

```
## Warning in grepl(db, input): input string 26 is invalid in this locale
## Warning in grepl(db, input): input string 27 is invalid in this locale
## Warning in grepl(db, input): input string 28 is invalid in this locale
## Warning in grepl(db, input): input string 31 is invalid in this locale
## Warning in grepl(db, input): input string 36 is invalid in this locale
## Warning in grep("^\\\\bibliography.+", input, value = TRUE): input
string 26 is invalid in this locale
## Warning in grep("^\\\\bibliography.+", input, value = TRUE): input
string 27 is invalid in this locale
## Warning in grep("^\\\\bibliography.+", input, value = TRUE): input
string 28 is invalid in this locale
## Warning in grep("^\\\\bibliography.+", input, value = TRUE): input
string 31 is invalid in this locale
## Warning in grep("^\\\\bibliography.+", input, value = TRUE): input
string 36 is invalid in this locale
```

РОЗДІЛ 1

МОДЕЛЮВАННЯ ТА ПРАКТИЧНЕ ЗАСТОСУВАННЯ РОЗРОБЛЕНИХ МЕТОДІВ ТА АРХІТЕКТУР

1.1. Моделювання самонавчанняї нейро-фаззі системи, що еволюціонує

1.1.1. Придумати назву1. Одна з основних переваг, притаманних пропонуваній самонавчанняї нейро-фаззі системі, що еволюціонує, полягає в автоматичному визначенні оптимальної кількості кластерів та значення фазифікатору на кожному етапі оброблення даних. Першу серію експериментів було проведено на штучно зсинтезованих наборах даних з різним ступенем розмитості та перекриття класів аби дослідити вплив значення параметру фазифікації на якість кластерування в режимі реального часу відвідно до обраного критерію дійсності.

= 6

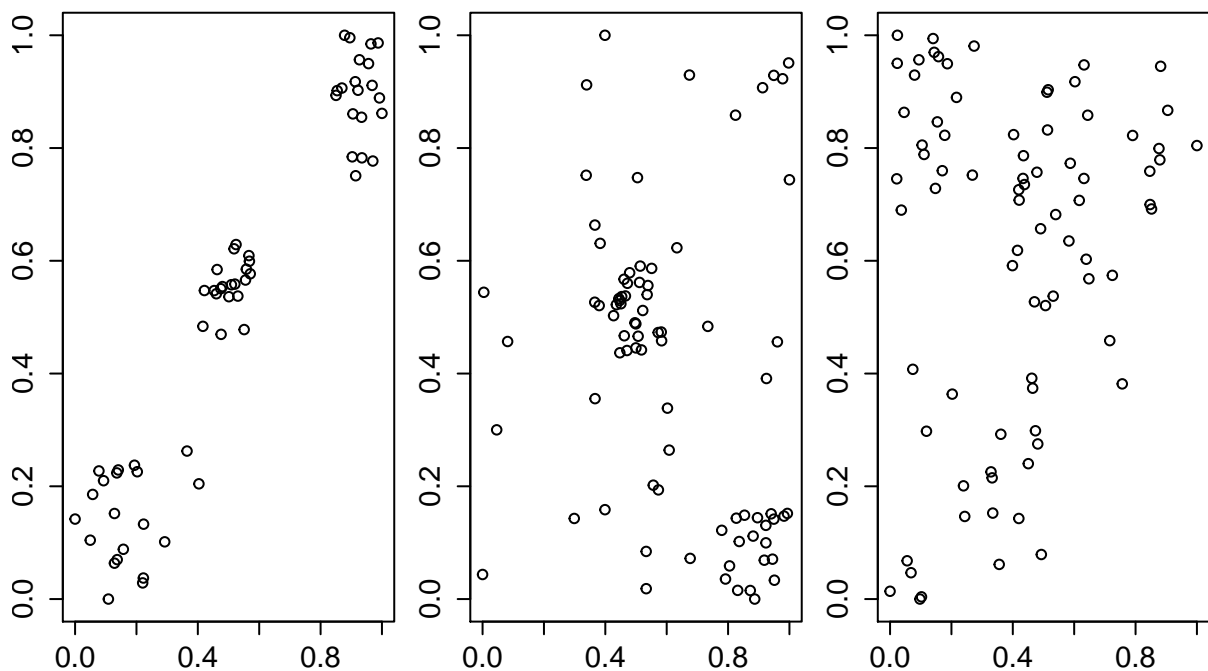


Рис. 1.1. Штучно сгенеровані набори даних

Кожен з наборів даних, що їх наведено на рис. 1.1 містить 80 спостережень з 2 ознаками (для наглядності) у кожному спостереженні. Тестові дані були сгенеровані таким чином, аби у першому наборі класи були чітко розподілені (crisp dataset), у другому наборі кластерні межі були дещо розмиті (fuzzy dataset), у третьому випадку класи сильно перетиналися (extra fuzzy dataset). Логічно припустити, що система, яка тестується, обере менше значення параметру фазифікації для першого датасету та більше для останнього, де межі класів спостережень є більш розмитими.

Спостереження надходили до нейро-фаззі мережі у послідовному режимі, вагові коефіцієнти нейронів були проініціалізовані використовуючи пакету модифікацію обраного алгоритму кластерування на датасеті з довільних двадцятьох спостережень відповідного набору даних (адже система, як і класичний fuzzy c-means, **чутлива до ініціалізації**). Локально оптимальні кількість кластерів та значення параметру фазифікації обумовлювалися максимальним середнім значенням рекурентних коефіцієнту розбиття РС (??) та Ксі-Бені індексу (??): $\max \frac{PC_j^{[m]} + 1 - XB_j^{[m]}}{2}$ (у данному випадку використовувалося від'ємне значення Ксі-Бені індексу $1 - XB(k)$, оскільки щоменше $XB_j^{[m]}$, толіпшим є розбиття даних на кластери).

Для першого набору даних (crisp dataset), як і передбачалося, оптимальним виявився другий каскад ($m = 3$) з трьома кластерами і нейроном-переможцем із найменшим значенням параметру фазифікації $\beta = 2$ (рис. 1.2). Така конфігурація є оптимальною відповідно до обох використовуваних індексів валідності – найменше значення Ксі-Бені індексу $XB_j^{[m]}$ та найбільший коефіцієнт розбиття $PC_j^{[m]}$:

$$PC_1^{[2]} = 0.9009951,$$

$$XB_1^{[2]} = 0.03349166.$$

Лише одне спостереження у цьому датасеті (його позначено багряним квадратом) не належить жодному кластерові з ступінем більшим від 0.6. Індокси

| | | | | |
|----------------------|-------------|-------------|-------------|-------------|
| Каскад 1 ($m = 2$) | $\beta = 2$ | $\beta = 3$ | $\beta = 4$ | $\beta = 5$ |
| Коефіцієнт розбиття | 0.91758 | 0.7446 | 0.64787 | 0.59236 |
| Індекс Ксі-Бені | 0.052129 | 0.061034 | 0.092235 | 0.1294 |
| Каскад 2 ($m = 3$) | $\beta = 2$ | $\beta = 3$ | $\beta = 4$ | $\beta = 5$ |
| Коефіцієнт розбиття | 0.92643 | 0.6609 | 0.50214 | 0.43305 |
| Індекс Ксі-Бені | 0.027232 | 0.06872 | 0.17281 | 0.26914 |
| Каскад 3 ($m = 4$) | $\beta = 2$ | $\beta = 3$ | $\beta = 4$ | $\beta = 5$ |
| Коефіцієнт розбиття | 0.87218 | 0.5256 | 0.37605 | 0.31993 |
| Індекс Ксі-Бені | 0.15687 | 0.4153 | 0.84699 | 1.1765 |
| Каскад 4 ($m = 5$) | $\beta = 2$ | $\beta = 3$ | $\beta = 4$ | $\beta = 5$ |
| Коефіцієнт розбиття | 0.73909 | 0.45445 | 0.32428 | 0.27063 |
| Індекс Ксі-Бені | 0.12985 | 0.30637 | 0.68584 | 1.0551 |

Таблиця 1.1

Індекси валідності (датасет 1)

валідності нейронів системи наведені у таблиці 5.1.

Для набору даних з середньою вираженістю класів найліпшим виявився нейрон другого каскаду ($m = 3$) і фаззифікатором $\beta = 3$ (таблиця 5.2).

