Моделі «копула» в задачах баєсівської класифікації при машинному навчанні

Виконав
студент групи КА-51м
Харченко Володимир

Задача класифікації

- маємо навчальну вибірку з N об'єктів, кожен з яких належить одному з K класів і характеризується набором m числових ознак $a_1, a_2, \ldots a_m$
- n_k - число об'єктів k-ого класу, так що $N = \sum_{k=1}^{K} n_k$
- значення j-ого ознаки i-ого об'єкта з k-ого класу позначимо x_{ijk}

Тоді цей об'єкт можна охарактеризувати вектором-рядком $x_{ik} = (x_{i1k}, \ldots, x_{ijk}, \ldots, x_{imk})$. Цей рядок будемо розглядати як i-у реалізацію векторної випадкової величини ξ_k, що підкоряється розподілу ймовірностей з щільністю $p\,(x_1, \ldots, x_m \mid k)$, своєю для кожного класу k.

Нехай тепер спостерігається об'єкт, для якого необхідно визначити, до якого класу він належить. Об'єкт характеризується тільки набором m числових ознак $x_1, \ldots, x_m$.

Постановка задачі

1. Побудувати баєсівські класифікатори на основі різних сімейств копул (гаусівські, Стьюдента, Клейтона, Гумбеля, Франка).

2. Для чотирьох виборок даних:

- обрати сімейство копул і параметри копула-функції, що найбільш точно відображають залежність між даними;
- оцінити точність такого вибору;
- застосувати класифікатори на основі копула-функцій;
- порівняти класифікатори за точністю класифікації.

Копула-функція

Копула – це функція , яка має наступні властивості:

a) Для кожного u з I^n:

- $C(u) = 0$, якщо хоча б одна координата u нульова;
- $C(u) = u_k$, якщо всі координати u дорівнюють 1, крім u_k.

b) Для кожного a, b з I^n таких, що $a \leq b$:

$$V_C([a, b]) \geq 0$$

де $V_H(b) = \sum sign(c)H(c)$, H - дійсна n-мірна функція,

$\mathrm{B} = [a, b] -$ n-мірний куб.

Теорема Скляра

Нехай – n-мірна функція розподілу з маргіналами $F_1, F_2, \dots F_n$. Тоді існує n-копула C така, що $\forall x \in \mathbb{R}^n$:

$$H(x_1, x_2, \dots x_n) = \mathrm{C}(F_1(x_1), F_2(x_2), \dots F_n(x_n))$$

Модифікація баєсівського класифікатора

- Класичний баєсівський класифікатор

$$\hat{k} = \underset{k}{\operatorname{argmax}} \ln P_k p(x_1, \ldots, x_m \mid k) \,/\, p(x_1, \ldots, x_m)$$

- Класифікатор на основі копула-функції

$$\hat{k} = \underset{k}{\operatorname{argmax}} \left(P_k \, c^k({F_1}^k(x_1 \mid \theta_1^k), \ldots, {F_m}^k(x_m \mid \theta_m^k)) \prod_{i=1}^{m} {F_i}^k(x_i \mid \theta_i^k) \right)$$

Вибірки даних

- Іриси (150 зразків, 4 ознаки, 3 класи)
- Насіння (210 зразків, 7 ознак, 3 класи)
- Вино (4184 зразків, 8 ознак, 3 класи)
- Vertebral (310 зразків, 6 ознак, 3 класи)

Методи вибору параметрів копул

Метод виводу функцій маргінальних розподілів (ВФМР)

$$l(\xi) = \sum_{t=1}^{T} \ln c\left(F_1(x_1^t \mid \xi_1), \dots, F_d(x_d^t \mid \xi_d), \alpha\right) + \sum_{t=1}^{T} \sum_{i=1}^{d} \ln f_i\left(x_i^t \mid \xi_i\right),$$

де $\xi = (\xi_1, \dots, \xi_d)$ представляє параметри маргінальних розподілів і α представляє параметри копула-функції.

