Présentation et généralisation d'une méthode naïve, inédite, de classification en taxonomie

Maurice Maeck[1]

Résumé

Dans la présente étude, il est démontré qu'une approche multiparamétrique naïve des problèmes de classification (hors clustering), menée en optimisant un nombre restreint de grandeurs (pouvoirs discriminants et limites interclasses), peut constituer une aide appréciable à l'interprétation des résultats produits par des algorithmes sophistiqués dans le domaine de la taxonomie.

Une fonction discriminante optimale, linéaire ou non, est obtenue en considérant les moyennes et les écarts-types des classes, des limites interclasses et le pouvoir discriminant des paramètres.

Une double optimisation est réalisée au moyen du solveur d'Excel selon une méthode de gradients (GRG, Generalized Reduced Gradient). Ce procédé se révèle supérieur à l'utilisation de limites interclasses basées sur les coefficients de Student tels qu'ils furent pris en compte dans une précédente publication (Maeck, 2018).

Après une présentation détaillée, la méthode est appliquée à quelques cas de classification en taxonomie végétale et animale et, ensuite, étendue à des bases de données plus générales pour en comparer les performances avec des outils actuels de classification.

Mots clés :classification, taxonomie, biométrie, statistique multiparamétrique, bases de données, pouvoir discriminant, limites interclasses

Summary

In the present study, it is shown that a naïve multiparametric approach to classification problems (excluding clustering), carried out by optimizing a limited number of quantities (discriminating powers and interclass boundaries), can constitute an appreciable aid to the interpretation of the results produced by sophisticated algorithms in the field of taxonomy.

An optimal discriminant function, linear or not, is obtained by considering the means and the standard deviations of the classes, the interclass limits and the discriminating power of the parameters.

A double optimization is carried out using the Excel solver according to a gradients method (GRG, Generalized Reduced Gradient). This method proves to be superior to the use of interclass boundaries based on Student's coefficients as they were taken into account in a previous publication (Maeck, 2018).

After a detailed presentation, the method is applied to a few classification cases in plant and animal taxonomy and then extended to more general databases to compare performance with current classification tools.

Keywords :classification, taxonomy, biometrics, multiparametric statistics, discriminant power, databases, interclass boundaries

[1] Rue du bois de Malmarais 26, 6567 Labuissière. Courriel (e-mail) : maeckma@gmail.com

Introduction

La notion de base de données recouvre des situations très différentes (Wikipédia, 2019) dont la plus élémentaire, qui nous concerne ici, est celle d'un tableau de données, numériques ou non, où chaque entrée (ligne) correspond à une série d'observations relative à un individu associé à une classe et plusieurs descripteurs (paramètres, variables) disposés en colonnes. Le Tableau 1 ci-dessous prend en compte 3 classes (mammifère, oiseau, insecte) et 4 descripteurs (P1 = masse, P2 = nombre de pattes, P3 = présence de plumes et P4 = nombre de dents). La colonne des classes est muette ; elle invite le lecteur à la réflexion et à la remplir selon son appréhension de la biologie animale en tenant compte d'une nécessaire variabilité naturelle (ou non). Les grandeurs non numériques seront associées arbitrairement à des entiers (pour la présence de plumes, par exemple : non = 0, oui = 1) de manière à pouvoir calculer une moyenne et une dispersion.

Individu	Classe	P1	P2	P3		P4	Commentaire
observé		Masse /kg	Nombre de pattes	Présence de plumes		Nombre de dents	
1		70	4	non	0	32	Facile
2		0,030	2	oui	1	0	
3		0,002	6	non	0	0	
4		2000	4	non	0	6	Braconnage
5		2000	4	non	0	4	
6		1,00	2	oui	1	0	Facile
7		82,0	3	non	0	15	Facile
8		4000	0	non	0	0	Archimède
9		43,0	2	oui	1	0	Facile
10		1,00	4	non	0	0	Bec plat
11		0,01	8	non	0	0	Tisseuse
12		0,020	2	non	0	0	Antennes
13		0,070	6	non	0	0	Le poids lourd
14		0,0001	4	non	0	0	Guerrières
15		0,300	2	non	0	0	Rhea
16		1,50	0	oui	1	0	Accident

Tableau 1. Structure de la mini-base de données construite au titre d'exemple.

Une fois la colonne des classes remplie, quelques observations s'imposent par simple inspection visuelle :

1. Le troisième descripteur (P3) isole (= discrimine) presque parfaitement la classe des oiseaux ; il est, par contre, complètement incapable de séparer les deux autres classes.
2. Le deuxième descripteur (P2) discrimine assez bien les trois classes mais peut se révéler sujet à des accidents de vie.
3. Le premier descripteur (P1) isole correctement les insectes en dessous du gramme ; il n'est d'aucune utilité pour des masses supérieures.
4. Le quatrième descripteur (P4) discrimine correctement les mammifères mais évolue fréquemment au cours de la vie.
5. Le onzième individu est un intrus.

Une démarche quantitative permettra d'objectiver ces observations. Dans ce but il convient, pour chaque descripteur, d'établir les paramètres de position (moyennes, médianes …) et de largeur (étendues, écarts-types …) des trois classes proposées. Ainsi, pour P1, par exemple :

- Moyenne des mammifères = (70 + 2000 + 2000 +82 + 4000 + 1)/6 = 1359 kg
  Etendue des mammifères = 4000 - 1 = 3999 kg
  Ecart-type des mammifères = $[(70 - 1359)^2 + (2000 - 1359)^2 + (2000 - 1359)^2 + (82 - 1359)^2 + (4000 - 1359)^2 + (1 - 1359)^2]^{½}/\sqrt{5}$ = 1608 kg

- Moyenne des oiseaux = 9,17 kg
  Etendue des oiseaux = 42,97 kg
  Ecart-type des oiseaux = 18,92 kg

- Moyenne des insectes = 0,0230 kg
  Etendue des insectes = 0,0699 kg
  Ecart-type des insectes = 0,0326 kg

Conformément aux hypothèses usuelles on considère que les masses attendues pour une classe donnée se situent dans l'intervalle de la moyenne plus ou moins deux écarts-types qui correspond à un taux de confiance de 95 % :

masse des mammifères = 1359 ± 3216 kg → $0 \leq 1359 \leq 4575$ kg
masse des oiseaux = 9,17 ± 37,84 kg → $0 \leq 9,17 \leq 47,01$ kg
masse des insectes = 0,0230 ± 0,0652 kg → $0 \leq 0,0230 \leq 0,0882$ kg

En conclusion

1. Toutes les masses jusqu'à 4575 kg peuvent être attribuées à un mammifère ce qui recouvre les deux autres classes.
2. Une masse supérieure à 47,01 kg exclut un oiseau.
3. Une masse supérieure à 0,0882 kg exclut un insecte.

Comme observé précédemment, la masse d'un individu est très peu discriminante malgré des moyennes fort différentes ; la faute en incombe aux largeurs des distributions comme en attestent des écarts-types supérieurs aux moyennes.

Le nombre de pattes (P2) semble plus pertinent avec, souvent, 4 pattes pour un mammifère, 2 pattes pour un oiseau et 6 pattes pour un insecte. Le calcul précédent est repris pour ce paramètre.

Moyenne mammifères = 3,17	Moyenne oiseaux = 1,60	Moyenne insectes = 4,50
Ecart-type mammifères = 1,60	Ecart-type oiseaux = 0,89	Ecart-type insectes = 1,91

nombre de pattes des mammifères = 3,17 ± 3,20 → $0 \leq 3,17 \leq 6,37$
nombre de pattes des oiseaux = 1,60 ± 1,78 → $0 \leq 1,60 \leq 3,38$
nombre de pattes des insectes = 4,50 ± 1,91 → $2,59 \leq 4,50 \leq 6,41$

Ici également, les classes se recouvrent d'une manière appréciable. En nombres entiers et en tenant compte des contraintes biologiques, un individu sera classé comme mammifère s'il présente de 0 à 4 pattes, ce sera un oiseau entre 0 et 2 pattes et un insecte entre 3 et 6 pattes.

Ce sont les accidents de vie, des pattes perdues, qui brouillent la classification pour ce descripteur.

La variable P3 (présence de plumes) sera-t-elle plus discriminante ?

Elle devrait, en principe, discriminer complètement les oiseaux sans séparer les deux autres classes.

Moyenne mammifères = 0,00	Moyenne oiseaux = 0,80	Moyenne insectes = 0,00
Ecart-type mammifères = 0,00	Ecart-type oiseaux = 0,45	Ecart-type insectes = 0,00

Comme l'échantillon ne comporte pas de mammifères ou d'insectes à plumes, ces deux classes sont confondues. La présence d'un (célèbre) oiseau nu empêche cependant une complète séparation des oiseaux puisque la borne inférieure (0,80 - 2.0,45 = - 0,1) englobe la valeur nulle qui sanctionne l'absence de plumes. Seuls les individus pourvus d'au moins une plume pourront être classifiés comme des oiseaux.

Il reste à examiner le pouvoir discriminant du dernier paramètre (P4).

Moyenne mammifères = 9,50	Moyenne oiseaux = 0,00	Moyenne insectes = 0,00
Ecart-type mammifères = 12,32	Ecart-type oiseaux = 0,00	Ecart-type insectes = 0,00

La situation est similaire à celle du calcul précédent avec, cette fois les classes des oiseaux et des insectes confondues. L'énorme écart-type pour les mammifères ne permet pas d'isoler complètement cette classe. Il est simplement possible de conclure qu'un individu, pourvu d'au moins une dent sera un mammifère.