Як показано на рис. 1.3, декілька спостережень у центрі (позначені багряними квадратами) можна віднести до 2 кластерів з відносно високим ступінем належності, проте більшість спостережень можна чітко розкластеризувати, що ілюструється високим значенням коефіцієнту розбиття, та дуже низьким Ксі-Бені індексом:

$$PC_2^{[2]} = 0.9727868,$$

$$XB_2^{[2]} = 0.087474.$$

Для набору з найменш чіткими межами класів (таблиця 5.3), система обрала нейроном-переможцем вузол третього каскаду ($m = 4$) з високим параметором фазифікації $\beta = 4$:

$$PC_3^{[3]} = 0.335525,$$

$$XB_3^{[3]} = 0.2128333.$$

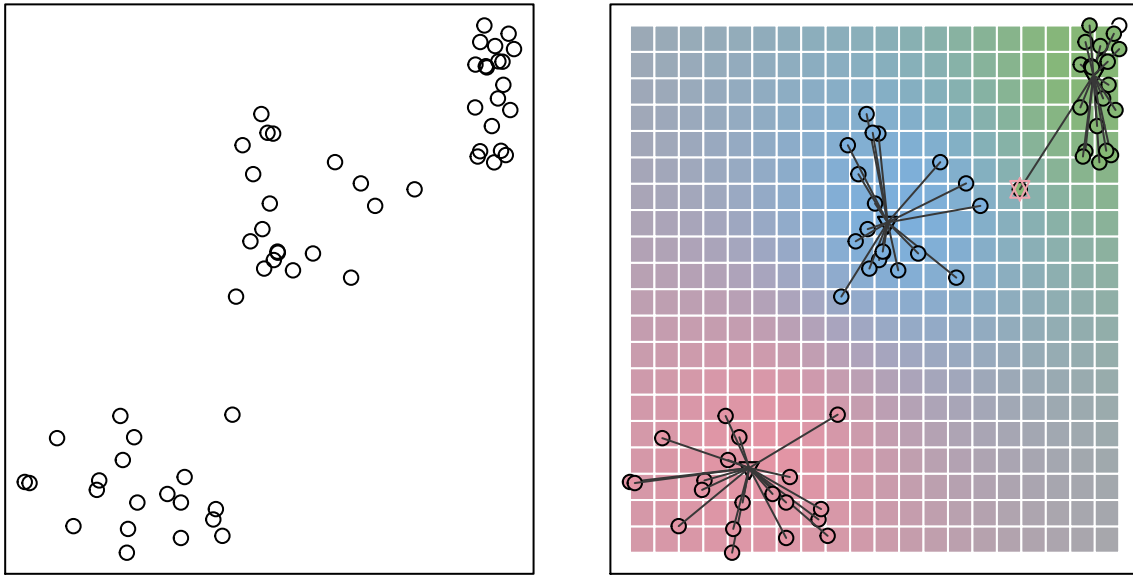


Рис. 1.2. Набір даних з чітко вираженими класами (Crisp dataset)

| Каскад 1 ($m = 2$) | $\beta = 2$ | $\beta = 3$ | $\beta = 4$ | $\beta = 5$ |
|----------------------|-------------|-------------|-------------|-------------|
| Коефіцієнт розбиття | 0.78414 | 0.58928 | 0.53853 | 0.52239 |
| Індекс Ксі-Бені | 0.16668 | 0.30834 | 0.3745 | 0.38723 |
| Каскад 2 ($m = 3$) | $\beta = 2$ | $\beta = 3$ | $\beta = 4$ | $\beta = 5$ |
| Коефіцієнт розбиття | 50084 | 0.71164 | 0.97275 | 0.4191 |
| Індекс Ксі-Бені | 0.009751 | 0.031235 | 0.087474 | 0.1323 |
| Каскад 3 ($m = 4$) | $\beta = 2$ | $\beta = 3$ | $\beta = 4$ | $\beta = 5$ |
| Коефіцієнт розбиття | 0.91888 | 0.47532 | 0.32777 | 0.28912 |
| Індекс Ксі-Бені | 0.052563 | 0.1757 | 0.27516 | 0.33766 |
| Каскад 4 ($m = 5$) | $\beta = 2$ | $\beta = 3$ | $\beta = 4$ | $\beta = 5$ |
| Коефіцієнт розбиття | 0.85618 | 0.34327 | 0.24778 | 0.22445 |
| Індекс Ксі-Бені | 0.048316 | 0.19887 | 0.34307 | 0.41228 |
| Каскад 5 ($m = 6$) | $\beta = 2$ | $\beta = 3$ | $\beta = 4$ | $\beta = 5$ |
| Коефіцієнт розбиття | 0.81295 | 0.30709 | 0.21636 | 0.19214 |
| Індекс Ксі-Бені | 0.060896 | 0.19702 | 0.31393 | 0.38668 |

Таблиця 1.2

Індекси валідності (датасет 2)

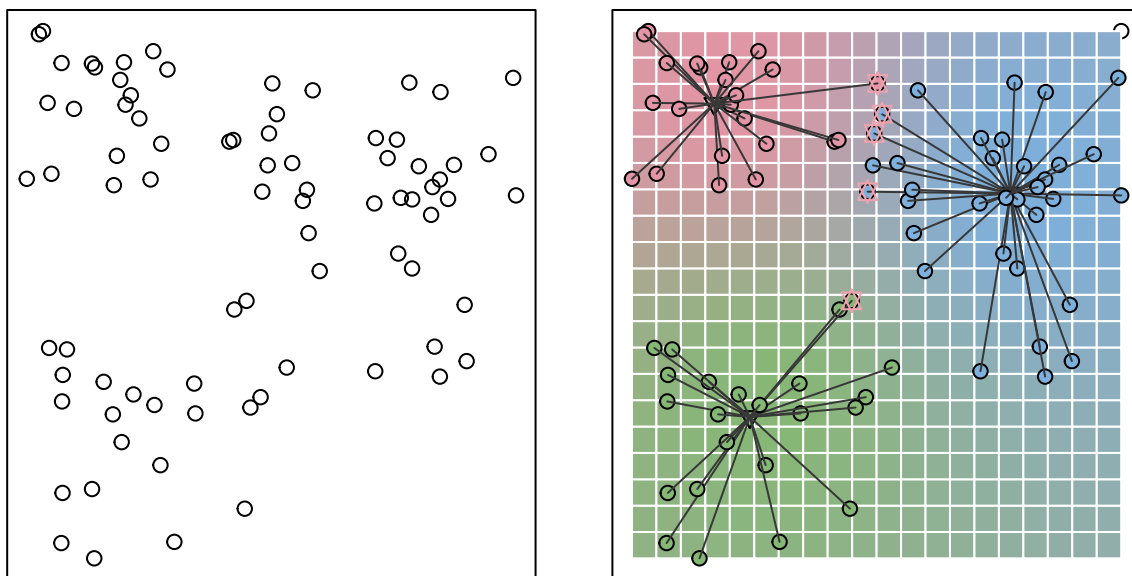


Рис. 1.3. Набір даних з нечіткими межами класів (fuzzy dataset)

На рис. 1.4 спостереження, для яких ступінь належності до будь-якого кластеру не перевищує 0.6, позначені багряними квадратами. Як і очікувалося, для цього набору даних кількість таких спостережень значно вища від попередніх датасетів з більш компактними та «чіткими» класами.

Для наглядності у всіх наведених рисунках кольором позначені не тільки розкластеровані спостереження і центри кластерів, а й задній план (фон) малюнків, що дозволяє візуально визначити, до якого кластеру система віднесла б нові спостереження. Не дивно, що, тоді як для перших двох датасетів важко визначити домінуючий колір, оскільки кластери їх спостережень більш менш компактні та явно виражені, для останнього набору даних домінуючий колір – сірий, сформований кольорами усіх кластерів, що ілюструє великий ступінь перекриття класів і, відповідно, високе значення оптимального параметру фазифікації β , що обрала система.

Ця низка експериментів проілюструвала як важливо вірно визначати параметр фазифікації, оптимальне значення якого у випадку оброблення даних у послідовному режимі з високою вирогідністю змінюється у часі, а саме зда-

| | | | | |
|----------------------|-------------|-------------|-------------|-------------|
| Каскад 1 ($m = 2$) | $\beta = 2$ | $\beta = 3$ | $\beta = 4$ | $\beta = 5$ |
| Коефіцієнт розбиття | 0.85094 | 0.71415 | 0.61734 | 0.57085 |
| Індекс Ксі-Бені | 0.10584 | 0.11462 | 0.13797 | 0.16101 |
| Каскад 2 ($m = 3$) | $\beta = 2$ | $\beta = 3$ | $\beta = 4$ | $\beta = 5$ |
| Коефіцієнт розбиття | 0.61668 | 0.42848 | 0.37779 | 0.35884 |
| Індекс Ксі-Бені | 0.1754 | 0.20364 | 0.22364 | 0.23995 |
| Каскад 3 ($m = 4$) | $\beta = 2$ | $\beta = 3$ | $\beta = 4$ | $\beta = 5$ |
| Коефіцієнт розбиття | 0.33458 | 0.44082 | 0.79405 | 0.29615 |
| Індекс Ксі-Бені | 0.20989 | 0.129 | 0.051039 | 0.26282 |
| Каскад 4 ($m = 5$) | $\beta = 2$ | $\beta = 3$ | $\beta = 4$ | $\beta = 5$ |
| Коефіцієнт розбиття | 0.50244 | 0.33067 | 0.26029 | 0.23318 |
| Індекс Ксі-Бені | 0.37268 | 0.61417 | 0.79695 | 0.93626 |
| Каскад 5 ($m = 6$) | $\beta = 2$ | $\beta = 3$ | $\beta = 4$ | $\beta = 5$ |
| Коефіцієнт розбиття | 0.53279 | 0.29731 | 0.22648 | 0.19858 |
| Індекс Ксі-Бені | 0.27407 | 0.47298 | 0.60569 | 0.70716 |