Оцінювання проводиться в два етапи:

$$\hat{\xi}_i = \arg\max_{\xi \in \Xi} \sum_{t=1}^{T} \ln f_i\left(x_i^t \mid \xi_i\right)$$

$$\hat{\alpha} = \arg\max_{\alpha} \sum_{t=1}^{T} \ln c\left(F_1(x_1^t \mid \xi_1), \dots, F_d(x_d^t \mid \xi_d), \alpha\right)$$

Методи вибору параметрів копул

Канонічний метод максимальної правдоподібності (КММП)

Емпіричні функції розподілу:

$$\hat{F}_{x_i}(x) = \frac{1}{T+1}\sum_{t=1}^{T} I(x_i^t \leq x)$$

Оцінка параметрів:

$$\hat{\alpha} = \arg\max_{\alpha} \sum_{t=1}^{T} \ln c\,(u_1^t, \dots, u_d^t, \alpha),$$

де $(u_1^t, \dots, u_d^t)_{t=1}^{T} = \left(\hat{F}_1(x_1^t), \dots, \hat{F}_d(x_d^t)\right)_{t=1}^{T}$

Критерії точності оцінювання

- Інформаційний критерій Акайке (ІКА)

$$\text{ІКА} = \ 2k\frac{T}{T-k-1} - 2l$$

- Баєсівський інформаційний критерій (БІК)

$$\text{БІК} = k\log(T) - 2l$$

- Критерій SnC

$$SnC = \sum_{t=1}^{T}\left(\hat{C}(e_1^t, \dots, e_d^t) - \ \Pi(e_1^t, \dots, e_d^t)\right)^2$$

Використані канонічні розподіли

- бета-розподіл;
- експоненційний;
- нормальний;
- гама-розподіл;
- зворотний гауссівський;
- логістичний;
- лог-логістичний;
- лог-нормальний;
- Нормальний;
- Стьюдента;
- Вейбулла.

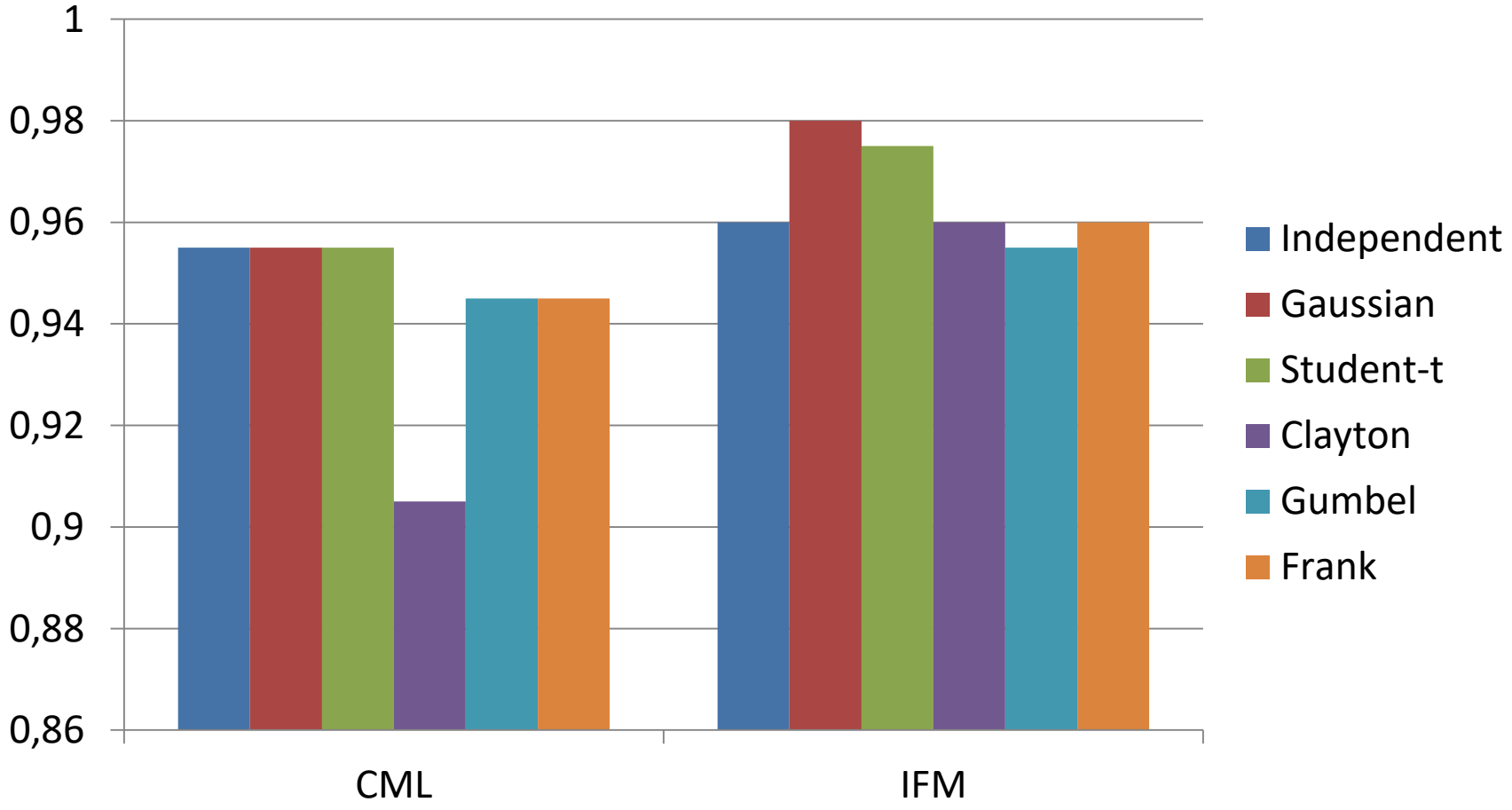
Використані копула-функції

- Гаусівські;
- Стьюдента;
- Клейтона;
- Гумбеля;
- Франка.

