Роботу виконала студентка

Національного університету «Одеська політехніка»

"Прикладної математики"

Матиченко А.Д.

Оцінка

ЗАВДАННЯ 1

Порівняти кластерізації 1 та 2 за допомогою точності, повноти та F-міри.

КЛАСТЕРІЗАЦІЯ 1

Полученные кластеры	T_1	T_2	T_3	$\|C_i\|$
C_1	45	10	3	58
C_2	0	50	0	50
C_3	12	0	30	42
$\|T_i\|$	57	60	33	$n=150$

КЛАСТЕРІЗАЦІЯ 2

Полученные кластеры	T_1	T_2	T_3	$\|C_i\|$
C_1	40	10	0	50
C_2	5	44	0	49
C_3	12	6	0	18
$\|T_i\|$	57	60	0	$n=117$

Використовуючи F -міру визначити у якої кластеризації якість вища.

Розв'язання

ЗНАЧЕННЯ ДЛЯ ПЕРШЕГО КЛАСТЕРА

	Точність	Повнота	Гармонічне середнє
i	$P_r(C_i)=\frac{1}{\lvert C_i\rvert}\max\limits_{j=1,k}\{n_{ij}\}$	$R_c(C_i)=\frac{1}{\lvert T_{j\max}\rvert}\max\limits_{j=1,k}\{n_{ij}\}$	$F(C_i)=\frac{2P_r(C_i)R_c(C_i)}{P_r(C_i)+R_c(C_i)}$
1	$P_r(C_1)=\frac{1}{\lvert C_1\rvert}=\frac{45}{58}=0.78$	$Rc(C_1)=\frac{1}{\lvert T_{1\max}\rvert}=\frac{45}{57}=0.79$	$F(C_1)=\frac{2\cdot 0.78\cdot 0.79}{0.78+0.79}=0.78$
2	$P_r(C_2)=\frac{1}{\lvert C_2\rvert}=\frac{50}{50}=1$	$Rc(C_2)=\frac{1}{\lvert T_{2\max}\rvert}=\frac{50}{60}=0.84$	$F(C_2)=\frac{2\cdot 1\cdot 0.84}{1+0.84}=0.91$
3	$P_r(C_3)=\frac{1}{\lvert C_3\rvert}=\frac{30}{42}=0.71$	$Rc(C_3)=\frac{1}{\lvert T_{3\max}\rvert}=\frac{30}{33}=0.9$	$F(C_3)=\frac{2\cdot 0.71\cdot 0.9}{0.71+0.9}=0.79$
$F=\frac{1}{r}\sum\limits_{i=1}^{n}F(C_i)=\frac{1}{3}(0.78+0.91+0.79)=0.83$			

ЗНАЧЕННЯ ДЛЯ ДРУГОГО КЛАСТЕРА

	Точність	Повнота	Гармонічне середнє
i	$P_r(C_i)=\frac{1}{\lvert C_i\rvert}\max\limits_{j=1,k}\{n_{ij}\}$	$R_c(C_i)=\frac{1}{\lvert T_{j\max}\rvert}\max\limits_{j=1,k}\{n_{ij}\}$	$F(C_i)=\frac{2P_r(C_i)R_c(C_i)}{P_r(C_i)+R_c(C_i)}$
1	$P_r(C_1)=\frac{1}{\lvert C_1\rvert}=\frac{40}{50}=0.8$	$Rc(C_1)=\frac{1}{\lvert T_{1\max}\rvert}=\frac{40}{57}=0.7$	$F(C_1)=\frac{2\cdot 0.7\cdot 0.8}{0.8+0.7}=0.75$
2	$P_r(C_2)=\frac{1}{\lvert C_2\rvert}=\frac{44}{49}=0.89$	$Rc(C_2)=\frac{1}{\lvert T_{2\max}\rvert}=\frac{44}{60}=0.74$	$F(C_2)=\frac{2\cdot 0.89\cdot 0.74}{0.89+0.74}=0.8$
3	$P_r(C_3)=\frac{1}{\lvert C_3\rvert}=\frac{12}{18}=0.67$	$Rc(C_3)=\frac{1}{\lvert T_{3\max}\rvert}=0$	$F(C_3)=\frac{2\cdot 0.67\cdot 0}{0.67+0}=0$
$F=\frac{1}{r}\sum\limits_{i=1}^{n}F(C_i)=\frac{1}{3}(0.75+0.8+0)=0.3125$			

Відповідь: Перша кластеризація за якістю краща, оскільки значення міри для неї вище.