Таблиця 1.3

Індекси валідності (датасет 3)

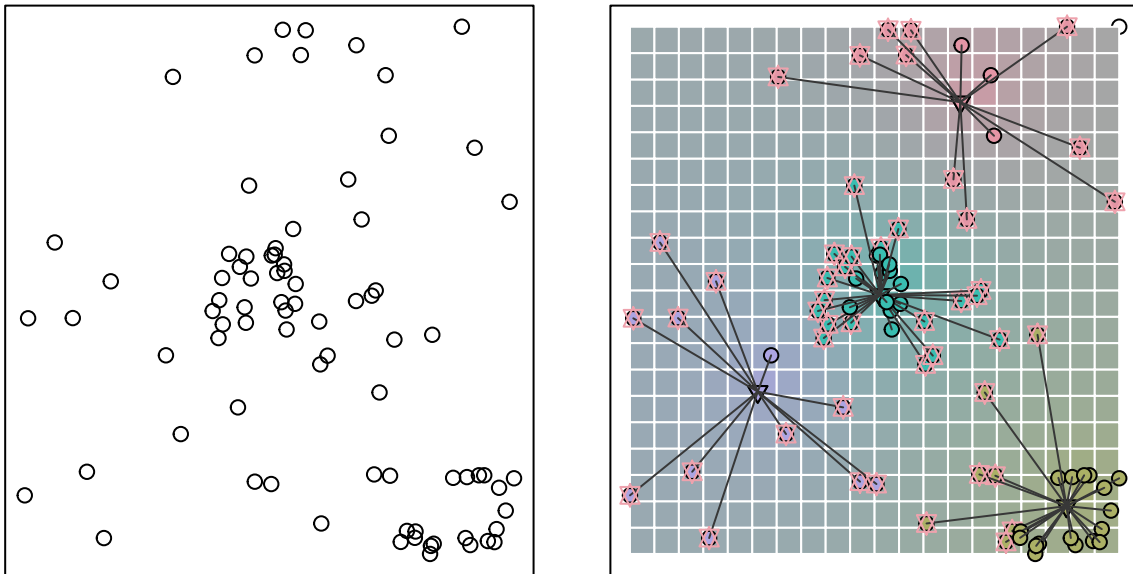


Рис. 1.4. Набір даних з класами, що перетинаються (extra fuzzy dataset)

тність визначати оптимальне значення цього параметру в онлайн режимі є відмінною особливістю запропонованої самонавчаної нейро-системи.

1.1.2. Придумати назву2. Наступна низка експериментів була проведена на наборі даних «Іриси Фішера» (Fisher's Iris data set).

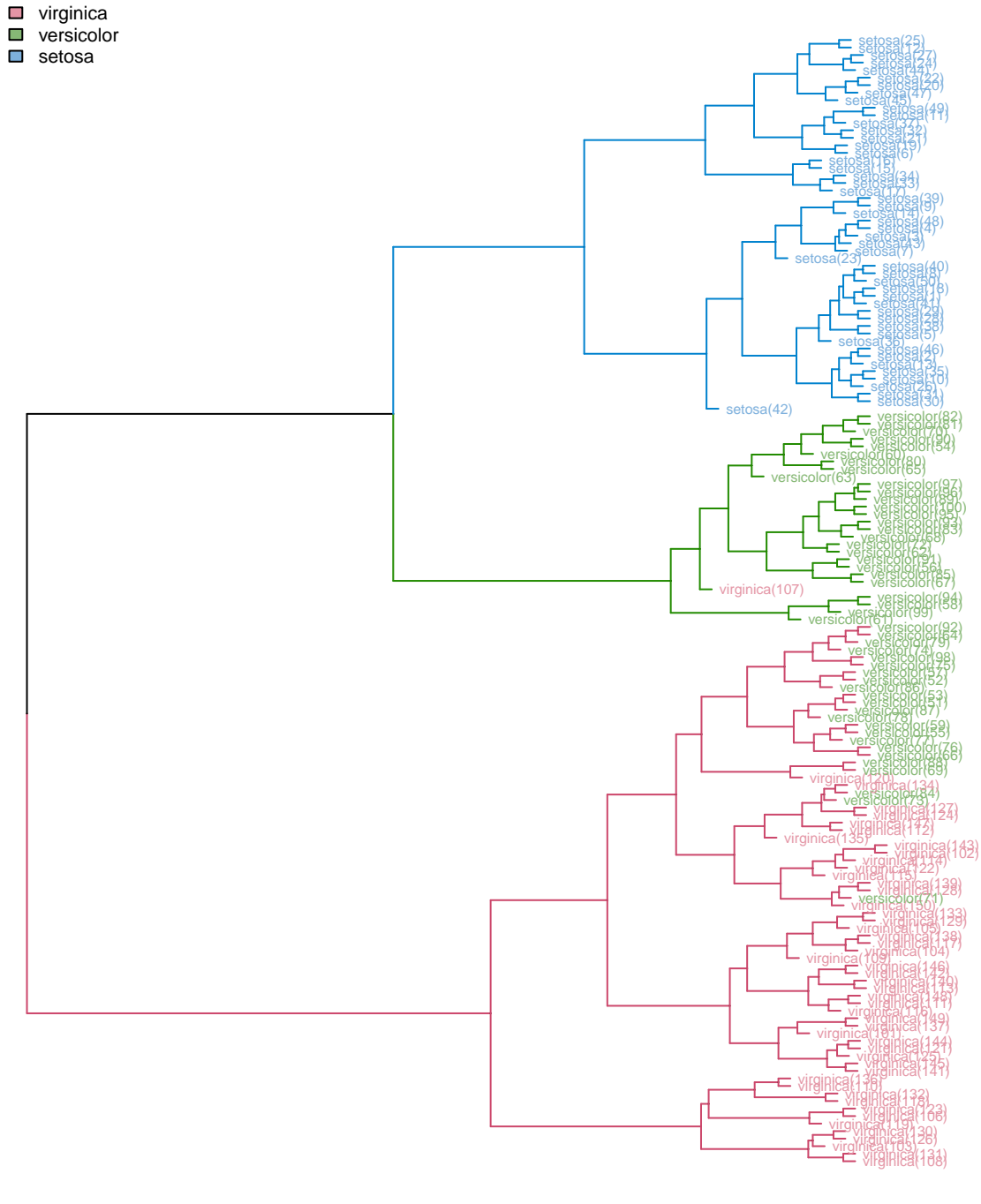


Рис. 1.5. Ієрархічне класування датасету «Іриси Фішера»

Це багатовимірний датасет для задачі класифікації, на прикладі якого англійський статистик та біолог Рональд Фішер в 1936 році продемонстрував

роботу розробленого ним методу дискримінантного аналізу. Іноді його також називають «Ірисами Андерсона» (через те, що дані були зібрані американським ботаніком Едгаром Андерсоном). Цей набір даних став класичним і часто використовується в літературі для ілюстрації роботи різних статистичних алгоритмів.