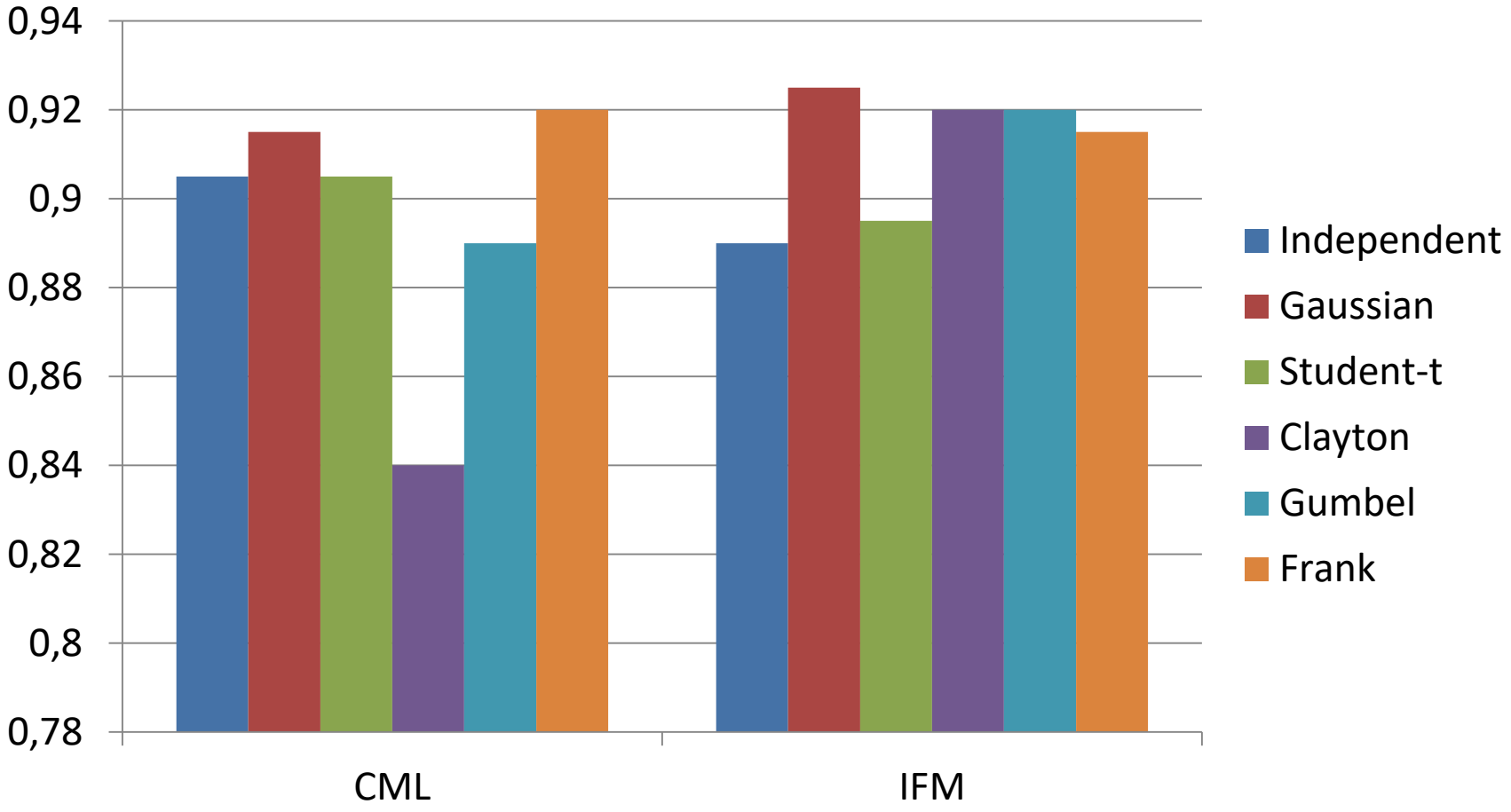
Алгоритм роботи з вибірками

- Для набору стандартних розподілів були отримані емпіричні оцінки параметрів за кожним з класів навчальної вибірки.
- Було обрано маргінальні розподіли, оцінка параметрів яких була проведена найкращим чином (за найменшим значенням БІК).
- Для обраних маргінальних розподілів за допомогою ВМР було оцінено оптимальні параметри для різних сімейств копул.
- Аналогічне оцінювання проведено для методу КММП.
- Наведено порівняння статистичних характеристик отриманих оцінок, з яких можна робити висновки про доцільність застосування для класифікації копул різних сімейств.

