MATHÉMATIQUES ET SCIENCES HUMAINES

SIMON REGNIER

Stabilité d'un opérateur de classification

Mathématiques et sciences humaines, tome 60 (1977), p. 21-30
<http://www.numdam.org/item?id=MSH_1977__60__21_0>

Math. Sci. hum. (*15*^{e} année, n°60, *1977*, p.21-30)

STABILITE D'UN OPERATEUR DE CLASSIFICATION

Simon REGNIER*

1. INTRODUCTION

Les algorithmes de classification concurrents sont aujourd'hui nombreux, mais on n'a presque jamais de raison *théorique* de choisir l'un plutôt que l'autre.

Il est démontré par exemple** que la méthode du maximum de vraisemblance appliquée au modèle statistique le plus général est inféconde : la vraisemblance est seulement fonction des effectifs des classes formées !

Privés de cette voie royale, nous proposons ici d'imposer *a priori* que l'algorithme, ou plutôt l'opérateur de classification possède certaine propriété souhaitable, dans l'esprit du théorème de Arrow. La propriété demandée est d'abord :
que la suppression (ou l'adjonction) d'un *objet arbitraire* à l'ensemble E que l'on veut classer ne perturbe PAS DU TOUT le résultat. On verra que c'est trop demander ! Deuxième impasse.

Les voies de recherche qui restent ouvertes (dans le même esprit) sont alors deux formes de stabilité approximative :
a) l'exigence de stabilité statistique, dans un modèle probabiliste.
b) l'exigence de stabilité à ε-près, quand l'ensemble des partitions de E est muni d'une métrique appropriée.

---

* Maison des Sciences de l'Homme, Service de Mathématiques Appliquées et de Calcul

** S. REGNIER, "Non fécondité du modèle statistique général de la classification automatique", Acte du Colloque international "Archéologie et Calculateurs", Marseille 7-12 Avril 1969, Editions du CNRS (1970).

2. OPERATEUR ET ALGORITHME

Nous appellerons opérateur de classification une fonction qui associe à tout ensemble fini E, muni d'une structure S d'un type fixé, une classe de partitions optimales de E.

$$(E,S) \longrightarrow T(E,S) = C \subseteq P(E)$$

où E = ensemble fini "d'objets"

S = structure de E

$P(E)$ = ensemble des partitions de E.

Si C est réduit à une partition X, T est dit univoque.
La plupart des algorithmes de classification connus définissent autant d'opérateurs. Mais la notion d'opérateur est plus large. Par exemple C peut être définie comme la classe des partitions Y qui maximisent une certaine fonction H(X,S), définie pour tout X de P(E), et dépendant de S. H définit l'opérateur T, sans que l'on connaisse nécessairement un algorithme de calcul de C.
En particulier si H est linéaire

$$H(X) = \sum_{\{i,j\} \subset E} X_{ij}\,(s_{ij} - 1/2)$$

où $X_{ij} \begin{cases} = 1 \text{ pour i et j réunis} \\ = 0 \text{ pour i et j séparés} \end{cases}$ par $[X \in R^{n(n-1)/2}]$

C est une face d'un polyèdre P_n, fermeture convexe de P(E) dans $R^{n(n-1)/2}$.
Calculer C relève de la programmation linéaire, mais le polyèdre étant connu par ses sommets, et la proportion d'arêtes parmi les paires de sommets étant forte, on ne connaît pas d'algorithme de calcul de C pour n grand, sinon par génération de toutes les partitions.

3. STABILITE STRICTE

L'ensemble d'objets E qu'il faut classer présente souvent un caractère contingent. Ce peut-être la partie observée d'une population plus vaste.
Il serait souhaitable, dans cette perspective, que l'opérateur T soit stable par rapport à E, c'est-à-dire que la suppression d'une partie de E conduise à des résultats concordants. Pour formaliser cette exigence, nous demanderons que :
Si F est une partie de E, où S définit une structure restreinte de même nature S_F, permettant de former : $T(F,S_F) = C_F$, classe de partitions Y optimales de F, chaque Y soit la restriction à F, $X_{/F}$, d'une partition optimale de E, $X \in C$; et plus brièvement que $C_F = C_{/F}$.

