Projet de conception et de prog...

Mickael LE DENMAT

Université Versailles Saint-Quentin en Yvelines
Investigating feature selection techniques to improve data mining tasks

27 janvier 2023

Table des matières

UVSQ université PARIS-SACLAY UFR des Sciences CAMPUS DE VERSAILLES

Data Mining

- Le monde d'aujourd'hui : Beaucoup de données !

UVSQ université PARIS-SACLAY UFR des Sciences CAMPUS DE VERSAILLES

Data Mining

- Le monde d'aujourd'hui : Beaucoup de données !
- Problème de prise de décision.

UVSQ université PARIS-SACLAY UFR des Sciences CAMPUS DE VERSAILLES

Data Mining

- Le monde d'aujourd'hui : Beaucoup de données !
- Problème de prise de décision.

Solution

Data Mining : la pratique consistant à rechercher automatiquement de grandes quantités de données afin de découvrir des tendances et des modèles qui vont au delà de la simple analyse. [Oracle,]

UVSQ université PARIS-SACLAY UFR des Sciences CAMPUS DE VERSAILLES

Feature selection

- Complexité du monde réel.

Feature selection

- Complexité du monde réel.
- Accélérer la prise de décision.

Feature selection

- Complexité du monde réel.
- Accélérer la prise de décision.

Solution

Feature selection => Rough Set Theory.

UVSQ université PARIS-SACLAY UFR des Sciences CAMPUS DE VERSAILLES

Rough Set Theory

UVSQ université PARIS-SACLAY UFR des Sciences CAMPUS DE VERSAILLES

Un système d'information

Patient	Headache	Muscle-Pain	Temperature
o_1	Yes	Yes	Very High
o_2	Yes	No	High
o_3	Yes	No	High
o_4	No	Yes	Normal
o_5	No	Yes	High
o_6	No	Yes	Very High

Table – Exemple d'un système d'information

Un système de décision

Patient	Headache	Muscle-Pain	Temperature	Flu
o_1	Yes	Yes	Very High	Yes
o_2	Yes	No	High	Yes
o_3	Yes	No	High	No
o_4	No	Yes	Normal	No
o_5	No	Yes	High	Yes
o_6	No	Yes	Very High	Yes

Table – Exemple d'un système d'information

Indiscernibility Relation

Définition

Soit I = (U, A), avec I un système d'information, U un ensemble d'objets et A un ensemble d'attributs. Avec n'importe quel sous ensemble $P \subseteq A$. Il existe une relation d'équivalence, noté IND(P), définit comme :

$$IND(P) = \{(x, y) \in U^2 | \forall a \in P, a(x) = a(y)\} \quad (1)$$

UVSQ université PARIS-SACLAY UFR des Sciences CAMPUS DE VERSAILLES

Indiscernibility Relation

Exemple

Patient	Headache	Muscle-Pain	Temperature	Flu
o_1	Yes	Yes	Very High	Yes
o_2	Yes	No	High	Yes
o_3	Yes	No	High	No
o_4	No	Yes	Normal	No
o_5	No	Yes	High	Yes
o_6	No	Yes	Very High	Yes

Rough Set

Soit un ensemble X d'objets cibles tel que $X \subseteq U$.

- Représenter X avec P (un sous ensemble d'attributs).

UVSQ université PARIS-SACLAY UFR des Sciences CAMPUS DE VERSAILLES

Rough Set

Soit un ensemble X d'objets cibles tel que $X \subseteq U$.

- Représenter X avec P (un sous ensemble d'attributs).
- X n'a qu'une seule classe => classe d'équivalence de P.

UVSQ université PARIS-SACLAY UFR des Sciences CAMPUS DE VERSAILLES

Rough Set

Soit un ensemble X d'objets cibles tel que $X \subseteq U$.

- Représenter X avec P (un sous ensemble d'attributs).
- X n'a qu'une seule classe => classe d'équivalence de P.

Problème

On ne peut pas calculer précisément X.
On doit l'approximer.

UVSQ université PARIS-SACLAY UFR des Sciences CAMPUS DE VERSAILLES

Approximations

B-Lower Approximation

$$\underline{B}(X) = \{o_j | [o_j]_B \subseteq X\} \tag{2}$$

C'est l'ensemble complet des objets dans U/P qui peuvent être classés dans X sans ambiguïté.

Approximations

B-Lower Approximation

$$\underline{B}(X) = \{o_j | [o_j]_B \subseteq X\} \tag{3}$$

C'est l'ensemble complet des objets dans U/P qui peuvent être classés dans X sans ambiguïté.

B-Upper Approximation

$$\bar{B}(X) = \{o_j | [o_j]_B \cap X \neq \emptyset\} \tag{4}$$

C'est l'ensemble des objets dans U/P qui peuvent être classés dans X. Ce sont des objets possèdent un objets indiscernables.

Approximation (suite)

B-boundary Approximation

$$BN_B(X) = \bar{B}(X) - \underline{B}(X) \tag{5}$$

Positive & Negative Region

Définition

Positive Region

$$POS_C\{(d)\} = \bigcup_{X \in U/\{(d)\}} \bar{C}(X) \tag{6}$$

Positive & Negative Region

Définition

Positive Region

$$POS_C\{(d)\} = \bigcup_{X \in U/\{(d)\}} \bar{C}(X) \tag{7}$$

Negative Region

$$NEG_C\{(d)\} = U - \bigcup_{X \in U/\{(d)\}} \underline{C}(X) \tag{8}$$

Reduct & Core

Définition

Un *reduct* est un sous ensemble minimal d'attributs ayant la même *Positive Region* que l'ensemble des attributs.

UVSQ université PARIS-SACLAY UFR des Sciences CAMPUS DE VERSAILLES

Reduct & Core

Définition

Un *reduct* est un sous ensemble minimal d'attributs ayant la même *Positive Region* que l'ensemble des attributs.

Définition

Un *core* est un ensemble d'attributs indépendant incluant tous les *reduct*.

UVSQ université PARIS-SACLAY UFR des Sciences CAMPUS DE VERSAILLES

Références

Oracle.
Qu'est-ce que le data mining ?

Merci pour votre attention !
Des questions ?

UVSQ université PARIS-SACLAY UFR des Sciences CAMPUS DE VERSAILLES