Les quatre descripteurs sont de (très) médiocres facteurs discriminants pour les trois classes sur la base des 15 entrées pertinentes. Mais alors :

> *Comment est-il possible qu'un lecteur attentif puisse classifier correctement (presque) tous les individus même sans le recours à la colonne des commentaires ?*

Cet « exploit » résulte d'une mise en collaboration visuelle des descripteurs lorsque le regard se pose sur la ligne d'un individu ! C'est la prise en compte simultanée, éventuellement assortie d'un facteur de confiance, des quatre paramètres qui mène à une classification presque parfaite. La transposition mathématique de cette collaboration revient à construire une fonction discriminante (F) qui remplacera les quatre valeurs individuelles d'une entrée. Une simple combinaison linéaire améliore déjà souvent la classification.

$$F = a_1.P1 + a_2.P2 + a_3.P3 + a_4.P4 \qquad (a_1, a_2, a_3 \text{ et } a_4 \text{ sont des constantes})$$

Les constantes de la fonction discriminante sont déterminées par optimisation d'un critère de classification bien choisi. La méthode que nous proposons adopte souvent une combinaison linéaire assortie de l'optimisation de son pouvoir discriminant défini ci-dessous.

Le pouvoir discriminant d'un descripteur (définition quantitative)

La base de données « Iris » (Dua & Karra Taniskidou, 2017) sera exploitée pour réaliser la présentation détaillée de notre méthode. On y distingue 3 classes (*Iris setosa*, *Iris versicolor*, *Iris virginica*), 4 descripteurs (longueur des pétales, largeur des pétales, longueur des sépales, largeur des sépales) et 50 entrées (individus) pour chaque classe soit un total de 150 entrées.

Pour les besoins d'une première optimisation, il est nécessaire de définir un critère de pouvoir de discrimination (Δ) des classes pour chaque descripteur. Dans le cas de deux classes, nous proposons l'expression sans dimension suivante qui respecte le sens physique des grandeurs statistiques mises en jeu (moyenne et écart-type)

$$\Delta = \frac{|m_1 - m_2|}{\sqrt{s_1^2 + s_2^2}}$$

où m_1 et m_2 sont les moyennes des deux classes
s_1 et s_2, les écarts-types associés

Ce choix exprime que plus un descripteur conduit à des moyennes différentes, plus il sera discriminant. Il en va de même, en raison inverse, des écarts-types puisque deux distributions, même fort écartées, pourront se recouvrir si leurs largeurs sont suffisamment importantes. La Figure 1 présente trois simulations basées sur des distributions normales (100 entrées par classe) de manière à évaluer la valeur d'un pouvoir discriminant associé à une bonne séparation des classes.

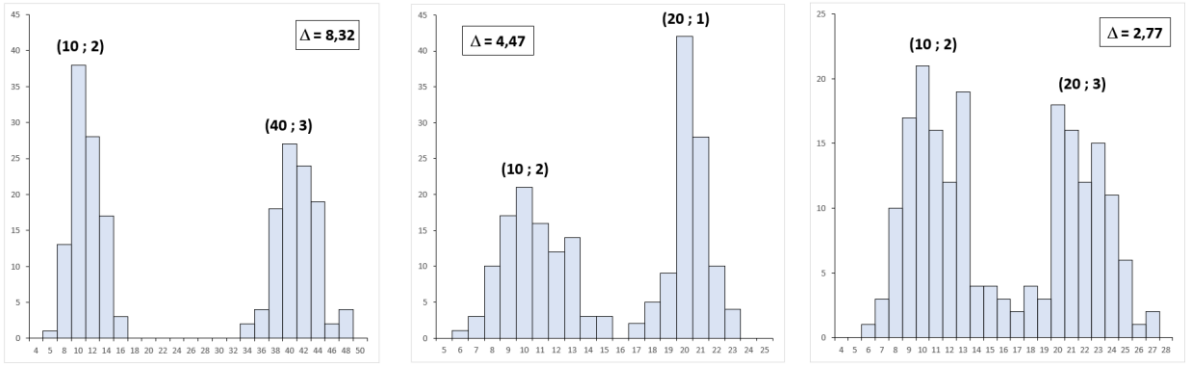


Fig. 1. Visualisation de la relation entre la valeur du pouvoir discriminant d'un descripteur et le recouvrement des histogrammes de deux classes pour une simulation de distributions normales de paramètres (moyenne ; écart-type).

Il apparaît clairement qu'un pouvoir discriminant supérieur à 4,5 est nécessaire à une bonne séparation des classes. Cette limite n'est qu'une évaluation car elle dépend évidemment des caractéristiques des distributions des mesures des classes.

La première étape de la méthode proposée consiste à classer les descripteurs par ordre décroissant de leurs pouvoirs discriminants. Ceci est fondamental car, dans la combinaison (linéaire) subséquente, la contribution du descripteur le plus discriminant sera fixée à l'unité de sorte qu'il fasse office de contribution de référence ; les optimisations se révèlent « flottantes » et divergentes sans cette précaution qui apparaît également dans l'article de Fisher (Fisher, 1936) relatif à la base de données des iris.

Lorsqu'il y a plus de deux classes, le classement est basé sur la somme de Δ1, Δ2, Δ3 …

A l'inspection du Tableau 2, il apparaît immédiatement que les deux descripteurs liés aux pétales sont les plus discriminants et que la discrimination entre *Iris setosa* et *Iris. versicolor* est bien meilleur que celle entre *Iris versicolor* et *Iris virginica.*

Raw data for the Iris Database					
#	**Species**	**Petal lenght /cm**	**Petal witdth /cm**	**Sepal lenght /cm**	**Sepal width /cm**
1	setosa	1.40	0.20	5.10	3.50
	⋮	⋮	⋮	⋮	⋮
50	setosa	1.40	0.20	5.00	3.30
51	versicolor	4.70	1.40	7.00	3.20
	⋮	⋮	⋮	⋮	⋮
100	versicolor	4.10	1.30	5.70	2.80
101	virginica	6.00	2.50	6.30	3.30
	⋮	⋮	⋮	⋮	⋮
150	virginica	5.10	1.80	5.90	3.00
Discrimination analysis for the Iris Database					
	Mean setosa =	1.46	0.25	5.01	3.43
	SD setosa =	0.17	0.11	0.35	0.38
	Δ1 =	**5.59**	**4.82**	**1.49**	**1.34**
	Mean versicolor =	4.26	1.33	5.94	2.77
	SD versicolor =	0.47	0.20	0.52	0.31
	Δ2 =	**1.78**	**2.07**	**0.80**	**0.45**
	Mean virginica =	5.55	2.03	6.59	2.97
	SD virginica =	0.55	0.27	0.64	0.32
	Sum Δ =	**7.37**	**6.89**	**2.29**	**1.79**

Tableau 2. Classement des descripteurs en fonction de la somme de leurs pouvoirs discriminants.

La mise en collaboration des deux premiers descripteurs (pétales) devrait conduire à une séparation complète de *Iris setosa* tandis que la discrimination entre les deux autres espèces restera probablement problématique.

Le pouvoir discriminant d'une combinaison (linéaire) des descripteurs

La première optimisation a pour objectif de maximiser le pouvoir discriminant d'une combinaison (linéaire) des descripteurs. Le Tableau 3 rend compte du résultat dans le cas des iris de Fischer. L'optimisation ne porte d'abord que sur a2 seul, puis sur le couple a2 et a3 et, finalement, sur les trois coefficients a2, a3 et a4. Avant d'être sollicités, tous les coefficients ont été mis à zéro.

#	Species	Coefficients	Combinations
1	setosa	a1 = 1,00	-3,65
2	setosa	a2 = 1,78	-3,15
3	setosa	a3 = -0,51	-3,30
4	setosa	a4 = -0,80	-2,97
5	setosa		-3,68
6	setosa		-3,47
7	setosa		-3,14
8	setosa		-3,42
9	setosa		-2,81

Mean setosa =	-3,40
SD setosa =	0,46
Δ1 =	**6,88**
Mean versicolor =	1,37
SD versicolor =	0,52
Δ2 =	**2,67**
Mean virginica =	3,41
SD virginica =	0,56
Sum Δ =	**9,55**

Lorsque, comme ici, plus de deux classes sont mises en jeu, l'optimisation porte sur la somme des pouvoirs discriminants des descrip-teurs. Pour le premier indi-vidu, on a bien :

1,00.1,40 + 1,78.0,20 - 0,51.5,10 - 0,80.3,50 = -3,65

Tableau 3. Première optimisation pour la base de données des iris.

Il est remarquable (et inattendu) que le deuxième descripteur (largeurs des pétales) prenne le pas sur le premier (longueur des pétales) dans la combinaison linéaire.

A ce stade, à l'issue de la première optimisation, une bonne séparation d'*Iris setosa* semble acquise.

L'optimisation des limites interclasses

La nécessité pour notre méthode de proposer une classification sur la base d'une série de données des descripteurs (= une entrée) exige de définir des limites interclasses comme le montre la Figure 2 de sorte qu'une valeur inférieure (supérieure) à une limite donnée soit sanctionnée par une classification dans la classe de gauche (droite).

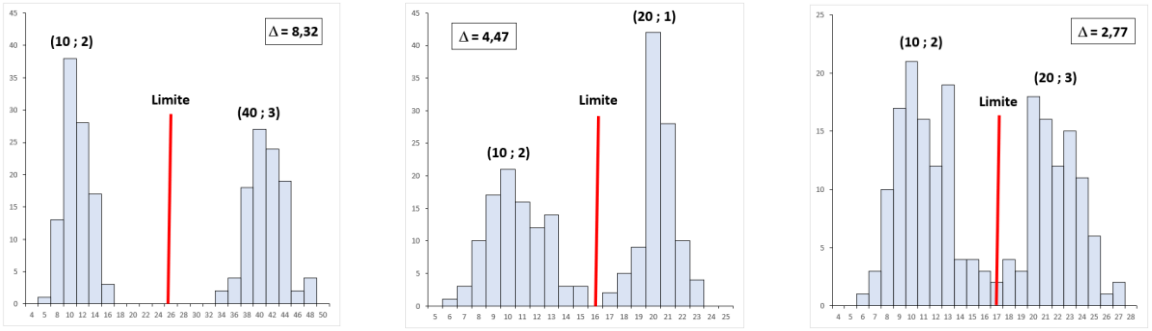


Fig. 2. Trois cas de disposition d'une limite en fonction du recouvrement des histogrammes de deuxclasses adjacentes

- si Δ est grand : il existe toute une gamme de valeurs acceptables, le logiciel n'aboutira pas nécessairement à une disposition symétrique entre les deux histogrammes mais la convergence est très robuste vis-à-vis des valeurs initiales des limites
- si Δ est critique : la convergence, souvent robuste, se limite à un intervalle restreint
- si Δ est petit : l'optimisation doit prendre en compte des contraintes sur les limites pour éviter une divergence ; les valeurs acceptables pour une limite sont restreintes à l'intervalle entre les deux moyennes des classes adjacentes

Cette seconde optimisation a pour objectif de minimiser le nombre d'erreurs de classification en faisant varier les limites. Pour la classification des iris de Fisher, on obtient un résidu de deux erreurs comme le Tableau 4 l'indique.