Les algorithmes de classification connus prennent en charge sur E des structures initiales extrêmement variées, mais les condensent en général à une structure numérique S qui est une matrice symétrique de nombres s_{ij}, associés à chaque paire d'objets i et j. S peut être une métrique ou une similarité. T peut ne tenir compte que de l'ordre relatif des nombres s_{ij}, appelé ordonnance. Quand S est une telle matrice, toutes les parties F de E possèdent une structure S_F homologue, et l'exigence de stabilité n'est remplie par aucun opérateur T satisfaisant. Nous allons établir ce résultat négatif dans le cadre d'une exigence moins restrictive : l'opérateur T est dit "une fois stable" si T opère seulement sur des ensembles E d'effectif n et n-1, et que T est stable pour les parties F à n-1 éléments, quand $|E| = n$.

4. OPERATEURS UNIVOQUES

Nous considérons d'abord un opérateur T univoque, $T(E,S) = X \in P(E)$.

La partition X est identifiée à la matrice $X_{ij} = \begin{matrix} 1 \text{ pour i et j réunis} \\ 0 \text{ pour i et j séparés} \end{matrix}$.

Ainsi sont définis n(n-1)/2 opérateurs booléens, par projection :

$$T_{ij}(S) = X_{ij} = 0 \text{ ou } 1.$$

LEMME : La fonction $T_{ij}(S)$ dépend seulement de s_{ij}.

<u>Preuve</u> : Soit k un objet différent de i et j. Prenons $F = E - \{k\}$, de n-1 objets. S_F est la sous-matrice déduite de S par suppression de la ligne k et la colonne k.

L'hypothèse de stabilité impose $\quad T_{ij}(E,S) = T_{ij}(F,S_F)$.

Le second membre ne peut dépendre d'aucun terme s_{ik}, s_{jk} ou s_{kl}. Le premier membre non plus.

Puisque k est quelconque dans $E - \{i,j\}$, $T_{ij}(E,S) = g_{ij}(s_{ij})$, une fonction d'un seul argument numérique s_{ij}.

Mais la fonction g_{ij} peut-elle dépendre elle-même de la paire d'objets ij ? Si oui, cela définit une structure supplémentaire sur E, indépendante de S, alors que cette matrice est censée prendre en compte toutes les relations entre objets qui conditionnent la classification X cherchée. g_{ij} variable sera donc considérée comme un cas pathologique que nous analyserons en second.

Dans le cas normal où $T_{ij}(S) = g(s_{ij})$, on voit facilement que l'application g de $\mathbb{R}$ dans $\{0,1\}$ doit être constante, car s'il existait a et b tels que $g(a) = 0$ et $g(b) = 1$ en prenant $s_{ij} = a$, $s_{ik} = s_{jk} = b$, la relation $X = T(S)$ ne serait pas transitive.

Ainsi, quand S est une matrice, les seuls opérateurs une fois stables et univoques sont

$$T(E,S) \equiv 0 \qquad \text{et } T(E,S) \equiv 1 \ .$$

Soient les deux opérateurs constants qui transforment chaque matrice S en une borne du treillis des partitions de E.

Dans le cas pathologique où g_{ij} dépend de ij, les paires ij où g n'est pas constant forment un graphe non orienté qui ne peut comporter de triangle pour les raisons ci-dessus, mais pas non plus de sommet de degré > 1, tel que

$i \begin{smallmatrix} \nearrow j \\ \searrow k \end{smallmatrix}$ car X_{jk} = constante = 0 ou 1 rendrait possible une non transitivité.

Ce graphe n'a donc que des arêtes isolées et définit une partition M en classes d'effectifs 1 et 2. Les paires où $X_{ij} = 1$ définissent une partition Z et M ne réunit que des paires de singletons de Z ($M \cap Z=0$).

Nous voyons que X = T(S) est une partition "peu variable" X = Z + Y où Y varie entre 0 et M, (pour l'addition des indicateurs X_{ij}).

5. OPERATEURS MULTIVOQUES

Si l'opérateur T est multivoque, $T(E,S) = C \subseteq P(E)$, l'intersection des partitions $X \in C$ définit un autre opérateur $V(E,S) = \bigcap_{C} X = Y$ qui est univoque et lui aussi une fois stable, car si $F = E - \{i\}$ et que $T(F,S_F) = C/_F$ l'intersection des restrictions $Y = X/_F$ est égale à la restriction de V, définie localement par

$$V_{ij} = \inf_{X \in C} (X_{ij}) .$$

De même $U_{ij} = \sup_{X \in C} (X_{ij})$ définit une relation une fois stable, mais non transitive. La fermeture transitive $\overline{U} = \bigcup_{X \in C} X$ n'est pas définie localement et l'on a seulement une inégalité $U_F \leqslant U/F$.