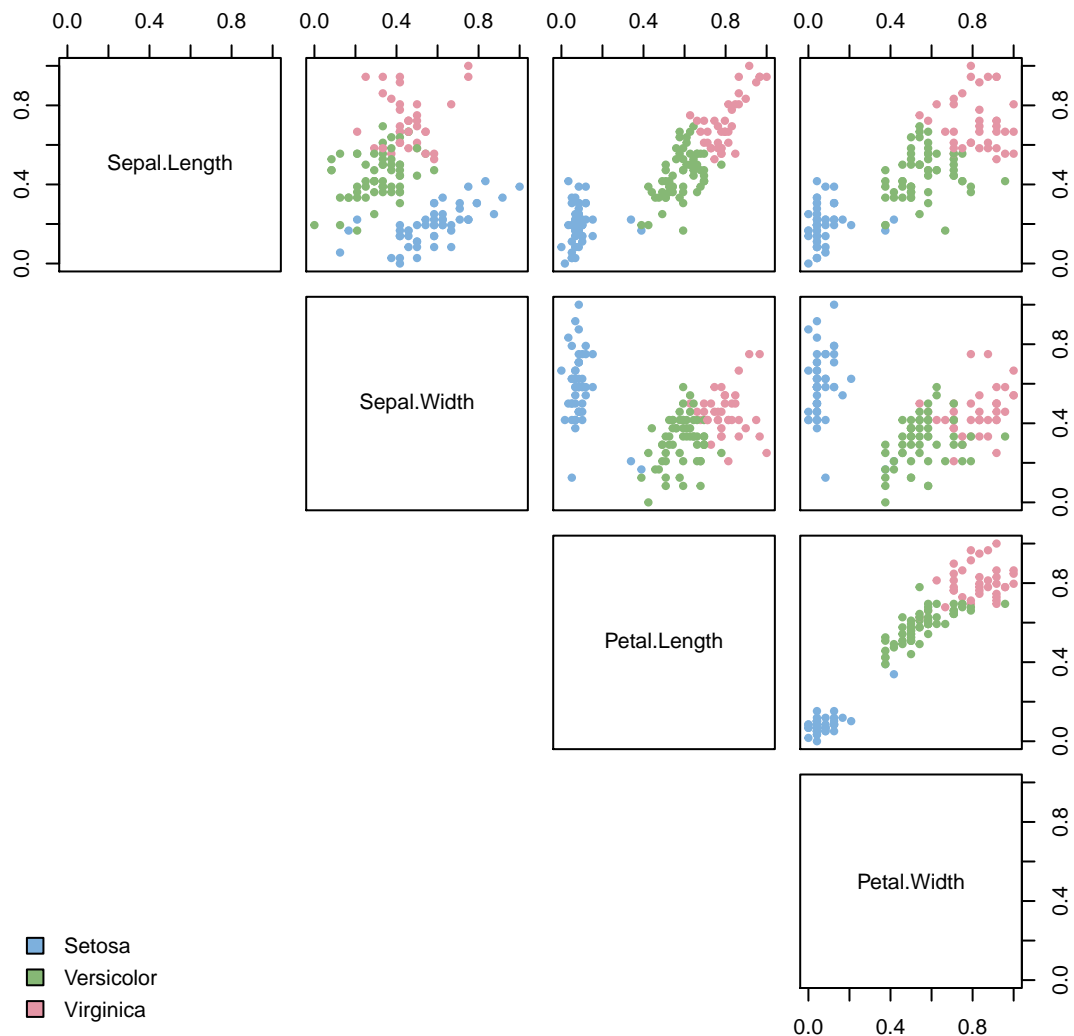


Рис. 1.6. Розкластерований датасет «Іриса Фішера» при $m = 3$, $\beta = 2$ (Точність кластерування – 96%)

Проте цей датасет рідко використовується у кластерному аналізі, адже межі класів «Versicolor» та «Virginica» не можна чітко визначити, ґрунтуючись на даних, що їх використовував Фішер (що легко продемонструвати за допомогою ієрархічного кластерування, рис. 1.5). Саме цим і цікавий для

нас цей набір даних: коли класичні методи чіткого кластерного аналізу не справляються з задачею, може стати у нагоді система, що реалізує нечітке кластерування зі змінним параметром фаззифікації та кількістю кластерів. Для більшості методів кластерного аналізу, зокрема для методу нечітких середніх (fuzzy c-means), необхідно заздалегідь задати кількість кластерів, і очевидним рішенням є прийняти $m = 3$, адже маємо три класи: Iris Verginica, Iris Versicolor та Iris Setosa (рис. 1.6).

| | $\beta = 2$ | $\beta = 3$ | $\beta = 4$ | $\beta = 5$ |
|-----|-------------|-------------|-------------|-------------|
| avg | 0.8313073 | 0.8741245 | 0.8709475 | 0.8888124 |
| min | 0.7859722 | 0.7533766 | 0.6615745 | 0.7656498 |
| max | 0.8534013 | 0.9166667 | 0.935051 | 0.9604701 |

Таблиця 1.4

Точність кластерування при $m = 3$

Точність кластерування за допомогою методу нечітких середніх за таких умов ($m = 3$, $\beta = 2$) рідко перевищує 83% (таблиця 5.4). (Оскільки для обраного датасету існують мітки з вірною класифікацією, ефективність кластеризації вимірювалася у відсотках точності щодо еталонного розбиття після дефаззифікації.) Проте, якщо не обмежувати пропоновану систему у кількості кластерів (система ініціалізується інтервалом допустимих значень m (кількість кластерів) та параметру фаззифікації β), вельми цікавими є результати кластерування нейронів кожного з каскадів.

У таблиці 5.5 наведена точність розбиття даних, коли $m \gg 3$ кластерів відповідно. Варто зазначити, що нейрони у пулі кожного каскаду реалізують метод нечітких середніх зі змінним значення фазифікатору, а отже є чутливими до довільно ініціалізованих цетрах кластерів, тому у таблицях наведені середня, мінімальна та максимальна точності кластерування (після дефаззифікації).

| $m = 7$ | $\beta = 2$ | $\beta = 3$ | $\beta = 4$ | $\beta = 5$ |
|----------|-------------|-------------|-------------|-------------|
| avg | 0.8972948 | 0.9150268 | 0.9242503 | 0.9178207 |
| min | 0.8536056 | 0.8461905 | 0.8723182 | 0.8600289 |
| max | 0.9621849 | 0.9736172 | 0.9810146 | 0.9663462 |
| $m = 8$ | $\beta = 2$ | $\beta = 3$ | $\beta = 4$ | $\beta = 5$ |
| avg | 0.9065560 | 0.9296311 | 0.9243606 | 0.9248976 |
| min | 0.8217056 | 0.8562179 | 0.8577202 | 0.8590278 |
| max | 0.9474588 | 0.9789402 | 0.9848214 | 0.9747899 |
| $m = 9$ | $\beta = 2$ | $\beta = 3$ | $\beta = 4$ | $\beta = 5$ |
| avg | 0.9258154 | 0.9282887 | 0.9308971 | 0.9229753 |
| min | 0.8689921 | 0.8270525 | 0.8684641 | 0.8556390 |
| max | 0.9849170 | 0.9806397 | 0.9664112 | 0.9748284 |
| $m = 10$ | $\beta = 2$ | $\beta = 3$ | $\beta = 4$ | $\beta = 5$ |
| avg | 0.9213191 | 0.9285106 | 0.9332528 | 0.9282907 |
| min | 0.8663370 | 0.8722271 | 0.8652272 | 0.8766667 |
| max | 0.9663420 | 0.9838095 | 0.9723656 | 0.9756335 |
| $m = 11$ | $\beta = 2$ | $\beta = 3$ | $\beta = 4$ | $\beta = 5$ |
| avg | 0.9295315 | 0.9408977 | 0.9317242 | 0.9295800 |
| min | 0.8520268 | 0.8964924 | 0.8890781 | 0.8788656 |
| max | 0.9716166 | 0.9848485 | 0.9704892 | 0.9798627 |
| $m = 12$ | $\beta = 2$ | $\beta = 3$ | $\beta = 4$ | $\beta = 5$ |
| avg | 0.9349407 | 0.9433244 | 0.9337934 | 0.9306632 |
| min | 0.8815133 | 0.8949802 | 0.8798160 | 0.8486111 |
| max | 0.9795274 | 0.9783497 | 0.9630952 | 0.9772727 |
| $m = 13$ | $\beta = 2$ | $\beta = 3$ | $\beta = 4$ | $\beta = 5$ |
| avg | 0.9420998 | 0.9398127 | 0.9375204 | 0.9357708 |
| min | 0.8823175 | 0.8614025 | 0.8882479 | 0.8828348 |
| max | 0.9807518 | 0.9788034 | 0.9753452 | 0.9748873 |

Таблиця 1.5

Точність кластерування для $m \in [7, 13]$, $\beta \in [2, 5]$

На рис 1.7 зображено залежність точності кластерування від кількості кластерів. Цікаво, що при, здавалося б, очевидному рішенні обрати кількість кластерів рівною трьом, отримуємо чиненаягіршу точнічть кластерування (при $\beta = 2$) після дефаззифікації щодо еталонного розбиття (Для порівнян-

ня на рис. 1.8 та рис. 1.9 наведені розбиття, що їх запропонували нейроні-переможці деяких каскадів, де $m \gg 3$).