#	Species	Coefficients	Combinations
1	setosa	a1 = 1,00	-3,65
2	setosa	a2 = 1,78	-3,15
3	setosa	a3 = -0,51	-3,30
4	setosa	a4 = -0,80	-2,97
5	setosa		-3,68
6	setosa		-3,47
7	setosa		-3,14
8	setosa		-3,42
9	setosa		-2,81
10	setosa		-3,30
11	setosa		-3,86
12	setosa		-3,21

Mean setosa =	-3,40
SD setosa =	0,46
Δ1 =	**6,88**
Mean versicolor =	1,37
SD versicolor =	0,52
Δ2 =	**2,67**
Mean virginica =	3,41
SD virginica =	0,56
Sum Δ =	**9,55**

Limit1 =	**-2,26**
Limit2 =	**2,44**

Classifications	
setosa	1
setosa	1
setosa	1
setosa	1
setosa	1
setosa	1
setosa	1
setosa	1
setosa	1
setosa	1
setosa	1
setosa	1

Entries	Errors
150	2
Success %	**98,7**

Success %	
setosa	**100,0**
versicolor	**98,0**
virginica	**98,0**

Tableau 4. Résultats de classification des iris de Fischer après la seconde optimisation.

Sans surprise, tous les *Iris setosa* sont correctement classés alors que les deux erreurs se répartissent équitablement entre *Iris versicolor* et *Iris virginica*. Le taux global de réussite de plus de 98 % classe la méthode proposée parmi les meilleurs scores actuels (Zielesny, 2016) (Lantz, 2019). La mise en œuvre de fonctions discriminantes non linéaires n'a pas été systématiquement explorée ; elle constitue probablement un axe d'amélioration intéressant qui sera abordé ultérieurement.

La classification des *Platantheras* « intermédiaires »

(Esposito et al., 2018) (Ambroise et al., 2020)

Ces articles s'interrogent sur la caractérisation et l'identification d'individus paraissant intermédiaires, observés sur deux sites sympatriques (Botton et Bois Niau) en compagnie d'individus des parents putatifs (*P. bifolia* et *P. chlorantha*). Pour la présente analyse, les paramètres des deux classes parentes sont déterminés d'après des mesures réalisées sur deux sites allopatriques (Navaugle pour *P. bifolia* et Transinnes pour *P. chlorantha*) jugés homogènes et représentatifs.

La base de données, à 2 classes et 4 descripteurs, se compose de 26 entrées pour *P. bifolia* et 41 entrées pour *P. chlorantha*. Le classement des descripteurs selon leur pouvoir discriminant montre d'emblée qu'une très nette séparation des deux classes sera de mise (cf. Tableau 5).

	Caud.lenght	Visc.dist	Lab.lenght	Spur.length
Mean bifolia =	0,22	0,25	9,69	20,82
SD bifolia =	0,08	0,11	1,43	2,38
Discrim =	**8,28**	**7,54**	**1,76**	**1,34**
Mean chlorantha =	1,79	3,90	13,76	25,72
SD chlorantha =	0,17	0,47	1,81	2,76

Tableau 5. Classement des descripteurs selon leur pouvoir discriminant (Δ).

Le Tableau 6 confirme cette prévision puisqu'aucune erreur n'apparaît après la double optimisation.

Coefficients		Combinations
a1 =	1,000	0,70
a2 =	0,333	0,53
a3 =	-0,006	0,66
a4 =	0,017	0,56
		0,56
		0,64
		0,58
		0,48

Mean bifolia =	0,60
SD bifolia =	0,09
Δ =	**10,93**
Mean chlorantha =	3,46
SD chlorantha =	0,24
Limit =	**2,66**

Classifications	
bifolia	1
bifolia	1
bifolia	1
bifolia	1
bifolia	1
bifolia	1
bifolia	1
bifolia	1

Entries	Errors
67	**0**
Success %	**100,0**

Success %	
bifolia	**100,0**
chlorantha	**100,0**

Tableau 6. Résultats de classification des *Platantheras* allopatriques après la seconde optimisation.

Il reste maintenant à proposer à la classification les 51 *P. intermedia* (33 individus pour Botton et 18 individus pour Niau) … qui sont tous identifiés comme des *P. bifolia* ainsi que le montre l'extrait du Tableau 7 !

Species	Caud.lenght	Visc.dist	Lab.lenght	Spur.length	Coefficients	Combinaisons.	Class.
intermedia	0,80	1,50	11,34	30,19	a1 = 1,000	1,76	bifolia
intermedia	0,60	1,12	11,49	29,52	a2 = 0,333	1,42	bifolia
intermedia	0,60	1,52	15,20	37,50	a3 = - 0,006	1,67	bifolia
intermedia	0,70	1,31	13,83	34,40	a4 = 0,017	1,65	bifolia
intermedia	0,70	1,17	12,39	30,43		1,55	bifolia
intermedia	0,60	1,02	12,36	31,92	**Limit = 2,66**	1,42	bifolia

Tableau 7. Classification des *P. intermedia* des deux sites (Botton et Bois Niau).

Contrairement aux conceptions antérieures citées dans Esposito et al., (2018) et Ambroise et al., (2020), les individus paraissant intermédiaires ne sont pas des hybrides, mais bien des représentants de l'espèce *P. bifolia*.

Par contre, dans l'étude de Durka et al. (2017) et Baum & Baum (2017), on a bien un taxon différent, non hybride également. Il serait intéressant de soumettre les données de ces auteurs à l'analyse menée ici.

Le sexage biométrique du Pinson des arbres (*Fringilla cælebs*) (Le Boulengé et al., 2018)

La détermination des paramètres de classe et de la limite porte donc sur 639 entrées avec un bilan initial, rapporté dans le Tableau 8, qui laisse entrevoir quelques difficultés de sexage liées au modeste pouvoir discriminant des deux descripteurs.

Mean LA =	84,06	20,69
SD LA =	1,83	1,46
Δ =	**2,01**	**1,09**
Mean MC =	90,08	22,99
SD MC =	2,38	1,52

#	sexe	LA /mm	MC /g
1	F	83,0	20,7
2	F	84,0	19,6
3	F	83,0	19,5

Dans cet article, les auteurs s'interrogent sur la possibilité d'un sexage biométrique des passereaux sans dimorphisme sexuel visible. Dans une première partie, le projet se concentre sur le Pinson des arbres dont le sexe est identifiable visuellement. La base de données, aimablement mise à disposition par les auteurs, comporte 2 classes (M = mâle, F = femelle), 2 descripteurs numériques directement exploitables (LA = longueur de l'aile, MC = masse corporelle) et 839 entrées fiables. Une extraction aléatoire de 200 entrées est réservée comme échantillon test.

Tableau 8. Pouvoir discriminant des deux descripteurs.

L'analyse de cette base de données se révèle cependant étonnamment efficace (cf. Tableau 9).

#	sexe	Coefficients	Combinations
1	F	a1 = 1,0000	98,02
2	F	a2 = 0,7257	98,22
3	F		97,15
4	F		98,54
5	F		99,42
6	F		97,67
7	F		100,56

Mean F =	99,0725
SD F =	2,1391
Δ =	**2,2419**
Mean M =	106,766
SD M =	2,6833
Limit =	**102,46**

Classification	
F	1
F	1
F	1
F	1
F	1
F	1
F	1

Entries	Errors
639	35

Success %	
F	95,0
M	93,7

Tableau 9. Résultats de l'optimisation en vue du sexage biométrique des pinsons des arbres (*Fringilla cælebs*).

Un taux de succès global proche de 95 % avec seulement deux descripteurs modérément discriminants justifierait la recherche d'une troisième variable capable d'améliorer ce score.

Pour évaluer le pouvoir prédictif de la classification, il faut faire appel aux 200 entrées initialement sélectionnées de manière aléatoire au départ du fichier complet. Le Tableau 10 rapporte le résultat de cette classification.

#	sexe	LA	MC
Test File 200 entries			
1	F	83,1	20,5
2	F	82,7	20,7
3	F	83,4	19,7
4	F	83,1	19,5
5	F	87,0	19,0
6	F	83,5	23,0

Limit = 102,46	
Coefficients	
a1 =	1,0000
a2 =	0,7257

Combinaisons	Classifications	Success
98,019	F	1
97,724	F	1
97,741	F	1
97,295	F	1
100,780	F	1
100,191	F	1

Entries 200	Errors 17
% OK =	**91,5**

Success %	
F	**89,7**
M	**92,9**

Tableau 10. Classification du fichier test de la base de données des pinsons des arbres (*Fringilla cælebs*).

Le pouvoir prédictif diminue significativement pour ce fichier test ce qui jette le doute sur la possibilité d'un sexage biométrique de cette espèce.