La stabilité $V_F = V/F$ entraîne que $V(E,S) = 0$ ou 1 (nous excluons le cas pathologique).

Si $V(E,S) = 1$, T est univoque et $T(E,S) = 1$.

Si $V(E,S) = 0$, la classe $C = T(E,S)$ sépare toutes les paires en ce sens que : pour toute paire ij il existe un X pris dans C tel que $X_{ij} = 0$. Il est clair dans ces conditions que l'ensemble C de partitions "optimales" est trop dispersé pour donner des informations pertinentes sur l'ensemble E.

6. STABILITE APPROCHEE ET STABILITE STATISTIQUE

A défaut de stabilité stricte, on peut demander à l'opérateur T une stabilité approchée, car on peut de plusieurs façons définir une distance d entre partitions sur E et $F \subseteq E$. Alors quand :

$$T(E,S) = X \quad \text{est univoque}$$
$$T(F,S_F) = Y$$

il faudrait que la distance entre Y et X_F (restriction de X à F) soit petite quand $|E-F|$ est petit.

Si F est une partie aléatoire de E, par exemple F est un ensemble d'objets observés tirés au hasard d'une population E d'objets possibles, on demandera de même que la distance aléatoire entre Y et $X/_F$ soit petite en probabilité quand l'écart de E à F est petit, aléatoire ou non. Pour T multivoque, on considérera de même une distance entre $T(F,S_F) = C_F$ et $C/_F$.

Une convergence $Y \to X$ ou $C_F \to C$ quand $F \to E$ ne peut être demandée dans le contexte actuel. $F \to E$ n'a pas de sens dans les ensembles E finis, mais certains algorithmes s'appliquent aussi à des échantillons dans la population E, ou famille E_p de p éléments tirés de E. Et nous allons présenter une classe particulière des opérateurs doués d'une convergence associée à cette situation.

7. LES H-OPERATEURS, LEUR STABILITE STATISTIQUE

$$T : (E,S) \longrightarrow C_E \subset P(E)$$

est défini par des représentations a de chaque ensemble E dans un ensemble fixe F appelé "ensemble d'images" F, fini. S est induite par la structure suivante sur F : H(x,L) est une fonction numérique définie pour toute partition x de P(F) et toute loi de probabilité L sur F.

Le centre C_E est l'image réciproque par a d'une classe de partitions optimales de F :

$$C : \{x \in P(F) \mid H(x,L) \leqslant H(y,L) \text{ pour tout } y\}$$

soit alors E_p un échantillon de p objets dans E : $e_1, e_2, \ldots e_p$.
L'échantillon des images $a(e_k)$ définit sur F une loi de probabilité empirique $1_p = \frac{1}{p} \sum_1^p \delta(f_k)$ (moyenne des mesures de Dirac) d'où résulte un autre centre

$$C_p = \{y \in P(F) \mid H(y,1_p) = \text{minimum}\}.$$

Si les objets aléatoires e_p sont indépendants et de même loi et que chaque $a(e_p)$ suit dans F la loi L(*), $1_p(f)$, proportion aléatoire, $\longrightarrow L(f)$ quand $p \longrightarrow \infty$ convergence presque sûre, pour toute image $f \in F$. Alors si H(x,L) est une fonction continue des nombres L(f), $H(x,1_p) \longrightarrow H(x,L)$ presque sûrement, pour tout x et nous démontrons alors que

$$\text{Probabilité } [C_p \subseteq C] \longrightarrow 1$$
$$\text{Probabilité } [\limsup C_p \subset C] = 1 .$$

Ces inclusions dans P(F) traduisent des inclusions homologues dans P(E).

Nous allons établir ces propriétés dans un cadre plus général qui les met mieux en lumière. La première propriété s'applique seule dans les situations concrètes où l'on connaît un seul centre C_p, à partir d'un échantillon E_p.

