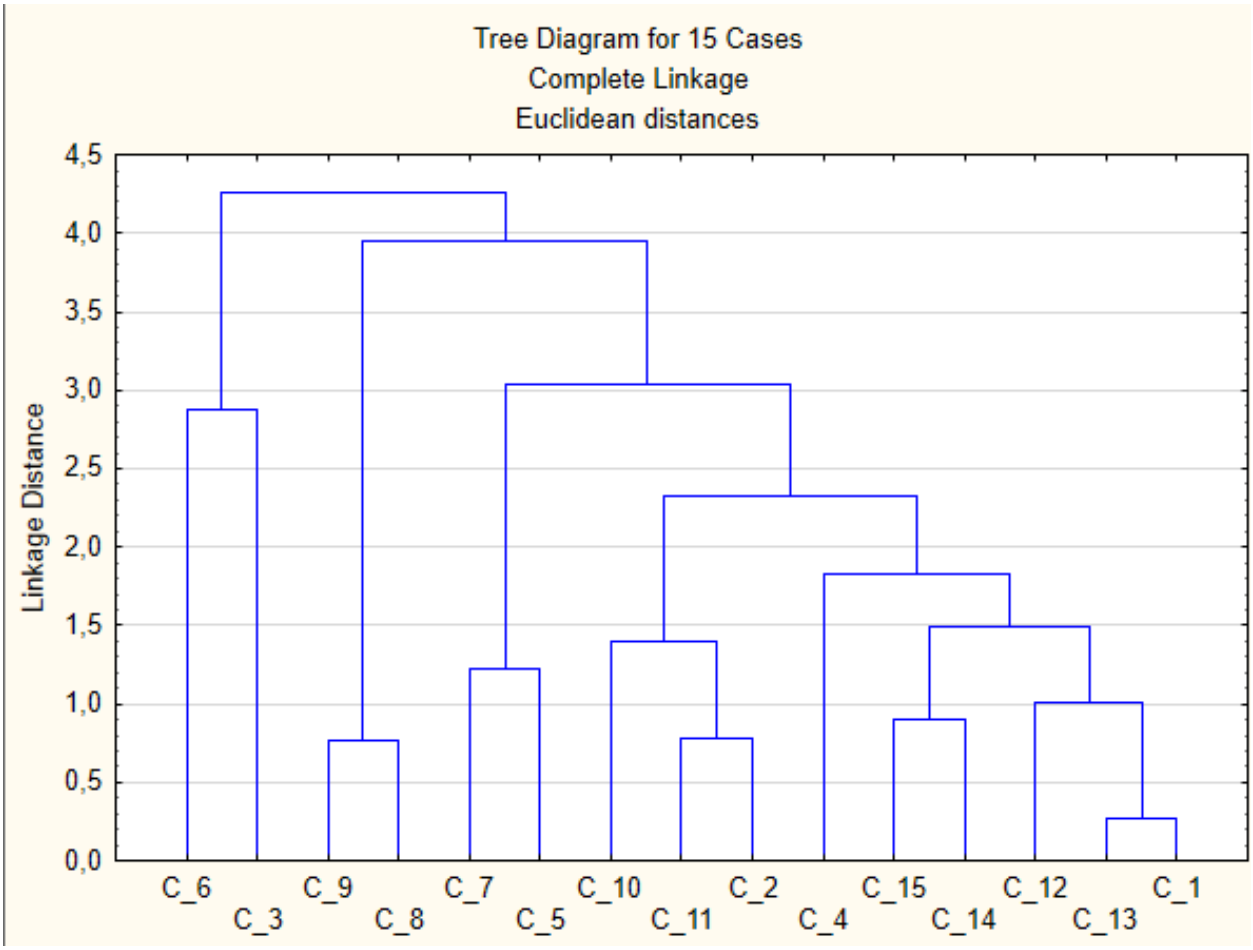
Данные взяты из БД СХ и РГ 1998-2002. https://rosinformagrotech.ru/db/bd-agrotekhnologij