ЗАВДАННЯ 2

Порівняти кластерізації за допомогою інформаційних показників.

КЛАСТЕРІЗАЦІЯ 1

Полученные кластеры	T_1	T_2	T_3	$\lvert C_i \rvert$
C_1	45	10	3	58
C_2	0	50	0	50
C_3	12	0	30	42
$\lvert T_i \rvert$	57	60	33	$n=150$

КЛАСТЕРІЗАЦІЯ 2

Полученные кластеры	T_1	T_2	T_3	$\lvert C_i \rvert$
C_1	40	10	0	50
C_2	5	44	0	49
C_3	12	6	0	18
$\lvert T_i \rvert$	57	60	0	$n=117$

Вираховуємо умовну ентропію, взаємну інформацію та нормалізовану взаємну інформацію. І потім порівнюємо умовну ентропію та взаємну нормалізовану інформацію.

Розв’язок

1. Визначимо інформаційні показники кластеризації один. Спочатку обчислимо умовну ентропію:

$$H(T \mid C) = -\sum_{i=1}^{r}\sum_{j=1}^{k} p_{ij} \log_2 \frac{p_{ij}}{p_{c_i}},\ p_{ij} = \frac{n_{ij}}{n}$$ за допомогою наступного коду:

```
temp=[]
result=0
for i in range(3):
    for j in range(3):
        if arr[i,j]==0:
            continue
        else:
```

```
            result+=arr[i,j]/arr[i,3]*np.log2((arr[i,j]/arr[i,3]))
    temp.append(result)
    result=0
res=sum([arr[i,3]/arr[3,3]*temp[i] for i in range(3)])
print(res)
```

$$H(T|C)=-\sum_{i=1}^{r}\sum_{j=1}^{k}p_{ij}\log_2\frac{p_{ij}}{p_{c_i}}=0.6$$

2. Взаємна інформація для кластеризації 1:

$H(T)=-\sum_{j=1}^{k}P_{T_j}\log_2 P_{T_j}=1.54$, $I(C,T)=H(T)-H(T|C)=0.6-0.41=0.93$ за допомогою наступного коду знайдемо:

```
result=0
i=0
for j in range(3):
    result+=arr[3,j]/arr[3,3]*np.log2((arr[3,j]/arr[3,3]))
    print(arr[3,j]/arr[3,3]*np.log2((arr[3,j]/arr[3,3])))
    i=i+1
print(result-res)
```

3. Нормалізована взаємна інформація $NMI(C,T)=\frac{I(C,T)}{\sqrt{H(C)\cdot H(T)}}=0.6$,

$H(C)=H(C_1,...,C_k)=-\sum_{I=1}^{K}P_{C_I}\log_2 P_{C_I}=1.57$ за допомогою наступного коду знайдемо:

```
C=0
for i in range(3):
    C+=arr[i,3]/arr[3,3]*np.log2(arr[i,3]/arr[3,3])
C=abs(C)
NMI=I/np.sqrt(abs(result)*C)
```

Аналогічно проробимо все це ж для другого кластеру. Запишемо результати у таблицю.

	Перший кластер	Другий
$H(T\|C)$	0.6	0.71
$H(T)$	1.54	0
$I(C,T)$	0.93	0
$H(C)$	1.57	1.46
$NMI(C,T)$	0.6	0

Порівняємо отримані показники.

Кластеризація	$H(T\|C)$	$NMI(C,T)$
1	0.6	0.6
2	0.71	0

Кластеризація 1 краще ніж 2, так як у неї вища нормалізація і менша умовна ентропія.