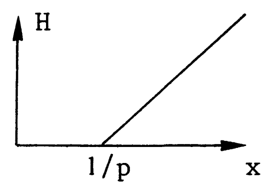
(*) L sera habituellement l'image par a de la loi uniforme sur E.

8. CENTRES D'UNE SUITE CONVERGENTE DE FONCTIONS

La définition de C à partir de H suggère la généralisation suivante :
J'appelle centre C de toute fonction numérique $H : X \to \mathbb{R}$ l'ensemble C, éventuellement vide, où H atteint sa borne inférieure.

$$C = \{x \in X \mid H(x) \leqslant H(y),\ \forall\, y \in X\}\ .$$

Si une suite de fonctions certaines H_p, tend vers H en chaque point de X, on voit facilement que les centres vérifient : $\limsup C_p \subset C$ et si X est fini $C_p \subset C$ pour p assez grand.
En effet, si $H(x) > H_p(y)$, pour p assez grand, j'ai $H_p(x) > H_p(y)$.
Donc la suite des C_p qui comprennent x ne peut être infinie : $x \notin \limsup C_p$.
Si X est fini, la suite des C_p qui contiennent au moins un x de X-C est également finie, parce que l'union d'une famille finie de parties finies est finie.
Ces deux résultats sont les seules propriétés générales dans le domaine des fonctions certaines, comme le montrent les contre-exemples suivants :



1) X = les rationnels compris entre 0 et 1
$H_p(x) = \sup\ (0,\ x-1/_p)$
Alors $C_p = [0,\frac{1}{p}] \supset C = \{0\}$, inclusion stricte pour tout p.

2) $H_p(x) = \frac{x}{p}$ $\qquad C = \underline{X} = [0,1]$ $\qquad$ et $\quad C_p = \{0\}$.
On voit aussi facilement que la suite C_p n'est pas en général convergente, et peut même être quelconque : (C_p) étant donnée on notera B_p les fonctions indicatrices et on posera : $H_p(x) = -\frac{1}{p}\ B_p(x)$, par exemple.

<u>Fonctions aléatoires à limite certaine</u>

Si les fonctions H_p sont aléatoires, la limite H restant une fonction certaine, on a d'abord un résultat indépendant du cardinal de $\underline{X}$:

LEMME

Si $x \in \underline{X} - C$

1) $H_p \to H$ en probabilité pour tout x $\Rightarrow$ $\Pr\,[x \in C_p] \longrightarrow 0$, si l'événement $x \in C_p$ est mesurable.

2) $H_p \longrightarrow H$ presque sûrement pour tout x $\Rightarrow$ $\Pr\,[\limsup C_p \ni x] = 0$.

En effet $\quad x \notin C$ entraîne $\exists\, y,a : H(y) = H(x) - 2a$ et $a > 0$;
alors $H_p(y) - H_p(x) = H_p(y) - H(y) + H(x) - H_p(x) - 2a$

(1) et $\quad \sup\limits_{p>n}\ [H_p(y) - H_p(x)] \ \leqslant\ \sup\ |H_p(y) - H(y)| + \sup\ |H_p(x) - H(x)| - 2a$

par suite

1) $x \in C_p \Rightarrow H_p(y) - H_p(x) \geqslant 0 \quad \forall y$, et en particulier :

$\Rightarrow$ ou bien : $H_p(y) - H(y) \geqslant a$, $H_p(x) - H(x) \geqslant a$,

ou bien : Les probabilités de ces deux événements : $1_p(x)$ et $1_p(y)$ tendent vers 0 lorsque $H_p(x) \longrightarrow H(x)$ en probabilité pour tout x, par suite : $\Pr(x \in C_p) \leqslant 1_p(x) + 1_p(y) \longrightarrow 0$.

2) $x \in \limsup C_p$ signifie : la suite des C_p qui contiennent x est infinie, soit $\forall n \quad \exists\, p > n \quad \forall y \quad H_p(y) \geqslant H_p(x)$

et entraîne $\forall y \quad \forall n \quad \exists\, p > n \quad H_p(y) \geqslant H_p(x)$

soit $\forall y \quad \limsup\, [H_p(y) - H_p(x)] \geqslant 0$.

Maintenant, les événements mesurables $E_n(y)$: $\sup\limits_{p>n} [H_p(y) - H_p(x)] \geqslant 0$

forment une suite décroissante : $E_{n+1} \Rightarrow E_n$ si bien que

$\Pr\, [\lim E_n] = \lim P\,(E_n)$, pour tout y.