2. Единицы измерения и расстояние между элементами

Стандартизируем данные

	1 СОЭ	2 лейкоцмты	3 глюкоза
1	-0,5149987	-0,3035845	0,09725851
2	-0,5149987	1,24663428	1,17555939
3	2,38186883	-1,2724713	-0,0296004
4	0,69202946	-0,3035845	0,09725851
5	-0,7564043	-0,376251	1,36584777
6	1,89905758	1,56152247	-0,0296004
7	-0,5149987	-1,5631373	1,23898885
8	-0,0321874	-0,9091387	-1,9324843
9	-0,7564043	-1,1755826	-1,9324843
10	0,69202946	0,95596826	0,79498261
11	-0,6357015	1,12552344	0,41440583
12	-0,7564043	0,05974802	-0,790754
13	-0,7564043	-0,4246954	0,09725851
14	0,0885154	0,47152488	-0,0296004
15	-0,5149987	0,90752392	-0,5370361

3. Иерархический алгоритм



Посчитаем расстояния по формуле Евклида:

$$d(\overrightarrow{x},\ \overrightarrow{y})=\sqrt{(x_1-y_1)^2+(x_2-y_2)^2+\cdots+(x_m-y_m)^2}$$

Суть иерархической кластеризации состоит в последовательном объединении меньших кластеров в большие. Алгоритм иерархической кластеризации начинает работу с формирование несвязанных кластеров, помещая каждый из объектов в отдельный кластер.

Определяется способ интерпретации матрицы расстояний, чтобы объединить два или более кластера следующего уровня. Процесс повторяется, пока не получим резкого скачка в минимальном расстоянии.

Существет 3 метода поиска необходимых расстояний:

- Принцип «ближнего соседа»
- Принцип средних связей
- Принцип «дальнего соседа»

	Euclidean distances (an)														
Case No.	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_10	C_11	C_12	C_13	C_14	C_15
C_1	0,00	1,89	3,06	1,21	1,29	3,05	1,70	2,17	2,22	1,88	1,47	0,99	0,27	0,99	1,37
C_2	1,89	0,00	4,02	2,24	1,65	2,72	2,81	3,81	3,95	1,30	0,78	2,31	2,00	1,55	1,75
C_3	3,06	4,02	0,00	1,95	3,55	2,87	3,18	3,10	3,67	2,92	3,88	3,49	3,25	2,88	3,66
C_4	1,21	2,24	1,95	0,00	1,93	2,23	2,08	2,24	2,64	1,44	1,98	1,74	1,45	0,99	1,82
C_5	1,29	1,65	3,55	1,93	0,00	3,57	1,22	3,42	3,39	2,05	1,78	2,20	1,27	1,84	2,31
C_6	3,05	2,72	2,87	2,23	3,57	0,00	4,15	3,67	4,26	1,58	2,61	3,14	3,32	2,11	2,55
C_7	1,70	2,81	3,18	2,08	1,22	4,15	0,00	3,27	3,20	2,83	2,81	2,61	1,63	2,47	3,04
C_8	2,17	3,81	3,10	2,24	3,42	3,67	3,27	0,00	0,77	3,38	3,16	1,66	2,21	2,35	2,34
C_9	2,22	3,95	3,67	2,64	3,39	4,26	3,20	0,77	0,00	3,75	3,29	1,68	2,16	2,65	2,52
C_10	1,88	1,30	2,92	1,44	2,05	1,58	2,83	3,38	3,75	0,00	1,39	2,33	2,12	1,13	1,80
C_11	1,47	0,78	3,88	1,98	1,78	2,61	2,81	3,16	3,29	1,39	0,00	1,61	1,59	1,07	0,98
C_12	0,99	2,31	3,49	1,74	2,20	3,14	2,61	1,66	1,68	2,33	1,61	0,00	1,01	1,21	0,92
C_13	0,27	2,00	3,25	1,45	1,27	3,32	1,63	2,21	2,16	2,12	1,59	1,01	0,00	1,24	1,50
C_14	0,99	1,55	2,88	0,99	1,84	2,11	2,47	2,35	2,65	1,13	1,07	1,21	1,24	0,00	0,90
C_15	1,37	1,75	3,66	1,82	2,31	2,55	3,04	2,34	2,52	1,80	0,98	0,92	1,50	0,90	0,00

1) Получили матрицу 15 на 15.

2) Далее найдем минимальное расстояние. Это расстояние равно 0,223. И объединим 5 и 13 в один кластер.

3) Матрица теперь 14 на 14. Пересчитаем ее. Расстояние до нового кластера будем брать по принципу «ближнего соседа».

4) Повторяем весь алгоритм, пока не получим резкого скачка в минимальном расстоянии.