Attention : *Dans l'interprétation d'un taux du succès, il ne faut jamais perdre de vue qu'un descripteur totalement inadéquat aura, en moyenne, un taux de succès de 50 % pour deux classes, 33 % pour trois classes, 25 % pour 4 classes …*

Une tentative de sexage de Fauvettes à tête noire, *Sylvia atricapilla*, autre espèce dont le plumage diffère entre les sexes, d'après une extraction aimablement mise à disposition par le centre de baguage BeBirds de l'Institut royal

des Sciences naturelles de Belgique (2 classes, 6 descripteurs, 505 entrées) livre des résultats de classification encore bien plus médiocres (Tableau 11). Le sexage biométrique n'a, de toute évidence, pas encore trouvé ses descripteurs adéquats.

Sex	Coefficients	Combinations
F	a1 = 1,000	70,13
F	a2 = 0,229	67,99
F	a3 = 0,822	72,80
F	a4 = -0,333	73,44
F	a5 = 0,553	70,28
F	a6 = 0,719	69,10
F		70,97
F		70,66

Mean F =	70,30
SD F =	2,35
Δ =	**0,153**
Mean M =	69,79
SD M =	2,40
Limit =	**69,80**

Classifications	
F	1
M	0
F	1
F	1
F	1
M	0
F	1
F	1

Entries	Errors
505	**214**
Success %	**57,6**

Success %	
F	**56,7**
M	**58,6**

Tableau 11. Succès de classification pour la base de données des Fauvettes à tête noire (*Sylvia atricapilla*).

Classification des graines de trois variétés de blé (Seeds Database, cf. Annexe)

Les auteurs examinent des graines appartenant à trois variétés différentes de blé : Kama (1), Rosa (2) et Canadian (3). La base de données comporte 3 classes, 7 descripteurs et 70 entrées pour chaque variété. Une visualisation de haute qualité de la structure interne du noyau a été étudiée à l'aide d'une technique de rayons X doux. Le Tableau 12 se rapporte à la classification de cette base de données par la méthode proposée ici.

Coefficients	Combinations
a1 = 1,000	125,77
a2 = 11,690	125,73
a3 = -8,955	122,95
a4 = -12,420	119,77
a5 = 0,006	129,39
a6 = -0,531	123,17
a7 = 45,250	123,74
	120,97
	129,68
	129,42
	125,83
	122,28

Mean 3 =	112,16
SD 3 =	3,04
Δ1 =	**1,92**
Mean 1 =	122,41
SD 1 =	4,37
Δ2 =	**2,09**
Mean 2 =	135,75
SD 2 =	4,64
Sum Discrim. =	**4,02**
Limit1 =	**116,82**
Limit2 =	**129,71**

Classifications	
1	1
1	1
1	1
1	1
1	1
1	1
1	1
1	1
1	1
1	1
1	1
1	1

Entries	Errors
210	**21**
Success %	**90,0**

Success %	
1	**87,1**
2	**88,6**
3	**94,3**

Tableau 12. Classification de la base de données Seeds.

Il apparaît que les descripteurs négligent la discrimination entre les variétés 1 et 3. Ce point peut être partiellement corrigé en choisissant 1 comme cible préférentielle de la seconde optimisation. Ce choix conduit à des taux de succès équilibrés entre les trois variétés (1 : 87,1 % ; 2 : 88,6 % et 3 : 94,3 %) pour un succès global de 90,0 %.

La littérature (Charytanowicz et al., 2010) rapporte un taux de succès global de presque 92 % avec une répartition (1 :96 % ; 2 : 84 % et 3 : 96 %) légèrement moins homogène.

Détermination d'un type de forêt (Forest type mapping Database, cf. Annexe)

Cet ensemble de données contient des données d'apprentissage et de test issues d'une étude par télédétection ayant permis de cartographier différents types de forêts en fonction de leurs caractéristiques spectrales à des longueurs d'onde du visible au proche infrarouge, en utilisant les images satellite ASTER. Le résultat (une carte de type de forêt) peut être utilisé pour identifier et/ou quantifier les flux écologiques (stockage du carbone, protection contre l'érosion, par exemple) liés à la forêt.

Les auteurs (Johnson et al., 2012) distinguent 4 classes (forêt 'Sugi', forêt 'Hinoki', forêt 'mixte'/'feuillus' et autres terres non forestières). Neuf descripteurs sont extraits des images (198 entrées pour le fichier « Training » et 325 pour le fichier « Test »).

L’analyse de cette base de données apparaît dans le Tableau 13.

Mean =	27,80	49,68	51,05	55,68	24,29	91,66	53,31	76,53	93,47
SD =	1,87	3,93	2,33	3,08	1,78	4,81	8,35	9,29	7,72
Discrim1 =	**0,64**	**0,11**	**1,38**	**0,19**	**0,38**	**0,58**	**0,43**	**0,75**	**0,43**
Mean =	29,40	50,23	55,15	56,46	25,10	95,25	57,97	91,57	97,67
SD =	1,67	2,90	1,82	2,81	1,22	3,92	7,03	17,88	5,99
Discrim2 =	**2,22**	**1,95**	**1,02**	**0,95**	**0,32**	**0,57**	**0,48**	**0,32**	**0,26**
Mean =	48,31	63,19	68,41	60,06	26,07	103,31	65,08	98,00	103,62
SD =	8,35	5,99	12,92	2,55	2,81	13,68	13,22	8,65	21,87
Discrim3 =	**0,83**	**1,16**	**0,76**	**1,62**	**1,60**	**0,87**	**0,87**	**0,08**	**0,44**
Mean =	66,54	77,70	87,97	77,95	44,19	118,11	78,27	98,74	113,67
SD s	20,31	11,02	22,42	10,73	10,99	9,95	7,47	2,20	6,17
SUM =	3,69	3,22	3,16	2,76	2,29	2,02	1,77	1,15	1,13

#	Class	P2	P5	P3	P9	P8	P6	P1	P7	P4
1	d	36	59	57	60	27	101	39	93	91
2	d	49	64	66	59	28	106	57	114	103
3	d	39	61	58	57	26	99	40	89	82
4	d	32	53	53	50	22	97	34	59	97

#	Class	Coefficients	Combinations
1	d	a1 = 1,000	15,415
2	d	a2 = -0,103	19,906
3	d	a3 = -0,479	17,165
4	d	a4 = -0,636	15,143
5	d	a5 = 0,940	19,157
6	d	a6 = 0,375	23,659
7	d	a7 = -0,456	15,061
8	d	a8 = 0,040	23,167
9	d	a9 = 0,019	12,151
10	d		17,309
11	d		18,390
12	d		18,069
13	d		17,386
14	d		15,247
15	d		21,913
16	d		19,626

Mean h =	-10,659
SD h =	3,276
Discrim1 =	**1,870**
Mean s =	-1,546
SD s =	3,609
Discrim2 =	**4,230**
Mean d =	17,919
SD d =	2,856
Discrim3 =	**1,120**
Mean o =	28,667
SD o =	9,159
Sum Discrim. =	**7,220**

Limit1 =	**-5,017**
Limit2 =	**5,136**
Limit3 =	**21,975**

Classifications	
d	1
d	1
d	1
d	1
d	1
o	0
d	1
o	0
d	1
d	1
d	1
d	1
d	1
d	1
d	1
d	1

Entries	Errors
198	20
Success %	89,90

Success %	
d	92,59
h	97,92
o	81,08
s	86,44

#	Class	P2	P5	P3	P9	P8	P6	P1	P7	P4
					Test File					
1	d	51	69	68	67	31	111	67	136	115
2	d	42	66	63	59	28	108	63	111	97
3	d	59	70	84	58	29	104	59	92	93
4	d	44	59	65	59	26	104	57	98	107
5	d	36	57	60	56	28	102	45	83	97
6	d	35	61	56	53	23	105	36	63	90
7	d	34	55	56	55	25	99	50	84	98
8	d	26	52	47	48	22	94	31	53	90
9	d	29	48	54	56	25	94	68	89	112
10	d	36	57	56	57	27	95	55	92	86
11	d	40	58	60	61	28	105	48	95	108
12	d	54	70	79	66	30	109	62	118	105
13	d	56	73	75	58	27	103	60	102	85

Coefficients	
a1 =	1,000
a2 =	-0,103
a3 =	-0,479
a4 =	-0,636
a5 =	0,940
a6 =	0,375
a7 =	-0,456
a8 =	0,040
a9 =	0,019
Limit1 =	**-5,017**
Limit2 =	**5,136**
Limit3 =	**21,975**

Combinaisons	Classifications	Success
16,59	d	1
11,92	d	1
19,52	d	1
12,70	d	1
15,03	d	1
17,03	d	1
9,61	d	1
13,26	d	1
-3,97	s	0
8,33	d	1
16,18	d	1
14,54	d	1
18,06	d	1

Entries	Errors
325	62
Success %	80,9

Success %	
d	90,5
h	86,8
o	58,7
s	79,4

Tableau 13. Classification de la base de données « Forest type mapping » (Discrim = Δ).

La dégradation du succès de classification de 90 % à 80 % entre le fichier de training et celui de test doit être imputée à une mauvaise identification des terres non forestières.
La littérature mentionne des taux de succès entre 82,2 % et 85,9 % dans les meilleurs des cas.

La comestibilité de champignons (Mushroom Database, cf. Annexe)

Cette base de données entend prédire la comestibilité d’un champignon des genres *Agaricus* et *Lepiota* au départ de caractères morphologiques macroscopiques (cf. Annexe). Elle comporte deux classes (edible, poisonous), 22 descripteurs, tous non numériques, et 8124 entrées.

Le descripteur 16 ne contient que la valeur « p » et le descripteur 11 a des données manquantes dont le statut n’a pas encore été défini dans la présente étude. Ces deux paramètres ne sont pas pris en compte.

La première étape de l’analyse consiste à associer une valeur numérique à chaque qualité d’un descripteur en fonction, par exemple, de son classement alphabétique. Ainsi, pour la forme du chapeau :

bell = b = 1 ; conical = c = 2 ; flat = f = 3 ; knobbed = k = 4 ; sunken = s = 5 ; convex = x = 6

L’arbitraire de cette attribution des nombres peut paraître inquiétant ; il n’a cependant que peu d’influence sur le résultat final et est de pratique courante dans le domaine.