Dans le cas particulier où $H(y) = H(x) - 2a$, il résulte de (1) que :

$$\Pr\left[\sup_{p>n}\, [H_p(y) - H_p(x)] \geqslant 0\right] \leqslant \Pr\left\{\sup_{p>n} |H_p(y) - H(y)| \geqslant a\right\} +$$

$$\Pr\left\{\sup_{p>n} |H_p(x) - H(x)| \geqslant a\right\}$$ et ces deux probabilités tendent vers 0 si $H_p \longrightarrow H$ presque sûrement.

Par conséquent $\Pr\,\{\limsup\, [H_p(y) - H_p(x)] \geqslant 0\} = 0$ pour le point y choisi, et à fortiori $\Pr\,[x \in \limsup C\, p] = 0$.

De ce lemme résulte immédiatement :

1) Si $H_p \longrightarrow H$ en probabilité et $\underline{X}$ fini $\Pr\,[C_p \subset C] \longrightarrow 1$.

2) Si $H_p \longrightarrow H$ presque sûrement et $\underline{X}$ dénombrable $\Pr\,[\limsup C_p \subset C] = 1$.

En effet dans 1) l'événement complémentaire s'écrit :

$$A_p : \exists\, x \notin C \qquad x \in C_p$$

$$\Pr\,[A_p] \leqslant \sum_{x \in \underline{X} - C} \Pr(x \in C_p) \quad \text{(somme finie)}$$

et l'on a vu que ces probabilités tendent vers 0,ainsi $\Pr\,[A_p] \longrightarrow 0$.

Dans 2) l'événement complémentaire s'écrit :

$$A : \exists\, x \notin C \qquad x \in \limsup C_p$$

$$\Pr A \leqslant \sum_{x \in \underline{X} - C} [\, \Pr \;\; x \in \limsup C_p] \quad \text{(somme dénombrable)}$$

et l'on a vu que toutes ces probabilités sont nulles. Q.E.D.

Les deux restrictions sur le cardinal de X sont indispensables.

En effet :

1) En reprenant le 1er contre-exemple ci-dessus, où $\underline{X}$, dénombrable, = les rationnels de [0,1]

$$C_p \supset C \quad \forall p \qquad \text{et} \qquad Pr\,[C_p \subset C] = 0 \quad \forall p\ .$$

2) Soit U une variable répartie uniformément sur l'ensemble $\underline{X}$ = [0,1] continu.

La fonction : $H_p(x) = -1$ pour $x = 0$ ou U (quelque soit p)
$= 0$ sinon

et presque sûrement égale à

$H(x) = -1$ pour $x = 0$
$= 0$ sinon

On a alors $C_p = \{0,U\}$ quelque soit p

$C = \{0\}$ et par suite

et par suite $Pr\,[\lim\sup C_p \subset C] = Pr\,\{U = 0\} = 0$.

9. AXIOMES DES H-OPERATEURS

Etant donné un opérateur $T : (E,S) \longrightarrow C_E$, ou un algorithme, que faut-il pour que ce soit un H-opérateur ?

L'espace d'images F peut se présenter de lui-même, ensemble des codages possibles d'un objet ; sinon on peut toujours prendre F = E et a(e) = e. Ensuite T opère pour chaque probabilité rationnelle L sur F, s'il vérifie les trois axiomes suivants.

<u>A1. Axiome des images confondues</u>

$a(i) = a(j) \Rightarrow iXj,\ \forall X \in C$; les partitions optimales réunissent les objets qui ont même image. A1 signifie que C_E est image réciproque, par a, d'une classe C de partitions optimales de F. C est déterminé par le doublet (E,a) : $C = T'(E,a) \subseteq P(F)$.

<u>A2. Axiome de transport</u>

$T'(E,a) = T'(E',b)$ lorsqu'il existe une bijection C de E sur E' telle que $a = b \circ C$. Alors chaque élément de F est image du même nombre d'objets dans E et E'.

On dira que a et b ont les mêmes degrés $\hat{a} = \hat{b}$; où $\hat{a}(f) = |a^{-1}(f)|$ par définition.