Мы получили 6 кластеров:

1) 6
2) 3
3) 8,9
4) 2,10,11,14,15
5) 1,4,12,13
6) 5,7

4. Алгоритм к-средних

Что из себя представляет данный алгоритм?

1) Задаем произвольные начальные значения центров

2) На каждой итерации перевычисляется центр масс для каждого кластера, полученного на предыдущем шаге, затем векторы разбиваются на кластеры вновь в соответствии с тем, какой из новых центров оказался ближе по выбранной метрике.

3) Алгоритм завершается, когда на какой-то итерации не происходит изменения внутрикластерного расстояния

Case No.	Euclidean distances (an)														
	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_10	C_11	C_12	C_13	C_14	C_15
C_1	0,00	1,89	3,06	1,21	1,29	3,05	1,70	2,17	2,22	1,88	1,47	0,99	0,27	0,99	1,37
C_2	1,89	0,00	4,02	2,24	1,65	2,72	2,81	3,81	3,95	1,30	0,78	2,31	2,00	1,55	1,75
C_3	3,06	4,02	0,00	1,95	3,55	2,87	3,18	3,10	3,67	2,92	3,88	3,49	3,25	2,88	3,66
C_4	1,21	2,24	1,95	0,00	1,93	2,23	2,08	2,24	2,64	1,44	1,98	1,74	1,45	0,99	1,82
C_5	1,29	1,65	3,55	1,93	0,00	3,57	1,22	3,42	3,39	2,05	1,78	2,20	1,27	1,84	2,31
C_6	3,05	2,72	2,87	2,23	3,57	0,00	4,15	3,67	4,26	1,58	2,61	3,14	3,32	2,11	2,55
C_7	1,70	2,81	3,18	2,08	1,22	4,15	0,00	3,27	3,20	2,83	2,81	2,61	1,63	2,47	3,04
C_8	2,17	3,81	3,10	2,24	3,42	3,67	3,27	0,00	0,77	3,38	3,16	1,66	2,21	2,35	2,34
C_9	2,22	3,95	3,67	2,64	3,39	4,26	3,20	0,77	0,00	3,75	3,29	1,68	2,16	2,65	2,52
C_10	1,88	1,30	2,92	1,44	2,05	1,58	2,83	3,38	3,75	0,00	1,39	2,33	2,12	1,13	1,80
C_11	1,47	0,78	3,88	1,98	1,78	2,61	2,81	3,16	3,29	1,39	0,00	1,61	1,59	1,07	0,98
C_12	0,99	2,31	3,49	1,74	2,20	3,14	2,61	1,66	1,68	2,33	1,61	0,00	1,01	1,21	0,92
C_13	0,27	2,00	3,25	1,45	1,27	3,32	1,63	2,21	2,16	2,12	1,59	1,01	0,00	1,24	1,50
C_14	0,99	1,55	2,88	0,99	1,84	2,11	2,47	2,35	2,65	1,13	1,07	1,21	1,24	0,00	0,90
C_15	1,37	1,75	3,66	1,82	2,31	2,55	3,04	2,34	2,52	1,80	0,98	0,92	1,50	0,90	0,00

	1	2	3	4	5	6
центры:	1,89; 1,56; -0,029	2,38 ; -1,27; -0,029	-0,394; -1,042 ; -1,932	-0,177;0,94;0,36	-0,33; -0,243; -0,1247	-0,635 ; -0,969; 1,302

Заметим, что совпадает с результатом иерархического алгоритма

Сравнение результатов с помощью многочлена качества

Сумма внутриклассовых дисперсий:

$$Q_1(S) = \sum_{l=1}^{K} \sum_{X_i \in S_l} \varrho^2(X_i, \overline{X}_l),$$

И сумма попарных внутриклассовых расстояний:

$$Q_2(S) = \sum_{l=1}^{K} \sum_{X_i, X_j \in S_l} \varrho^2(X_i, X_j)$$

Оценим качество разбиения для кластеров.

Q1(S) = 0
Q2(S) = 0
Q3(S) = 0,297
Q4(S) = 3,427
Q5(S) = 2, 1657
Q6(S) = 0,7415

$\sum =$ 6,6325

Качество разбиения среднее.

Вывод

Мы получили 6 кластеров после кластеризации

1) 6
2) 3
3) 8,9
4) 2,10,11,14,15
5) 1,4,12,13
6) 5,7