Le tri des descripteurs qui apparaît dans le Tableau 14 n’augure, a priori, pas d’une bonne capacité de classification des entrées car le meilleur descripteur a un pouvoir descriptif inférieur à 0,9.

Mean e =	1,0684	7,6226	1,6540	4,0076	1,2852	2,7757	1,8073	4,2833	2,7985	2,1483	2,1255	2,6160	4,2015	7,0989	2,9316	7,0646	1,9544	1,6160	5,3346	4,2662	5,5817	1,0000
SD e =	0,2525	2,6982	0,4758	1,7179	0,4515	0,6129	0,9907	1,5508	0,6904	1,5882	0,3313	1,3012	1,7577	1,5174	0,3308	1,5305	0,2087	0,4864	1,5213	1,7414	2,4350	0,0000
Discrim =	**0,898**	**0,883**	**0,824**	**0,640**	**0,533**	**0,503**	**0,463**	**0,449**	**0,444**	**0,314**	**0,314**	**0,257**	**0,245**	**0,219**	**0,211**	**0,209**	**0,187**	**0,145**	**0,132**	**0,075**	**0,044**	**#DIV/0!**
Mean p =	1,5679	3,8636	1,1593	2,5230	1,0286	2,3596	1,2579	5,0317	2,3943	2,8958	2,0092	3,0552	5,0215	6,5128	3,0020	6,5046	1,9954	1,5148	4,9408	4,4362	5,4219	1,0000
SD p =	0,4954	3,2958	0,3660	1,5562	0,1667	0,5550	0,6550	0,6149	0,5927	1,7715	0,1658	1,1040	2,8480	2,2027	0,0452	2,2064	0,0677	0,4998	2,5720	1,4376	2,6576	0,0000

#	Class	P8	P9	P4	P19	P7	P12	P11	P21	P13	P22	P18	P2	P20	P14	P17	P15	P6	P10	P5	P1	P3	P16
5529	e	2	4	1	2	1	3		5	1	1	2	4	2	8	3	8	2	1	6	3	7	1
5352	e	2	4	1	2	1	3		5	1	1	2	4	2	8	3	8	2	1	6	6	7	1
4648	e	2	4	1	2	1	3		5	1	1	2	4	2	8	3	8	2	1	6	3	8	1
4534	e	2	4	1	2	1	3		5	1	1	2	4	2	8	3	8	2	1	6	6	8	1
4900	e	2	4	1	2	1	3		5	1	1	2	4	2	8	3	8	2	1	6	3	9	1
5189	e	2	4	1	2	1	3		5	1	1	2	4	2	8	3	8	2	1	6	6	9	1

Tableau 14. Tri des descripteurs de la base de données Mushroom en fonction de leur pouvoir discriminant.

Les deux optimisations démontrent cependant l'étonnante efficacité de collaboration de ces 20 descripteurs médiocres (cf. Tableau 15) puisque le succès global de classification est légèrement supérieur à 95 % qui est la valeur rapportée dans la littérature. Ainsi, si un nouvel indidu est classé par la combinaison linéaire trouvée, la probabilité d'une classification erronée est proche d'environ 5 %. Un taux de succès de 95 %, très honorable, n'est cependant pas encore suffisant pour abandonner la règle selon laquelle aucun critère macroscopique ne permet de déterminer à coup sûr la comestibilité d'un champignon.

Class	Coefficients	Combinations
e	a1 = 1,0000	1
e	a2 = -0,0161	1
e	a3 = -0,6874	1
e	a4 = 0,0504	1
e	a5 = -0,9938	1
e	a6 = -0,2835	1
e	a7 = -0,0115	1

Mean e =	-0,9722
SD e =	0,3811
Discrim =	**2,3731**
Mean p =	0,2862
SD p =	0,3687
Limit =	**-0,6317**

Classifications	
p	0
p	0
p	0
p	0
p	0
p	0
p	0

Entries	Errors
8124	376
Success %	**95,4**

Success %	
e	**94,8**
p	**96,0**

Tableau 15. Résultat d'optimisation pour la base de données Mushroom.

Salaire annuel inférieur ou supérieur à 50 k$ (Adult Database, cf. Annexe)

Après les iris de Fischer, c'est la base de données la plus sollicitée pour tester les algorithmes de classification. Elle comporte 2 classes, 14 descripteurs et 48842 entrées (cf. Annexe).
Les données inconnues sont éliminées comme dans la plupart des études de la littérature ; il reste ainsi 30162 entrées pour le fichier « Training » et 15060 entrées pour le fichier « Test » destiné à évaluer le pouvoir prédictif de la méthode.

Les paramètres 2, 4, 6, 8, 9, 10 et 14 ne sont pas de nature numérique. L'attribution d'un nombre à chaque qualité suit l'ordre alphabétique de cette qualité dans chaque descripteur.
Le tri des paramètres fait l'objet du Tableau 16.

Mean <=50K =	9,63	36,61	2,6	39	0,62	3,75	148,89	53,45	11,16	4,64	6,8	37,30	3,19	190339
SD <=50K =	2,41	13,46	1,5	12	0,49	1,62	936,39	310,27	4,07	0,87	4,0	6,19	0,92	106571
Discrim =	**0,585**	**0,434**	**0,423**	**0,396**	**0,390**	**0,365**	**0,263**	**0,210**	**0,140**	**0,123**	**0,085**	**0,039**	**0,029**	**0,015**
Mean >50K =	11,61	43,96	1,7	46	0,85	3,08	3937,68	193,75	11,86	4,78	7,3	37,63	3,23	188150
SD >50K =	2,37	10,27	1,6	11	0,36	0,86	14386,06	592,83	2,83	0,72	4,0	5,84	1,05	102822

#	Class	P5	P1	P8	P13	P10	P6	P11	P12	P4	P9	P7	P14	P2	P3
1	<=50K	10	25	3	40	0	5	0	0	16	2	3	1	3	228608
2	<=50K	2	37	1	40	1	3	0	0	4	2	3	1	3	191342
6	<=50K	9	36	1	40	1	3	0	0	12	2	10	1	3	116138

Tableau 16. Tri des paramètres pour la base de données « Adult ».

Il apparaît immédiatement que tous les descripteurs ont un pouvoir discriminant très pauvre.
La double optimisation proposée dans cette étude conduit donc à un taux de succès modéré rapporté dans le Tableau 17.

#	Class	Coefficients	Combinations
2	<=50K	a1 = 1,0000	12
16	<=50K	a2 = 0,1281	10
6	<=50K	a3 = -0,2615	17
18	<=50K	a4 = 0,0850	15
17	<=50K	a5 = 2,7182	14
9	<=50K	a6 = -0,7515	19
10	<=50K	a7 = 0,0001	19
12	<=50K	a8 = 0,0016	5
1	<=50K	a9 = -0,0131	13
13	<=50K	a10 = 0,3778	19
14	<=50K	a11 = 0,0187	20
36	<=50K	a12 = -0,0057	17
31	<=50K	a13 = -0,3261	10
60	<=50K	a14 = 8,1E-07	16

Mean <=50K = 16,5665
SD <=50K = 4,1447
Discrim = 0,9651
Mean >50K = 21,9945
SD >50K = 3,8018

Limit = 21,8654

Classifications	
<=50K	1
<=50K	1
<=50K	1
<=50K	1
<=50K	1
<=50K	1
<=50K	1
<=50K	1
<=50K	1
<=50K	1
<=50K	1
<=50K	1
<=50K	1
<=50K	1
<=50K	1

Entries	Errors
30162	5893
Success %	80,5

Success %	
<=50K	90,6
>50K	49,8

Tableau 17. Taux de succès de classification pour le fichier « Training » (Discrim = Δ).

Le Tableau 18 rapporte le résultat de classification du fichier « Test ».

#	Class	P5	P1	P8	P13	P10	P6	P11	P12	P4	P9	P7	P14	P2	P3
Adult Test File															
1	<=50K	7	25	4	40	1	5	0	0	2	3	7	39	3	226802
2	<=50K	9	38	1	50	1	3	0	0	12	5	5	39	3	89814
6	<=50K	6	34	2	30	1	5	0	0	1	5	8	39	3	198693
9	<=50K	10	24	5	40	0	5	0	0	16	5	8	39	3	369667
10	<=50K	4	55	1	10	1	3	0	0	6	5	3	39	3	104996
12	<=50K	13	36	1	40	1	3	0	0	10	5	1	39	1	212465
13	<=50K	9	26	2	39	0	5	0	0	12	5	1	39	3	82091

Limit = 21,8654	
Coefficients	
a1 = 1,0000	
a2 = 0,1281	
a3 = -0,2615	
a4 = 0,0850	
a5 = 2,7182	
a6 = -0,7515	
a7 = 0,0001	

Combinaisons
11,738
19,016
12,328
12,337
12,846
22,613
11,980

Classifications	Success
<=50K	1
<=50K	1
<=50K	1
<=50K	1
<=50K	1
>50K	0
<=50K	1

Entries	Errors
15060	3003
% OK =	80,1

Success %	
<=50K	90,1
>50K	49,4

Tableau 18. Taux de succès de classification du fichier « Test ».

Les taux de succès des deux fichiers sont comparables et font ressortir un bon score pour les salaires annuels inférieurs à 50 k$ au détriment des salaires supérieurs à 50 k$.