A_2 signifie que T'(E,a) peut s'acrire $\Theta(\hat{a})$ (F est sous-entendu) et que C, dans P(F) ne dépend que du nombre d'objets codés en f. $\hat{a}$ définit une mesure entière positive (et bornée) sur F, et une probabilité rationnelle :

$l(f) = a(f)/_{|E|}$. C ne dépendra que de cette loi l moyennant l'axiome suivant.

A3. Axiome des échantillons répétés

Si k est un entier > 0, $\Theta(k\hat{a}) = \Theta(\hat{a})$. Θ est homogène de degré 0, et $T(E_p) = T(E_{kp})$ si l'échantillon E_{kp} "répète" k fois chaque élément de E_p.
Ces trois axiomes sont visiblement des exigences naturelles exprimant que le codage a est pertinent pour la visée classificatoire poursuivie.

A4. Pour que T soit un H-opérateur, il faut encore une condition plus restrictive que $C = \varphi(1)$ soit, dans P(F), le centre d'une fonction H(x,1) [x ∈ P(F)] continue en 1 : $C = \{x \in P(F) \mid H(x,1) = \inf_x H(x,1)\}$.

Alors H sera prolongée par continuité pour les lois 1 non rationnelles, et les centres C ainsi définis vérifieront les propriétés du paragraphe 8.

10. STABILITE APPROCHEE DE CERTAINS H-OPERATEURS

Considérons une distance d entre partitions x et y ∈ P(F), et les centres C et C_p de H(x,1) et $H(x,1_p)$ envisagés comme fonctions de x. Si 1_p tend vers 1, $H(x,1_p) \longrightarrow H(x,1)$ pour chaque x.
Cependant la distance d(x,y) entre un x de C et un y de C_p peut ne pas tendre vers 0. On le conçoit en prenant x et y sur l'intervalle [0,1] = X, et $H(x) = 0$, $H_p(x) = x/_p$, alors $C = X$ et $C_p = \{0\}$ pour tout p.
La distance de Hausdorff entre C et C_p est 1 pour tout p : (celle entre y = 0 et x = 1).
Il faut qu'un axiome supplémentaire établisse un lien entre la distance d et les fonctions H et H_p. Sinon, même quand $C_p \subseteq C$ pour p grand, quand l'ensemble $C - C_p$ est non vide, ses éléments y sont à une distance arbitraire de C.
L'axiome le plus naturel nous semble être le suivant : nous dirons que la fonction H(x) est *à minimum métrisé* s'il existe une fonction de $\mathbb{R}$ dans $\mathbb{R}$: $a \longrightarrow \varphi(a) > 0$, croissante, telle que, C étant le centre $\{x \mid H(x) = \inf_y H(y)\}$.
$d(y,C) > a$ entraîne $H(y) > H(C) + \varphi(a)$.
En langage intuitif, sur une courbe ou une carte en relief, cela veut dire : il n'y a pas de point très bas et très éloigné de *tous* les points les plus bas.
Les fonctions $H(x,1) = \Sigma\, x_{fg}(s_{fg} - 1/2)\, 1_f\, 1_g$ mentionnées au paragraphe 2 sont à minimum métrisé pour la distance euclidienne :

$$d^2(x,y) = \Sigma\, (x_{fg} - y_{fg})^2\ 1_f\ 1_g$$

l'axiome du minimum métrisé, appliqué à H, suffit pour que, quand $p \longrightarrow \infty$, la distance de Hausdorff entre centres :

$d(C,C_p) \longrightarrow 0$, si la convergence $H_p \longrightarrow H$ est uniforme.
En effet $d(C,C_p) > a$ signifie, il existe $x \in C$ et $y \in C_p$ tels que $a < d(x,C_p)$ et $a < d(y,C)$.

Ce dernier point, joint à l'axiome envisagé, entraîne $H(y) > H(x) + \varphi(a)$. Ainsi $H_p(y) \leqslant H_p(x)$ n'est pas possible pour p grand.

11. CONCLUSIONS

La présente étude doit être clairement située.
Nous avons parlé de stabilité relativement à l'ensemble des objets que l'on veut classer. Une toute autre direction serait relative à la mesure des similarités entre objets, ou d'une autre structure initiale.

Cette étude établit l'impossibilité d'une stabilité stricte, mais laisse ouverte des possibilités de stabilité statistique, et surtout de stabilité approchée.