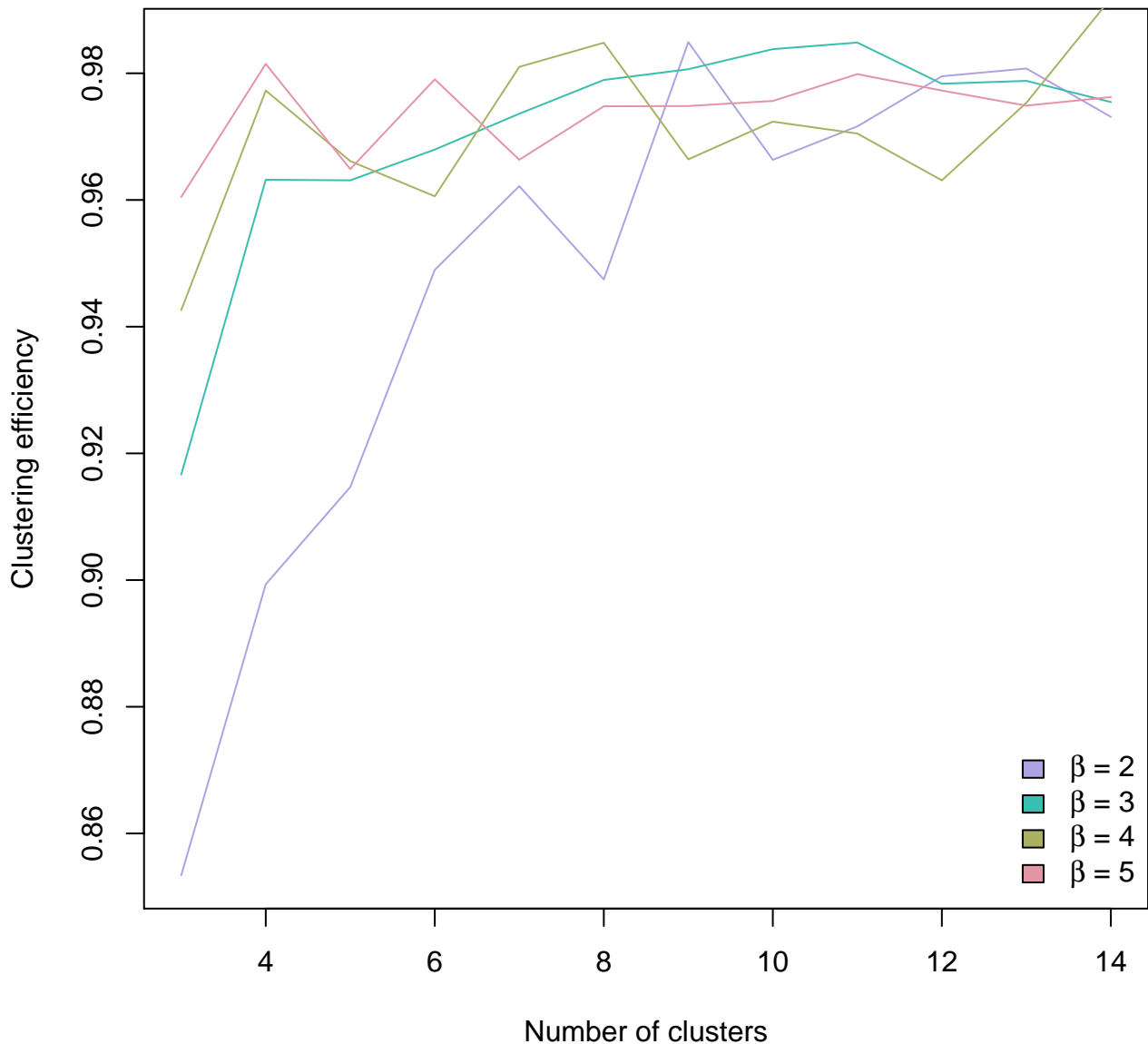


Рис. 1.7. Точність кластерування від кількості кластерів та параметру фаззифікації

Цьому легко знайти пояснення, адже метод нечітких k -середніх (а саме цей метод у цьому експерименті реалізують вузли пулів кожного каскаду) добре розпізнає кластери лише гіперсферичної форми. Проте кластер довільної (негіперсферичної) форми, можна розбити на декілька гіперсферичних

підкластерів, що й відбувається у каскадах, де $m > 3$, що пропонують розбиття на дрібні кластери. На рисунках 1.10 та 1.11 наведені розбиття деяких каскадів, де кількість кластерів більша від кількості класів еталонної вибірки; тут можна побачити, що декілька кластерів, що після дефазифікації будуть віднесені до одного класу, наприклад, Iris Virginica розташовані поруч один з одним, тобто є складовими більшого кластеру негіперсферичної форми.

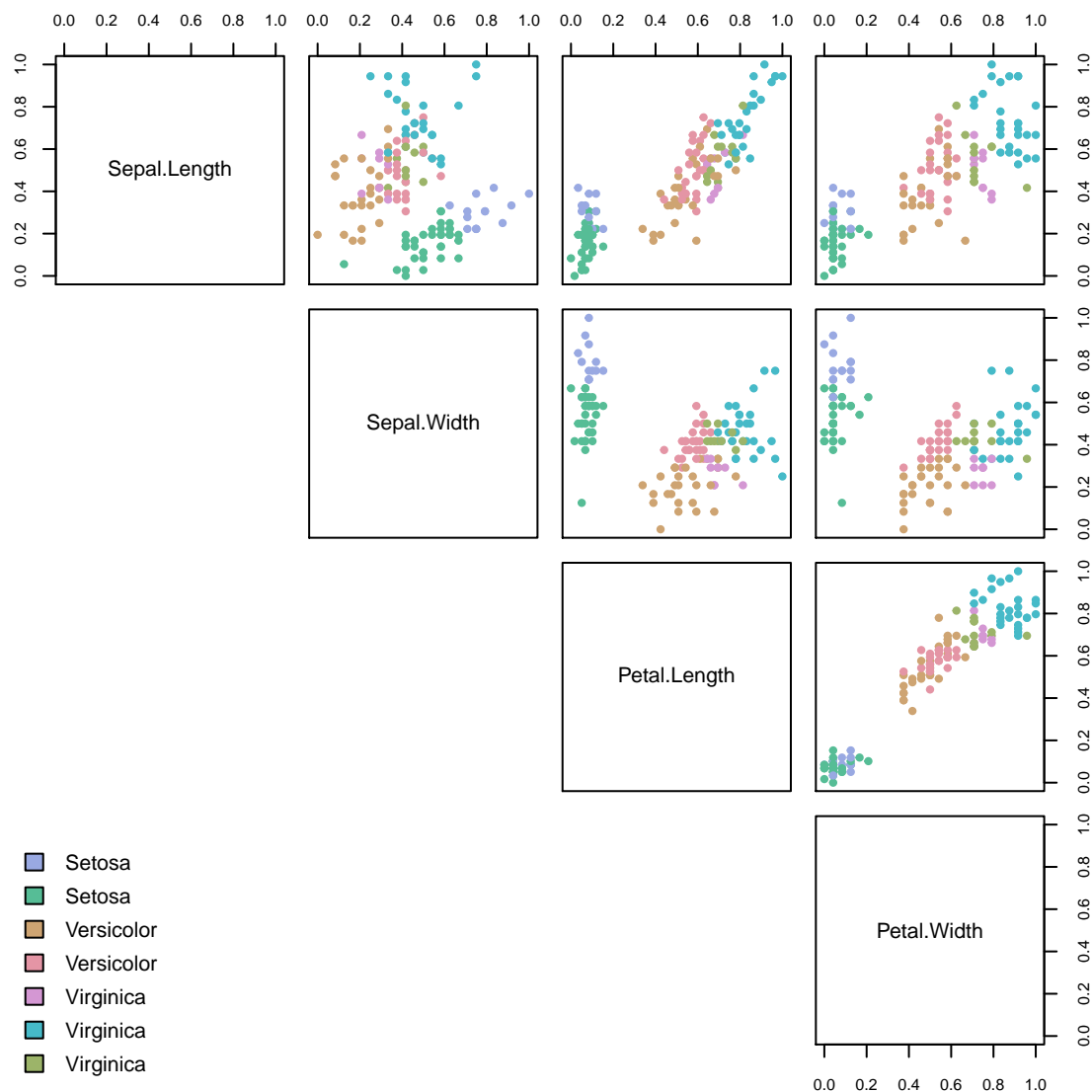


Рис. 1.8. Розкластерований датасет «Іриси Фішера» при $m = 7$, $\beta = 5$ (Точність кластерування $\approx 93\%$)

Таким чином, видається доречним, навіть у випадку, коли відоме еталонне розбиття датасету, дозволити системі обрати кінцеву кількість кластерів