Завдання 3

Порівняти кластерізації 1 та 2 із попереднього завдання за допомогою попарних показників.

КЛАСТЕРІЗАЦІЯ 1

Полученные кластеры	T_1	T_2	T_3	$\lvert C_i \rvert$
C_1	45	10	3	58
C_2	0	50	0	50
C_3	12	0	30	42
$\lvert T_i \rvert$	57	60	33	$n=150$

КЛАСТЕРІЗАЦІЯ 2

Полученные кластеры	T_1	T_2	T_3	$\lvert C_i \rvert$
C_1	40	10	0	50
C_2	5	44	0	49
C_3	12	6	0	18
$\lvert T_i \rvert$	57	60	0	$n=117$

Розв’язання

ДЛЯ ПЕРШОГО КЛАСТЕРУ

$$TP=\sum_{i=1}^{r}\sum_{j=1}^{k}\frac{n_{ij}\left(n_{ij}-1\right)}{2}=\frac{45\cdot 44}{2}+\frac{10\cdot 9}{2}+\frac{3\cdot 2}{2}+\frac{50\cdot 49}{2}+\frac{12\cdot 11}{2}+\frac{30\cdot 29}{2}=2764$$

$$FN=\sum_{j=1}^{k}\frac{m_j\left(m_j-1\right)}{2}-TP=\frac{57\left(57-1\right)}{2}+\frac{60\left(60-1\right)}{2}+\frac{33\left(33-1\right)}{2}-2764=1110$$

$$FP=\sum_{i=1}^{r}\frac{n_i\left(n_i-1\right)}{2}-TP=\frac{58\left(58-1\right)}{2}+\frac{50\left(50-1\right)}{2}+\frac{42\left(42-1\right)}{2}-TP=975$$

$$TN=N-TP-FP-FN=\frac{n\left(n-1\right)}{2}-TP-FP-FN=\frac{150\left(150-1\right)}{2}-2764-1110-975=6326$$

$$Jaccard = \frac{TP}{TP + FP + FN} = \frac{2764}{2764 + 975 + 1110} = \frac{2764}{4849} = 0.57$$

$$FM = \frac{TP}{\sqrt{(TP + FN)(TP + FP)}} = \frac{2764}{\sqrt{(2764 + 1110)(2764 + 975)}} = 0.73$$

ДЛЯ ДРУГОГО КЛАСТЕРУ

$$TP = \sum_{i=1}^{r} \sum_{j=1}^{k} \frac{n_{ij}(n_{ij} - 1)}{2} = \frac{40 \cdot 39}{2} + \frac{10 \cdot 9}{2} + \frac{5 \cdot 4}{2} + \frac{44 \cdot 43}{2} + \frac{12 \cdot 11}{2} + \frac{6 \cdot 5}{2} = 1796$$

$$FN = \sum_{j=1}^{k} \frac{m_j(m_j - 1)}{2} - TP = \frac{57(57 - 1)}{2} + \frac{60(60 - 1)}{2} - 1796 = 1570$$

$$FP = \sum_{i=1}^{r} \frac{n_i(n_i - 1)}{2} - TP = \frac{50(50 - 1)}{2} + \frac{49(49 - 1)}{2} + \frac{18(18 - 1)}{2} - TP = 2338 - 1796 = 542$$

$$TN = N - TP - FP - FN = \frac{n(n-1)}{2} - TP - FP - FN = \frac{117(117 - 1)}{2} - 2764 - 1110 - 975 = 4701$$

$$Jaccard = \frac{TP}{TP + FP + FN} = \frac{1796}{1796 + 1570 + 542} = 0.46$$

$$FM = \frac{TP}{\sqrt{(TP + FN)(TP + FP)}} = \frac{1796}{\sqrt{(1796 + 1570)(1796 + 542)}} = 0.65$$

Відповідь: характеристичні значення для першого кластеру більші, тому має більш високу якість кластерізації.