Il convient maintenant de comparer les taux de succès de la méthode proposée à ceux obtenus selon les algorithmes actuels. C'est ce que rapporte le Tableau 19 (cf. Annexe) :

Algorithm	Succès %
C4.5	84,46
Voted ID3 (0.6)	84,36
1R	80,46
HOODG	85,18
Naive-Bayes	83,88
OC1	84,96

Algorithm	Succès %
C4.5-auto	85,54
Voted ID3 (0.8)	82,53
NBTree	85,90
FSS Naive Bayes	85,95
Nearest-neighbor (1)	78,58
Pebls	crash

Algorithm	Succès %
C4.5 rules	85,06
T2	83,16
CN2	84,00
IDTM (Decision table)	85,54
Nearest-neighbor (3)	79,65

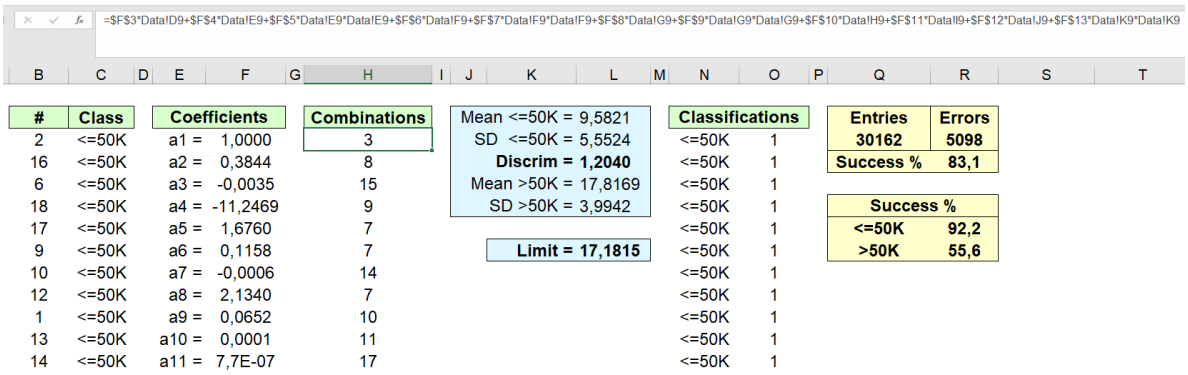
Tableau 19. Performances de plusieurs algorithmes classiques.

Les performances de la méthode proposée (80,5 % et 80,1 %) la situent dans les dernières positions du tableau. Il importe cependant de se souvenir que le calcul est basé sur une combinaison linéaire des contributions des paramètres alors que beaucoup de ces méthodes adoptent des fonctions discriminantes non linéaires.

Pour évaluer l'impact de cette limitation, le calcul a été repris en utilisant une fonction discriminante un peu plus sophistiquée :

$$a1.P5 + a2.P1 + a3.(P1)^2 + a4.P8 + a5.(P8)^2 + a6.P13 + a7.(P13)^2 + a8.P10 + a9.P6 + a10.P11 + a11.(P12)^2$$

Les Tableaux 20 et 21 rapportent les nouveaux résultats d'optimisation.



#	Class	Coefficients	Combinations
2	<=50K	a1 = 1,0000	3
16	<=50K	a2 = 0,3844	8
6	<=50K	a3 = -0,0035	15
18	<=50K	a4 = -11,2469	9
17	<=50K	a5 = 1,6760	7
9	<=50K	a6 = 0,1158	7
10	<=50K	a7 = -0,0006	14
12	<=50K	a8 = 2,1340	7
1	<=50K	a9 = 0,0652	10
13	<=50K	a10 = 0,0001	11
14	<=50K	a11 = 7,7E-07	17

Mean <=50K =	9,5821
SD <=50K =	5,5524
Discrim =	**1,2040**
Mean >50K =	17,8169
SD >50K =	3,9942
Limit =	**17,1815**

Classifications	
<=50K	1
<=50K	1
<=50K	1
<=50K	1
<=50K	1
<=50K	1
<=50K	1
<=50K	1
<=50K	1
<=50K	1
<=50K	1
<=50K	1

Entries	Errors
30162	5098
Success %	83,1

Success %	
<=50K	92,2
>50K	55,6

Tableau 20. Optimisation polynomiale partielle pour le fichier « Training ».

#	Class	P5	P1	P8	P13	P10	P6	P11	P12	P4	P9	P7	P14	P2	P3
Adult Test File															
1	<=50K	7	25	4	40	1	5	0	0	2	3	7	39	3	226802
2	<=50K	9	38	1	50	1	3	0	0	12	5	5	39	3	89814
6	<=50K	6	34	2	30	1	5	0	0	1	5	8	39	3	198693
9	<=50K	10	24	5	40	0	5	0	0	16	5	8	39	3	369667
10	<=50K	4	55	1	10	1	3	0	0	6	5	3	39	3	104996
12	<=50K	13	36	1	40	1	3	0	0	10	5	1	39	1	212465
13	<=50K	9	26	2	39	0	5	0	0	12	5	1	39	3	82091
17	<=50K	10	20	4	25	1	5	0	0	16	5	8	39	6	444554
18	<=50K	9	43	6	30	0	3	0	0	12	5	1	39	3	128354
19	<=50K	9	37	5	20	0	7	0	0	12	5	7	39	3	60548
22	<=50K	10	34	4	35	0	5	0	0	16	3	8	39	3	238588

Limit = 17,1815	
Coefficients	
a1 =	1,0000
a2 =	0,3844
a3 =	-0,0035
a4 =	-11,2469
a5 =	1,6760
a6 =	0,1158
a7 =	-0,0006
a8 =	2,1340
a9 =	0,0652
a10 =	0,0001
a11 =	7,7E-07

Combinaisons	Classifications	Success
2,405	<=50K	1
15,650	<=50K	1
4,661	<=50K	1
6,894	<=50K	1
8,487	<=50K	1
18,773	>50K	0
4,792	<=50K	1
3,109	<=50K	1
15,091	<=50K	1
6,664	<=50K	1
4,530	<=50K	1

Entries	Errors
15060	2589
% OK =	82,8

Success %	
<=50K	90,2
>50K	60,2

Tableau 21. Optimisation polynomiale partielle pour le fichier « Test ».

La démarche naïve proposée dans cette étude s'avère capable de conduire à des pourcentages de succès de classification similaires à ceux d'algorithmes sophistiqués à condition d'adopter une fonction discriminante adéquate.

C'est dans cet état d'esprit que la classification des iris de Fisher a été reconsidérée avec une fonction discriminante modifiée et une optimisation limitée au pouvoir discriminant entre *I. versicolor* et *I. virginica* (Lp = longueur des pétales, lp = largeur des pétales, Ls = longueur des sépales, ls = largeur des sépales) :

$$a1.Lp + a2.lp + a3.Ls + a4.ls + a5.(Ls)^2 + a6.(ls)^2 + a7.(Ls)^2.ls$$

Il ne rest alors plus qu'un seul iris mal classifié (cf. Tableau 22) !

Species	Coeff.	Comb.
setosa	a1 = 1,00	2,92
setosa	a2 = 1,66	4,76
setosa	a3 = 2,53	3,82
setosa	a4 = 3,44	4,36
setosa	a5 = -0,58	2,34
setosa	a6 = -1,59	1,85
setosa	a7 = 0,12	3,11
setosa		3,39
setosa		4,90
setosa		4,35
setosa		2,35
setosa		3,32
setosa		4,56

Mean setosa =	3,18
SD setosa =	1,46
Δ1 =	**4,36**
Mean versicolor =	9,86
SD versicolor =	0,46
Δ2 =	**2,93**
Mean virginica =	11,81
SD virginica =	0,48
Sum Δ =	**7,29**

Limit1 =	**6,69**
Limit2 =	**10,76**

Classifications	
setosa	1
setosa	1
setosa	1
setosa	1
setosa	1
setosa	1
setosa	1
setosa	1
setosa	1
setosa	1
setosa	1
setosa	1
setosa	1

Entries	Errors
150	1
Success %	99,3

Success %	
setosa	100,0
versicolor	98,0
virginica	100,0

Tableau 22. Optimisation polynomiale partielle pour les iris de Fisher.

Diagnostic du cancer du sein (Wisconsin Diagnostic Breast Cancer Database, cf. Annexe)

Le domaine médical regorge de publications relatives au diagnostic de maladies. Il s'agit ici d'évaluer la possibilité que l'imagerie d'une ponction de tumeur puisse conduire à une classification fiable comme tumeur bénigne ou maligne.

Actuellement, cette base donnée comporte 2 classes, 9 descripteurs et 699 entrées. Une fois les données manquantes éliminées, il reste 683 entrées qui ont été l'objet d'une extraction aléatoire de 150 entrées afin de constituer un fichier « Test ». Le training portera donc sur 533 entrées. Aucun descripteur n'est particulièrement discriminant ; ils ont tous un pouvoir discriminant inférieur à 2. Le Tableau 23 rapporte les résultats de classification pour les fichiers de training et de test.

Mean b = 1,33 1,41 1,29 2,10 2,96 1,23 1,33 2,11 1,07
SD b = 1,12 0,95 0,80 1,10 1,67 0,91 0,87 0,91 0,56
Discrim = 1,930 1,850 1,827 1,511 1,455 1,336 1,306 1,242 0,590
Mean m = 7,69 6,56 6,53 5,95 7,17 5,83 5,60 5,27 2,63
SD m = 3,10 2,61 2,75 2,30 2,36 3,32 3,15 2,37 2,57

#	Class	Class	P6	P3	P2	P7	P1	P8	P4	P5	P9
1	2	b	1	1	1	1	4	1	1	2	1
2	2	b	1	1	1	2	1	1	1	1	1
3	2	b	1	1	1	2	1	1	1	2	1

#	Class	Coefficients	Combinations
1	b	a1 = 1,0000	6
2	b	a2 = 0,2497	4
3	b	a3 = 0,5109	4
4	b	a4 = 0,1994	4
5	b	a5 = 0,8076	6
6	b	a6 = 0,4538	8
7	b	a7 = 0,2783	7
8	b	a8 = 0,2591	20
9	b	a9 = -0,0549	6

Mean b = 6,571
SD b = 2,658
Discrim = 3,1510
Mean m = 25,064
SD m = 5,232

Limit = 13,115

Classifications	
b	1
b	1
b	1
b	1
b	1
b	1
b	1
m	0
b	1

Entries	Errors
533	13
Success %	97,6

Success %	
b	97,1
m	98,4

#	Class	P6	P3	P2	P7	P1	P8	P4	P5	P9
Cancer Test File										
1	b	1	1	1	2	1	1	1	2	1
2	b	8	7	9	4	6	2	5	5	1
3	b	1	4	1	3	5	2	1	2	1
4	b	1	2	1	1	5	1	1	2	1
5	b	1	1	1	3	5	1	1	2	1
6	b	3	5	4	4	3	6	3	7	1
7	b	1	1	1	2	2	1	1	2	1
8	b	5	1	1	1	1	1	1	2	1
9	b	1	2	1	2	1	1	1	2	1

Limit = 13,1150 Coefficients	Combinaisons
a1 = 1,0000	4,162
a2 = 0,2497	23,529
a3 = 0,5109	8,795
a4 = 0,1994	7,443
a5 = 0,8076	7,592
a6 = 0,4538	14,829
a7 = 0,2783	4,970
a8 = 0,2591	7,963
a9 = -0,0549	4,412

Classifications	Success
b	1
m	0
b	1
b	1
b	1
m	0
b	1
b	1
b	1

Entries	Errors
150	3
% OK =	98,0

Success %	
b	96,8
m	100,0

Tableau 23. Classification (training + test) de la base de données de diagnostic du cancer du sein.