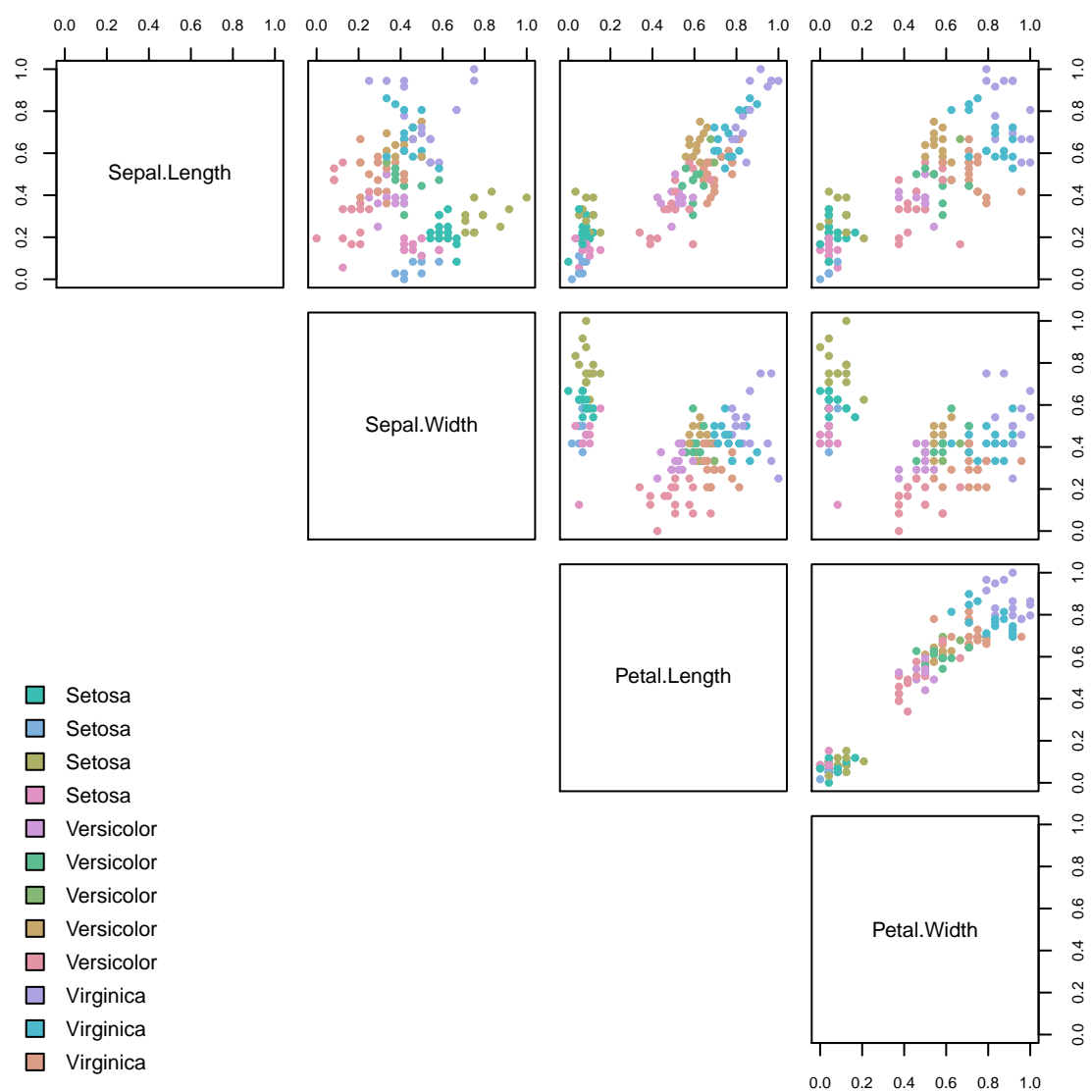


Рис. 1.9. Розкластерований датасет «Іриси Фішера» при $m = 12$, $\beta = 4$ (Точність кластерування $\approx 96\%$)

| | $\beta = 2$ | $\beta = 3$ | $\beta = 4$ | $\beta = 5$ |
|-----|-------------|-------------|-------------|-------------|
| avg | 0.9369168 | 0.9448829 | 0.9383179 | 0.9403416 |
| min | 0.8847819 | 0.8953380 | 0.8787879 | 0.9069805 |
| max | 0.9731262 | 0.9754579 | 0.9918301 | 0.9762515 |

Таблиця 1.6

Точність кластерування при $m = 14$

самостійно, особливо у випадку, коли вузли системи реалізують однаковий метод кластерування.

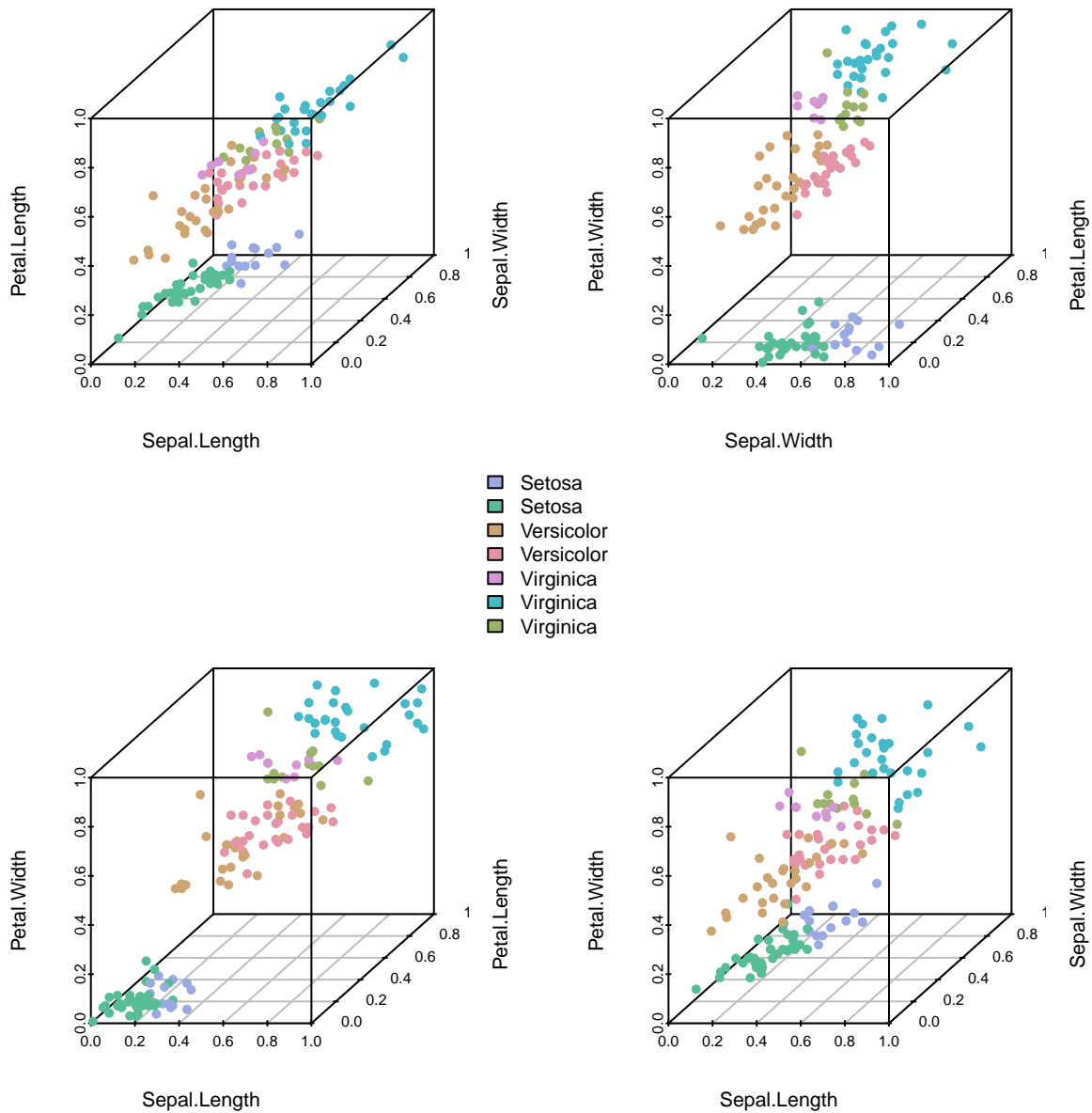


Рис. 1.10. Розкластерований датасет «Іриси Фішера» при $m = 7$, $\beta = 5$ (Точність кластерування $\approx 93\%$)

Варто зауважити, що у цьому випадку для визначення локально оптимального розбиття доцільно використовувати модифіковані індекси валідності, чи такі, що не залежать від відстані центрів кластерів, наприклад ті, що

ґрунтуються на щільності (density-based).

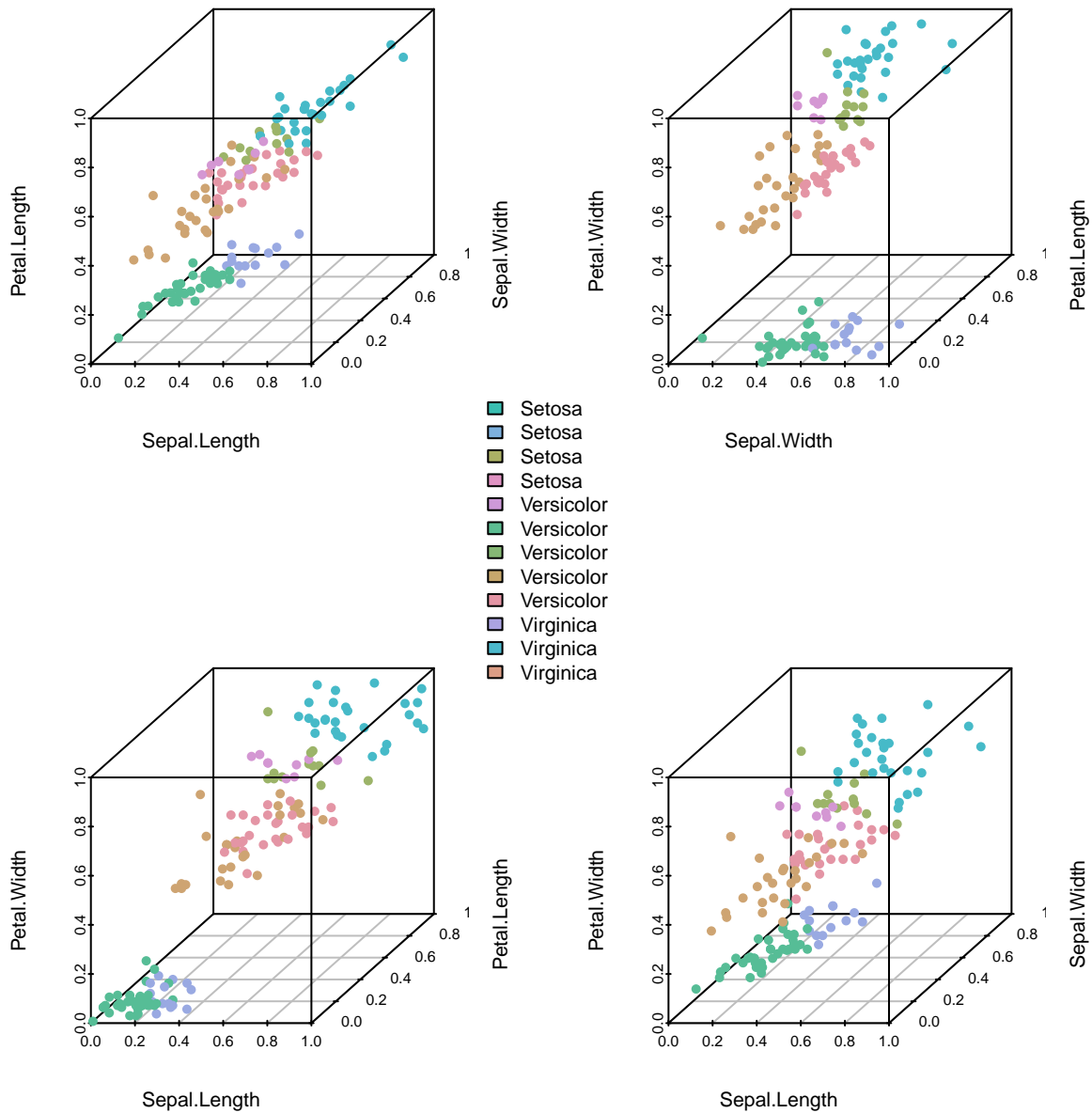


Рис. 1.11. Розкластерований датасет «Іриси Фішера» при $m = 12$, $\beta = 4$ (Точність кластерування $\approx 96\%$)

1.1.3. Придумати назву 3. Наступну серію експериментів було проведено на датасеті «Знання студентів про електричні машини постійного струму».