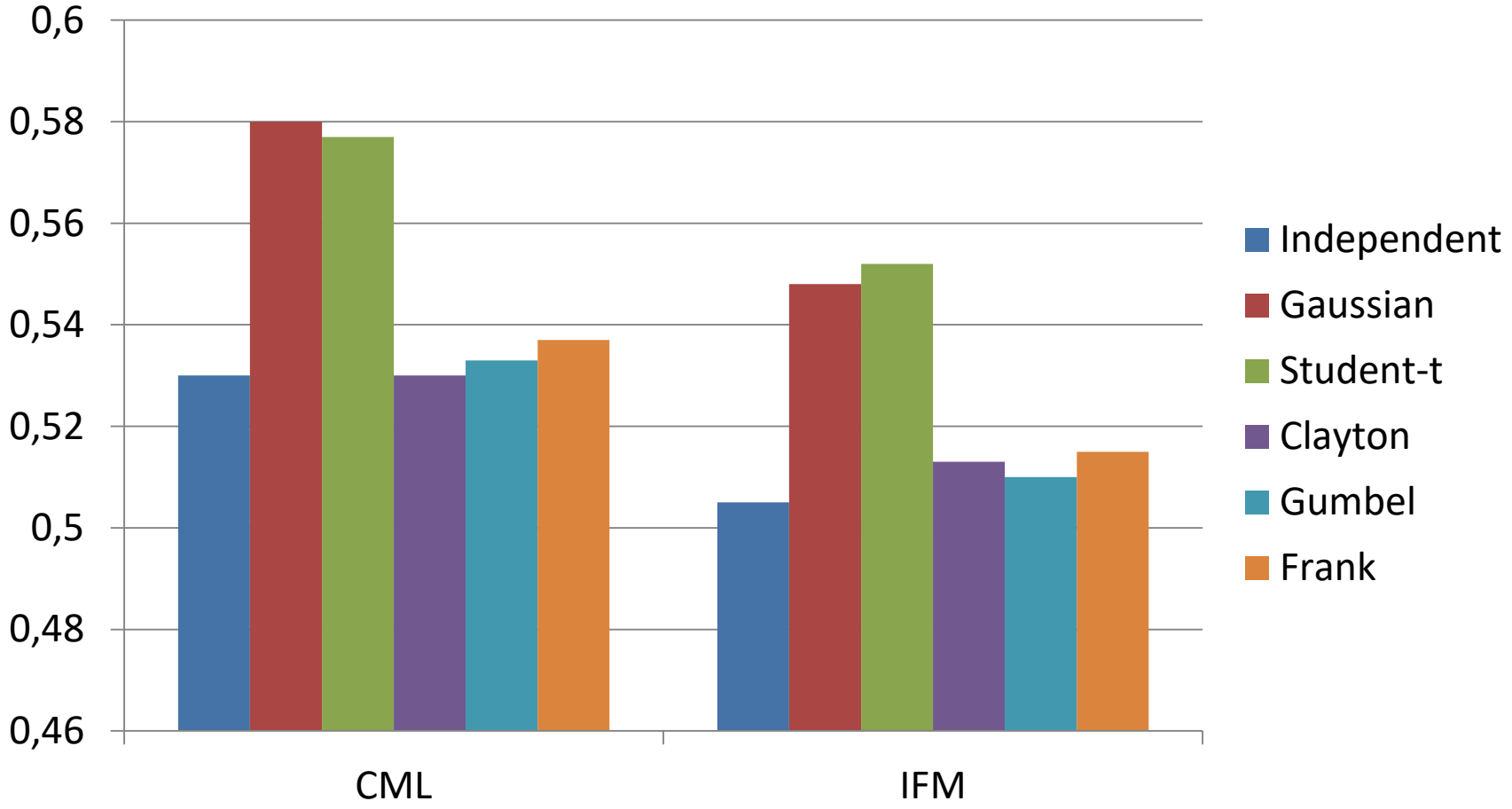
Результати класифікації вибірки «іриси»