La double optimisation de « Training » permet de classifier correctement 97,6 % des tumeurs et la classification de l'échantillon « Test » conduit à un taux de réussite de 98,0 %. Ces deux valeurs se comparent avantageusement à celles de la littérature (cf. Annexe).

Conclusions

La méthode de classification proposée dans la présente étude s'appuie sur deux optimisations successives et une fonction discriminante, linéaire ou non. La première optimisation vise à maximiser le pouvoir discriminant (Δ) de la fonction discriminante tandis que l'optimisation des limites interclasses cherche à minimiser le nombre d'erreurs de classification d'une base de données de training.

Outre la base de données des iris sollicitée pour la présentation de la méthode, sept autres études du domaine des sciences de la vie et une base de données particulièrement populaire ont été analysées avec des résultats sensiblement équivalents à ceux de la littérature.

A deux occasions il a été également montré que le choix d'une fonction discriminante non linéaire peut conduire à une appréciable amélioration du succès de classification, ce qui donne à penser que les deux critères d'optimisation retenus (pouvoir discriminant et limite interclasse) pourraient trouver leur place dans des démarches plus sophistiquées.

Remerciements :

Nos remerciements vont aux concepteurs, dépositaires et gestionnaires des bases de données qui ont été mises à notre disposition, et, en particulier, au Centre belge de baguage BeBirds de l'Institut Royal des Sciences Naturelles de Belgique (SPP Politique scientifique), à ses responsables et à tous les bagueurs bénévoles qui collectent les données et participent au financement du système. Ils s'adressent aussi à deux lecteurs, dont les avis et conseils ont été très appréciés.

Bibliographie

Ambroise, V., Esposito, F., Scopece, G. & Tyteca, D. 2020. Can phenotypic selection on floral traits explain the presence of enigmatic intermediate individuals in sympatric populations of *Platanthera bifolia* and *P. chlorantha* (Orchidaceae)? Plant Species Biol. 35 : 59-71.

Baum A., & Baum H. 2017. *Platanthera muelleri* - eine dritte Art in der *Platanthera bifolia*/*chlorantha* Gruppe in Mitteleuropa. Journal EuropäIscher Orchideen 49 : 133-152.

Charytanowicz, M., Niewczas, J., Kulczycki, P., Kowalski, A., Łukasik, S. & Zak, S. 2010. Complete Gradient Clustering Algorithm for Features Analysis of X-ray Images Information Technologies. Pp 15-24 in E. Pietka & J. Kawa, eds., Biomedicine. Springer-Verlag, Berlin-Heidelberg.

Dua, D. & Karra Taniskidou, E. 2017. UCI Machine Learning Repository. Iris Data Set. Irvine, University of California, School of Information and Computer Science. https://archive.ics.uci.edu/ml/index.php

Durka W., Baum A., Michalski, S.G. & Baum, H. 2017. Darwin's legacy in *Platanthera*: are there more than two species in the *Platanthera* bifolia/chlorantha group? Plant Systematics and Evolution 303 : 419-431.

Esposito, F., Vereecken, N.J., Gammella, M., Rinaldi, R., Laurent, P. & Tyteca, D. 2018. Characterization of sympatric *Platanthera bifolia* and *Platanthera chlorantha* (Orchidaceae) populations with intermediate plants. PeerJ 6:e4256; DOI 10.7717/peerj.4256.

Fisher, R.A. 1936. The use of multiple measurements in taxonomic problems. Annual Eugenics 7, Part II : 179-188.

Johnson, B., Tateishi, R. & Xie, Z. 2012. Using geographically-weighted variables for image classification. Remote Sensing Letters 3, 6 : 491-499.

Le Boulengé, E. & Burnel, A. 2018. Identification biométrique du sexe chez les passereaux. Aves 55 : 63-97.

Lantz, B. 2019. Machine Learning with R. Troisième éd. Packt Publishing.

Maeck, M. 2018. La classification des iris de Fisher revisitée. Les Naturalistes belges 99,3 : 1-17.

Wikipédia, 2019. Base de données https://fr.wikipedia.org/wiki/Base_de_donn%C3%A9es

Zielesny, A. 2016. From Curve Fitting to Machine Learning. An Illustrative Guide to Scientific Data Analysis and Computational Intelligence. Springer International Publishing Switzerland.

ANNEXE

Les bases de données exploitées dans cette étude sont hébergées sur le site de l'Université de Californie.
Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science
https://archive.ics.uci.edu/ml/index.php

Les iris de Fischer (Iris Plants Database)
Updated Sept 21 by C.Blake - Added discrepency information

Sources : (a) Creator: R.A. Fisher
(b) Donor: Michael Marshall (MARSHALL%PLU@io.arc.nasa.gov)
(c) Date: July, 1988

Past Usage : Publications: too many to mention !!!
Fisher,R.A. "The use of multiple measurements in taxonomic problems"
Annual Eugenics, 7, Part II, 179-188 (1936)

Relevant Information :
This is perhaps the best known database to be found in the pattern recognition literature. Fisher's paper is a classic in the field and is referenced frequently to this day. (See Duda & Hart, for example.)
The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant. One class is linearly separable from the other 2; the latter are NOT linearly separable from each other.
Predicted attribute : class of iris plant. This is an exceedingly simple domain.
Number of Instances : 150 (50 in each of three classes).
Number of Attributes : 4 numeric, predictive attributes and the class
Attribute Information : 1. sepal length in cm
2. sepal width in cm
3. petal length in cm
4. petal width in cm
5. class : Iris Setosa, Iris Versicolour, Iris Virginica
Missing Attribute Values: None

Trois variétés de graines **(**Seeds Database**)**

Source :
MaÅ‚gorzata Charytanowicz, Jerzy Niewczas
Institute of Mathematics and Computer Science, The John Paul II Catholic University of Lublin, KonstantynÃ³w 1 H, PL 20-708 Lublin, Poland
e-mail: {mchmat,jniewczas}@kul.lublin.pl
Piotr Kulczycki, Piotr A. Kowalski, Szymon Lukasik, Slawomir Zak
Department of Automatic Control and Information Technology, Cracow University of Technology, Warszawska 24, PL 31-155 Cracow, Poland and Systems Research Institute, Polish Academy of Sciences, Newelska 6, PL 01-447 Warsaw, Poland
e-mail: {kulczycki,pakowal,slukasik,slzak}@ibspan.waw.pl

Data Set Information :
The examined group comprised kernels belonging to three different varieties of wheat: Kama, Rosa and Canadian, 70 elements each, randomly selected for the experiment. High quality visualization of the internal kernel structure was detected using a soft X-ray technique. It is non-destructive and considerably cheaper than other more sophisticated imaging techniques like scanning microscopy or laser technology. The images were recorded on 13x18 cm X-ray KODAK plates. Studies were conducted using combine harvested wheat grain originating from experimental fields, explored at the Institute of Agrophysics of the Polish Academy of Sciences in Lublin.
The data set can be used for the tasks of classification and cluster analysis.

Attribute Information :
To construct the data, seven geometric parameters of wheat kernels were measured:
1. area A,
2. perimeter P,
3. compactness C = 4*pi*A/P^2,
4. length of kernel,
5. width of kernel,

6. asymmetry coefficient
7. length of kernel groove.
All of these parameters were real-valued continuous.

Relevant Papers :

M. Charytanowicz, J. Niewczas, P. Kulczycki, P.A. Kowalski, S. Lukasik, S. Zak, 'A Complete Gradient Clustering Algorithm for Features Analysis of X-ray Images', in: Information Technologies in Biomedicine, Ewa Pietka, Jacek Kawa (eds.), Springer-Verlag, Berlin-Heidelberg, 2010, pp. 15-24.

Citation Request :

Contributors gratefully acknowledge support of their work by the Institute of Agrophysics of the Polish Academy of Sciences in Lublin.

Détermination de quatre types de forêts (Forest type mapping Database)

Source:

Brian Johnson
johnson '**@**' iges.or.jp
Institute for Global Environmental Strategies

Data Set Information:

This data set contains training and testing data from a remote sensing study which mapped different forest types based on their spectral characteristics at visible-to-near infrared wavelengths, using ASTER satellite imagery. The output (forest type map) can be used to identify and/or quantify the ecosystem services (e.g. carbon storage, erosion protection) provided by the forest.

Attribute Information:

Class: 's' ('Sugi' forest), 'h' ('Hinoki' forest), 'd' ('Mixed deciduous' forest), 'o' ('Other' non-forest land)
b1 - b9: ASTER image bands containing spectral information in the green, red, and near infrared wavelengths for three dates (Sept. 26, 2010; March 19, 2011; May 08, 2011.

Relevant Papers :

Johnson, B., Tateishi, R., Xie, Z., 2012. Using geographically-weighted variables for image classification. Remote Sensing Letters, 3 (6), 491-499.