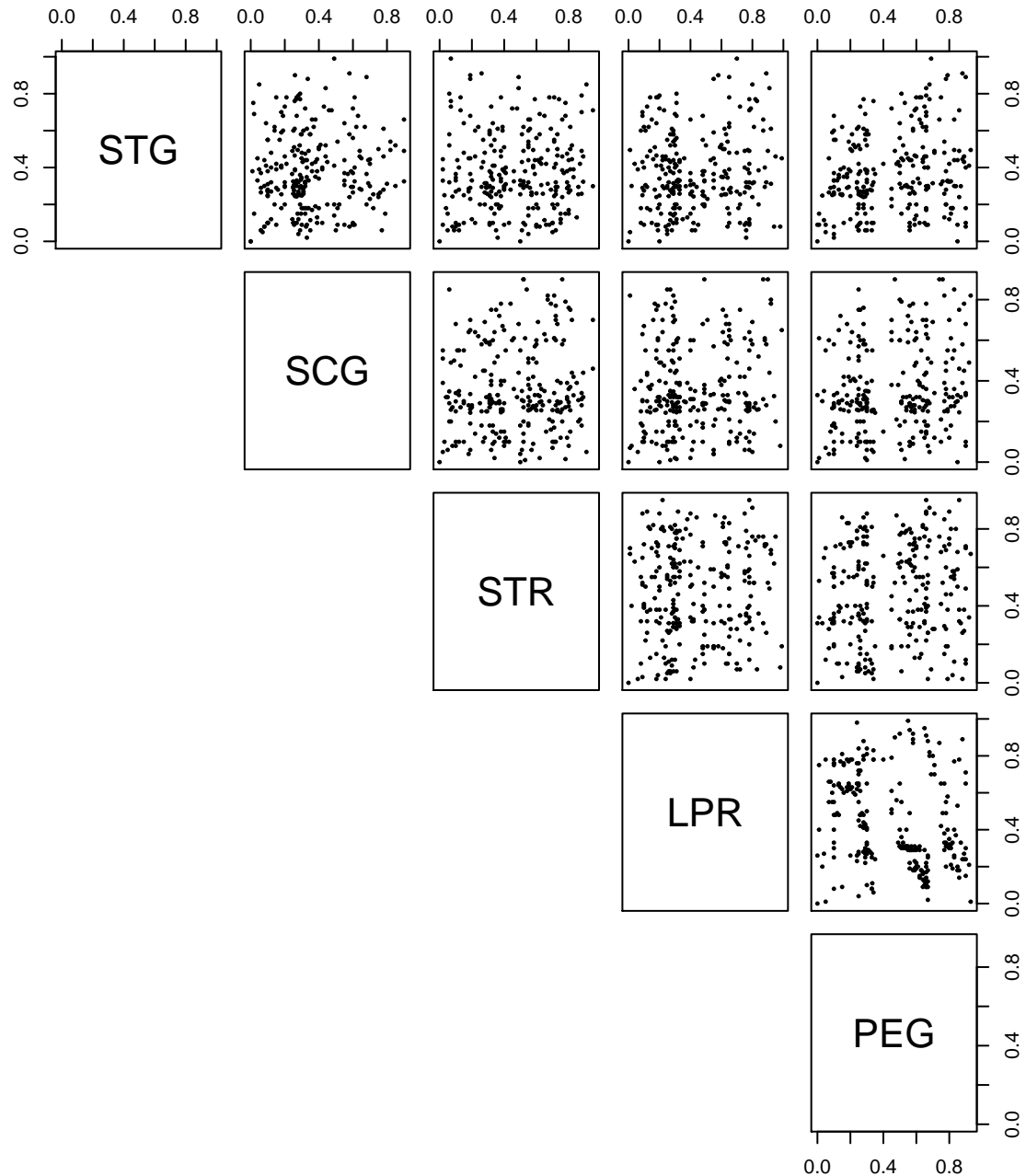


Рис. 1.12. Датасет «Знання студентів про електричні машини постійного струму»

Цей датасет було додано до UCI репозиторію у 2013 році, він містить 403 паттерни, кожен з п'ятьма атрибутами:

1. STG: кількість часу, що його витратив(витратила) студент(ка) на навчання цільового матеріалу,
2. SCG: Кількість повторювань навчання цільового матеріалу студен-

том(студенткою),

3. STR: Кількість часу, що його використав(використала) студент(ка) на навчання матеріалу, пов'язаного з цільовим матеріалом,
4. LPR: Оцінка, що її отримав(отримала) студент(ка) на іспиті з предмету, пов'язаного з цільовим предметом,
5. PEG: Оцінка, що її отримав(отримала) студент(ка) на іспиті з цільового предмету,

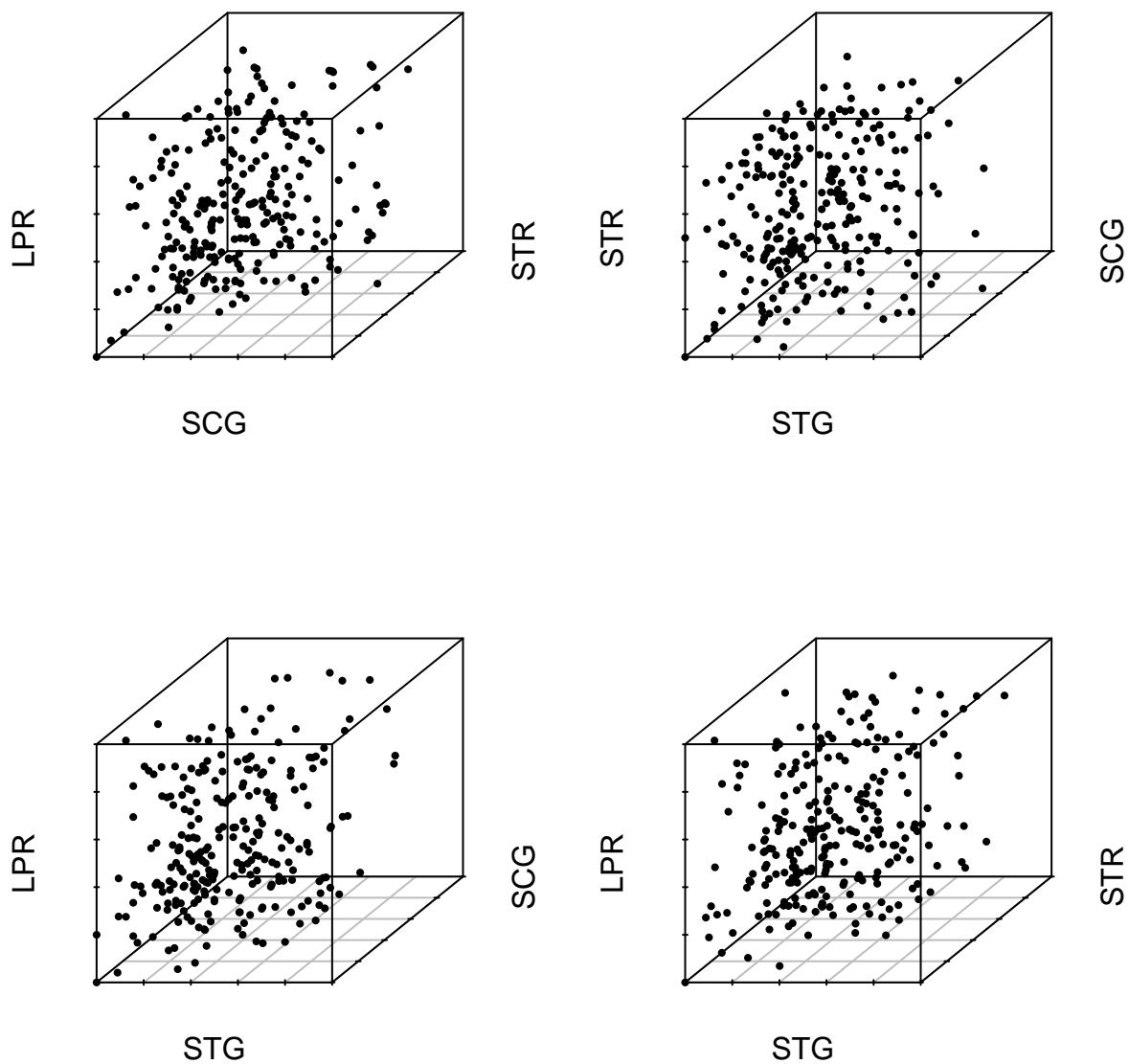


Рис. 1.13. Датасет «Знання студентів про електричні машини постійного струму»

Попарні графіки атрибутів наведено на рис. 1.12 та у тримірному просторі на рис. 1.13.