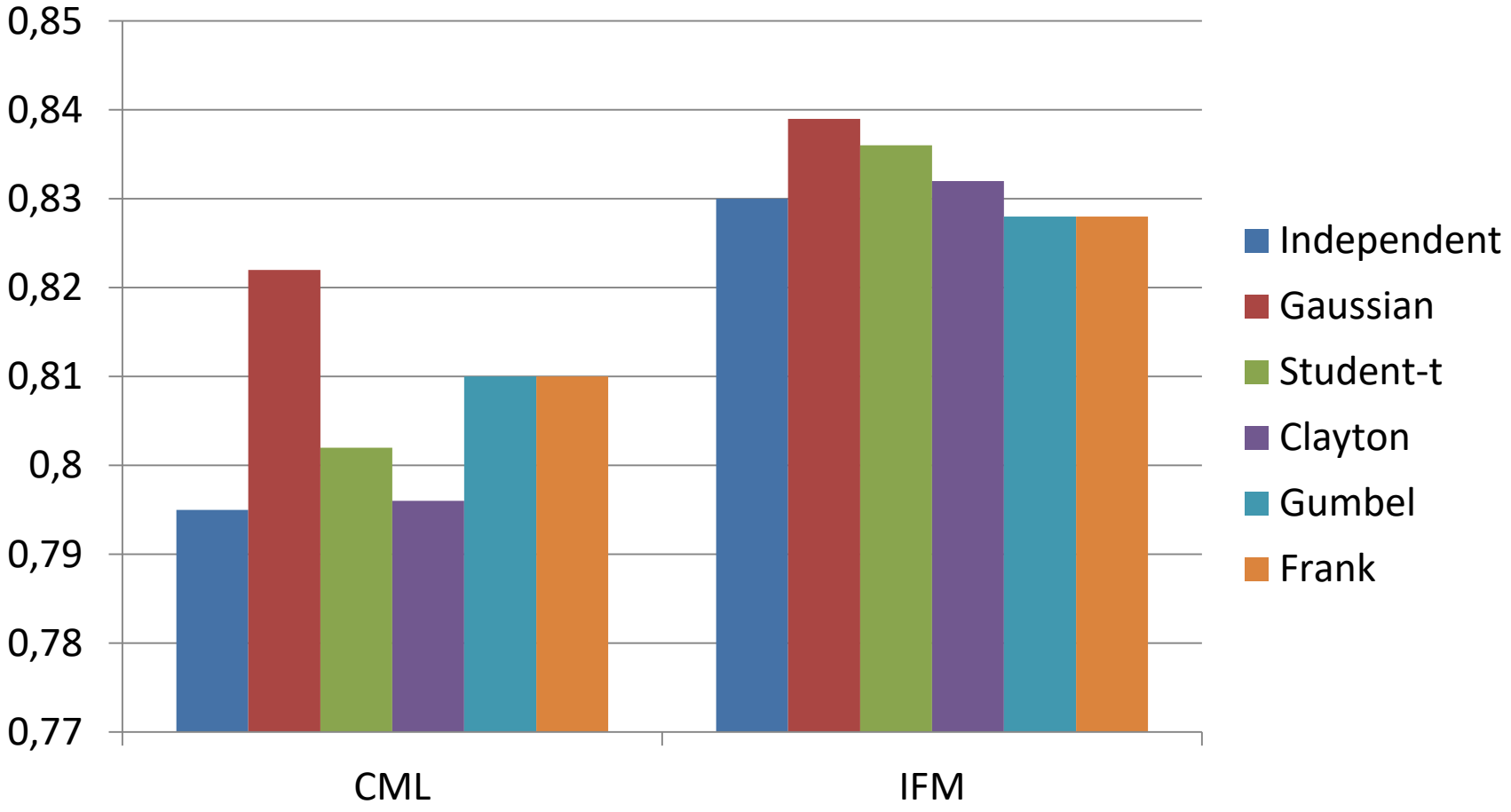
Результати класифікації вибірки «насіння»



Результати класифікації вибірки «вино»



Результати класифікації вибірки «Vertebral»



Висновки

- гауссівські копули і копули Стьюдента найточніше передають залежність між даними для більшості ознак і класів. Архімедові копули лише іноді можуть дати результат того ж порядку.
- метод вибору маргінальних розподілів загалом є більш ефективним за класичний метод максимальної правдоподібності. Тим не менш, КММП демонструє кращі результати у випадках, коли маргінальні розподіли не можуть бути встановлені з достатньою точністю.
- значення інформаційного критерію Акайке і критерію SnC в досліджуваних випадках корелюють між собою і, в основному, дають схожі результати.