Détermination de la comestibilité de champignons (Mushroom Database)

Sources :

(a) Mushroom records drawn from The Audubon Society Field Guide to North American Mushrooms (1981) G. H. Lincoff (Pres.), New York: Alfred A. Knopf
(b) Donor : Jeff Schlimmer (Jeffrey.Schlimmer@a.gp.cs.cmu.edu)
(c) Date: 27 April 1987

Past Usage :

1. Schlimmer,J.S. (1987). Concept Acquisition Through Representational Adjustment (Technical Report 87-19). Doctoral dissertation, Department of Information and Computer Science, University of California, Irvine. STAGGER: asymptoted to 95% classification accuracy after reviewing 1000 instances.
2. Iba,W., Wogulis,J., & Langley,P. (1988). Trading off Simplicity and Coverage in Incremental Concept Learning. In Proceedings of the 5th International Conference on Machine Learning, 73-79.
3. Ann Arbor, Michigan: Morgan Kaufmann. Approximately the same results with their HILLARY algorithm.

Relevant Information :

This data set includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family (pp. 500-525). Each species is identified as definitely edible, definitely poisonous, or of unknown edibility and not recommended. This latter class was combined with the poisonous one. The Guide clearly states that there is no simple rule for determining the edibility of a mushroom; no rule like ``leaflets three, let it be'' for Poisonous Oak and Ivy.
Number of Instances : 8124
Number of Attributes : 22 (all nominally valued)
Attribute Information: (classes: edible=e, poisonous=p)
1. cap-shape : bell=b, conical=c, convex=x, flat=f, knobbed=k, sunken=s
2. cap-surface : fibrous=f, grooves=g, scaly=y, smooth=s
3. cap-color : brown=n, buff=b, cinnamon=c, gray=g, green=r, pink=p, purple=u, red=e, white=w, yellow=y
4. bruises? : bruises=t, no=f
5. odor : almond=a, anise=l, creosote=c, fishy=y, foul=f, musty=m, none=n, pungent=p, spicy=s

6. gill-attachment : attached=a, descending=d, free=f, notched=n
7. gill-spacing : close=c, crowded=w, distant=d
8. gill-size : broad=b, narrow=n
9. gill-color : black=k, brown=n, buff=b, chocolate=h, gray=g, green=r, orange=o, pink=p, purple=u, red=e, white=w, yellow=y
10. stalk-shape : enlarging=e, tapering=t
11. stalk-root : bulbous=b, club=c, cup=u, equal=e, rhizomorphs=z, rooted=r, missing=?
12. stalk-surface-above-ring : fibrous=f, scaly=y, silky=k, smooth=s
13. stalk-surface-below-ring : fibrous=f, scaly=y, silky=k, smooth=s
14. stalk-color-above-ring : brown=n, buff=b, cinnamon=c, gray=g, orange=o, pink=p, red=e, white=w, yellow=y
15. stalk-color-below-ring : brown=n, buff=b, cinnamon=c, gray=g, orange=o, pink=p, red=e, white=w, yellow=y
16. veil-type : partial=p, universal=u
17. veil-color : brown=n, orange=o, white=w, yellow=y
18. ring-number : one=n, one=o, two=t
19. ring-type : cobwebby=c, evanescent=e, flaring=f, large=l, none=n, pendant=p, sheathing=s, zone=z
20. spore-print-color : black=k,brown=n, buff=b, chocolate=h, green=r, orange=o, purple=u, white=w, yellow=y
21. population : abundant=a, clustered=c, numerous=n, scattered=s, several=v, solitary=y
22. habitat : grasses=g,leaves=l,meadows=m,paths=p, urban=u,waste=w,woods=d

Missing Attribute Values : 2480 of them (denoted by "?"), all for attribute #11.

Class Distribution : edible: 4208 (51.8%)
poisonous: 3916 (48.2%)
total: 8124 instances

Prévision du salaire annuel (Adult Database)

Source :

Donor : Ronny Kohavi and Barry Becker
Data Mining and Visualization Silicon Graphics.
e-mail: ronnyk '@' live.com for questions.

Data Set Information :

Extraction was done by Barry Becker from the 1994 Census database. A set of reasonably clean records was extracted using the following conditions: ((AAGE>16) && (AGI>100) && (AFNLWGT>1)&& (HRSWK>0)). Prediction task is to determine whether a person makes over 50K a year.

Attribute Information:

Listing of attributes : Class : >50K, <=50K.

1. age: continuous
2. workclass : Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked
3. fnlwgt : continuous
4. education : Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool
5. education-num : continuous
6. marital-status : Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse
7. occupation : Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces
8. relationship : Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried
9. .race : White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black
10. sex : Female, Male
11. capital-gain : continuous
12. capital-loss : continuous
13. hours-per-week : continuous
14. native-country : United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands

Error Accuracy reported as follows, after removal of unknowns from train/test sets :

C4.5 : 84.46 ± 0.30 Naive-Bayes : 83.88 ± 0.30 NBTree : 85.90 ± 0.28

Following algorithms were later run with the following error rates, all after removal of unknowns and using the original train/test split. All these numbers are straight runs using MLC++ with default values.

Algorithm	Error	Algorithm	Error	Algorithm	Error
C4.5	15.54	C4.5-auto	14.46	C4.5 rules	14.94
Voted ID3 (0.6)	15.64	Voted ID3 (0.8)	16.47	T2	16.84
1R	19.54	NBTree	14.10	CN2	16.00
HOODG	14.82	FSS Naive Bayes	14.05	IDTM (Decision table)	14.46
Naive-Bayes	16.12	Nearest-neighbor (1)	21.42	Nearest-neighbor (3)	20.35
OC1	15.04	Pebls	Crashed. Unknown why (bounds WERE increased)		

Relevant Papers :

Ron Kohavi, "Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid", Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, 1996

Diagnostic d'un cancer du sein (Wisconsin Diagnostic Breast Cancer Database)

Source Information

a) Creators : Dr. William H. Wolberg, General Surgery Dept., University of Wisconsin, Clinical Sciences Center, Madison, WI 53792 - wolberg@eagle.surgery.wisc.edu
W. Nick Street, Computer Sciences Dept., University of Wisconsin, 1210 West Dayton St., Madison, WI 53706 - street@cs.wisc.edu 608-262-6619
Olvi L. Mangasarian, Computer Sciences Dept., University of Wisconsin, 1210 West Dayton St., Madison, WI 53706 - olvi@cs.wisc.edu

b) Donor : Nick Street

c) Date : November 1995

Relevant Information :

Number of Instances: 699 (as of 15 July 1992)

Number of Attributes: 10 plus the class attribute

Attribute Information: (class attribute has been moved to last column)

1. Sample code number	id number
2. Clump Thickness	1 - 10
3. Uniformity of Cell Size	1 - 10
4. Uniformity of Cell Shape	1 - 10
5. Marginal Adhesion	1 - 10
6. Single Epithelial Cell Size	1 - 10
7. Bare Nuclei	1 - 10
8. Bland Chromatin	1 - 10
9. Normal Nucleoli	1 - 10
10. Mitoses	1 - 10
11. Class:	(2 for benign, 4 for malignant)

Missing attribute values: 16

Accuracy

1. Two pairs of parallel hyperplanes were found to be consistent with 50% of the data.
   Accuracy on remaining 50% of dataset : 93.5%.
   Three pairs of parallel hyperplanes were found to be consistent with 67% of data.
   Accuracy on remaining 33% of dataset : 95.9%.
2. Zhang,~J. (1992). Selecting typical instances in instance-based learning.
   Size of data set: only 369 instances (at that point in time). Applied 4 instance-based learning algorithms.
   Collected classification results averaged over 10 trials.
   Best accuracy result : 1-nearest neighbor: 93.7%, trained on 200 instances, tested on the other 169.
3. Also of interest :
   Using only typical instances: 92.2% (storing only 23.1 instances) ; trained on 200 instances, tested on the other 169

Results

1. predicting field 2, diagnosis: B = benign, M = malignant
2. sets are linearly separable using all 30 input features
3. best predictive accuracy obtained using one separating plane in the 3-D space of Worst Area, Worst Smoothness and Mean Texture. Estimated accuracy 97.5% using repeated 10-fold crossvalidations. Classifier has correctly diagnosed 176 consecutive new patients as of November 1995.

Relevant information

Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. A few of the images can be found at :
http://www.cs.wisc.edu/~street/images/
Separating plane described above was obtained using Multisurface Method-Tree (MSM-T) [K. P. Bennett, "Decision Tree Construction Via Linear Programming." Proceedings of the 4th Midwest Artificial Intelligence and Cognitive Science Society, pp. 97-101, 1992], a classification method which uses linear programming to construct a decision tree. Relevant features were selected using an exhaustive search in the space of 1-4features and 1-3 separating planes.
The actual linear program used to obtain the separating plane in the 3-dimensional space is that described in :
[K. P. Bennett and O. L. Mangasarian: "Robust Linear Programming Discrimination of Two Linearly Inseparable Sets", Optimization Methods and Software 1, 1992, 23-34].
Number of instances : 569
Number of attributes : 32 (ID, diagnosis, 30 real-valued input features)
Attribute information :
1) ID number
2) Diagnosis (M = malignant, B = benign) (3-32)
Ten real-valued features are computed for each cell nucleus :

a) radius (mean of distances from center to points on the perimeter)
b) texture (standard deviation of gray-scale values)
c) perimeter
d) area
e) smoothness (local variation in radius lengths)
f) compactness (perimeter^2 / area - 1.0)
g) concavity (severity of concave portions of the contour)
h) concave points (number of concave portions of the contour)
i) symmetry
j) fractal dimension ("coastline approximation" - 1)

Several of the papers listed above contain detailed descriptions of how these features are computed.
The mean, standard error, and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius. All feature values are recoded with four significant digits.
Missing attribute values : none
Class distribution : 357 benign, 212 malignant