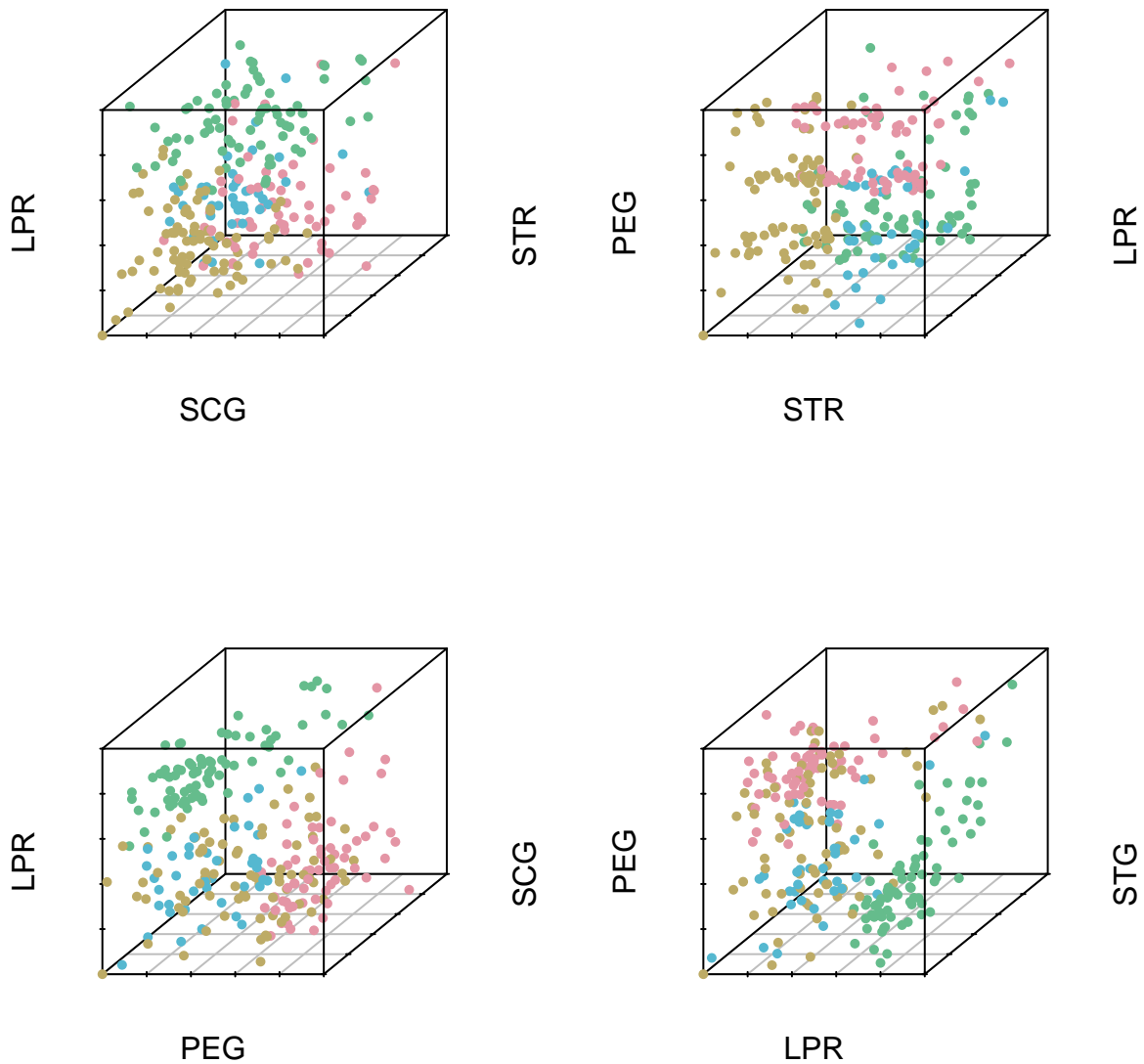


Рис. 1.14. Датасет «Знання студентів про електричні машини постійного струму»

Для цього експерименту нейрони-узагальнювачі керувалися рекурентним Ксі-Бені Індексом при визначанні локально-оптимального нейрона (з найліпшим параметром фаззифікації) та каскаду (з оптимальною кількістю кластерів).

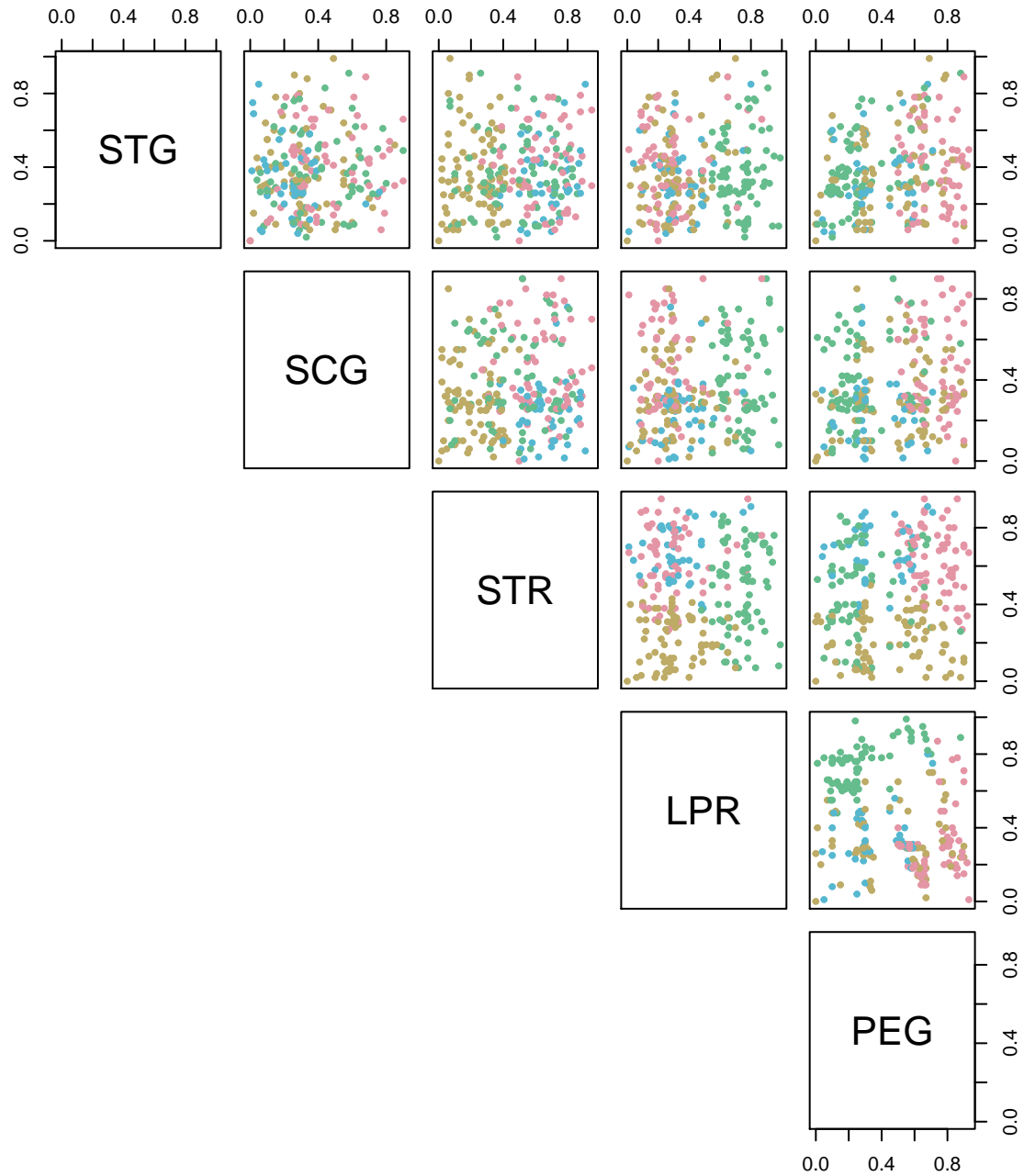


Рис. 1.15. Датасет «Знання студентів про електричні машини постійного струму»

Оптимальне розбиття, що його наведено на рис.1.14 та на рис.1.15, надав другий нейрон третього каскаду з $m = 4$, $\beta = 2$ з коефіцієнтом Ксі-Бені 0.38155.