Apprentissage et classification monotone

Laura Nguyen

LFI

19 juillet 2018

Classification monotone

- Classifieur monotone :
  - Exploiter au mieux l'éventuelle monotonie de la classe par rapport aux valeurs d'attributs
  - Classifieur aussi monotone que possible et maintenant de bonnes performances
- Concepts sémantiques : préférence, priorité, importance...
- Applications :
  - Prédiction du risque de faillite
  - Evaluation du prix de biens immobiliers
  - Cote de crédit
- Pazzani et al. (2001) :
  - Etude auprès des experts (médecins)
  - Gain d'interprétabilité
  - Performances équivalentes

Exemple de problème de classification monotone

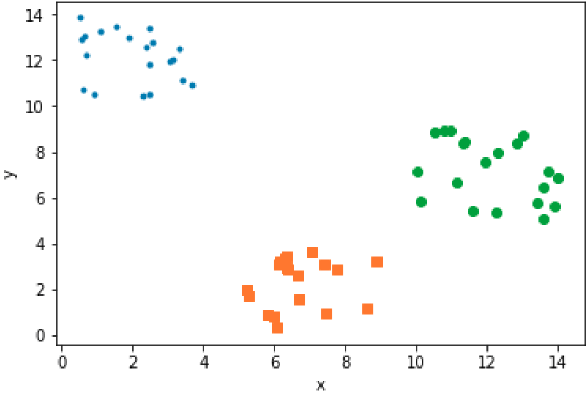
Potharst & Feelders (2002)

client	revenu	éducation	casier judiciaire	prêt
cl1	faible	faible	juste	non
cl2	faible	faible	excellent	faible
cl3	moyen	intermédiaire	excellent	intermédiaire
cl4	élevé	faible	excellent	élevé
cl5	élevé	intermédiaire	excellent	élevé

Formalisation

- Entrées :
  - n exemples : $\Omega = \{\omega_1, ..., \omega_n\}$
  - m attributs ordonnés : $A = \{a_1, ..., a_m\}$.
    Pour $j = 1, ..., m$:
    - $a_j : \Omega \rightarrow X_j$ avec X_j totalement ordonné
    - Espace de description $X = X_1 \times ... \times X_m$
  - k classes : $C = \{c_1, ..., c_k\}$
- Sortie :
  - Fonction de classification monotone $\lambda : X \rightarrow C$

Dataset artificiel à un attribut monotone



Classification par arbres de décision

- Pas d'hypothèse sur les données
- Exploiter au mieux l'éventuelle gradualité entre les valeurs d'attributs et la classe
- Construction :
  - Choisir a_j respectant le plus la contrainte de monotonie locale :

$$\forall \omega_i, \omega_h \in \Omega, a_j(\omega_i) \leq a_j(\omega_h) \Rightarrow \lambda(\omega_i) \leq \lambda(\omega_h)$$

  - Approche gloutonne : pas de garantie d'avoir un classifieur globalement monotone

Mesures de discrimination monotone

- Problème : incapacité des mesures de discrimination classiques à détecter la monotonie
- Chercher des mesures de discrimination sensibles à la monotonie et robustes au bruit non-monotone
  - Généralisation à rang de mesures classiques (Shannon, Gini)
  - Modèle de construction hiérarchique
  - Nouvelles mesures

Ensembles dominants

Pour $\omega_i \in \Omega$,

- Classes d'équivalence générées par :
  - un attribut a_j :

$$[\omega_i]_{a_j} = \{\omega_h \in \Omega : a_j(\omega_i) = a_j(\omega_h)\}$$

  - la classe λ :

$$[\omega_i]_\lambda = \{\omega_h \in \Omega : \lambda(\omega_i) = \lambda(\omega_h)\}$$

- Ensembles dominants générés par :
  - un attribut a_j :

$$[\omega_i]^{\leq}_{a_j} = \{\omega_h \in \Omega : a_j(\omega_i) \leq a_j(\omega_h)\}$$

  - la classe λ :

$$[\omega_i]^{\leq}_{\lambda} = \{\omega_h \in \Omega : \lambda(\omega_i) \leq \lambda(\omega_h)\}$$

Généralisation à rang de l'entropie de Shannon Hu et al. (2010)

- Entropie conditionnelle de Shannon :

$$H_s(\lambda|a_j) = \sum_{s=1}^{t_j} p_s \left(-\sum_{q=1}^{k} \frac{p_{q,s}}{p_s} \log \left(\frac{p_{q,s}}{p_s} \right) \right)$$
$$= \sum_{i=1}^{|\Omega|} \frac{1}{|\Omega|} \left(\log \left(\frac{|[\omega_i]_\lambda \cap [\omega_i]_{a_j}|}{|[\omega_i]_{a_j}|} \right) \right)$$

- Entropie de Shannon d'ordre :

$$H_s^*(\lambda|a_j) = \sum_{i=1}^{|\Omega|} \frac{1}{|\Omega|} \left(\log \left(\frac{|[\omega_i]_\lambda^{\leq} \cap [\omega_i]_{a_j}^{\leq}|}{|[\omega_i]_{a_j}^{\leq}|} \right) \right)$$

Modèle de construction hiérarchique de mesures de discrimination à rang Marsala & Petturiti (2015)

- Isoler les propriétés d'une telle mesure
- Créer de nouvelles mesures
- Structure fonctionnelle commune avec 3 fonctions
  Pour a_j fixé :
  - f^* : mesure de monotonie locale de l'objet
  - g^* : mesure d'agrégation de f^*
  - h^* : agrégation de mesures g^*

  Ecriture générique :

$$H^*(\lambda|a_j) = h^*(g^*(f^*(\omega_1)), ..., g^*(f^*(\omega_n)))$$

Couche f^*

Pour $a_j \in A$ fixé,

- $dsr(\omega_i) = \dfrac{|[\omega_i]^{\leq}_{\lambda} \cap [\omega_i]^{\leq}_{a_j}|}{|[\omega_i]^{\leq}_{a_j}|}$
- $mindsr(\omega_i) = \dfrac{min_{\omega_h \in [\omega_i]_{a_j}} |[\omega_h]^{\leq}_{\lambda} \cap [\omega_h]^{\leq}_{a_j}|}{|[\omega_i]^{\leq}_{a_j}|}$
- $maxdsr(\omega_i) = \dfrac{max_{\omega_h \in [\omega_i]_{a_j}} |[\omega_h]^{\leq}_{\lambda} \cap [\omega_h]^{\leq}_{a_j}|}{|[\omega_i]^{\leq}_{a_j}|}$
- $avgdsr(\omega_i) = \dfrac{\displaystyle\sum_{\omega_h \in [\omega_i]_{a_j}} \dfrac{|[\omega_h]^{\leq}_{\lambda} \cap [\omega_h]^{\leq}_{a_j}|}{|[\omega_i]_{a_j}|}}{|[\omega_i]^{\leq}_{a_j}|}$

Réécriture des mesures de discrimination à rang

$$H_S^*(\lambda|a_j) = \sum_{i=1}^{n} \frac{1}{n} \left(-\log\left(dsr\left(\omega_i\right)\right)\right)$$

- $f_S^* : dsr(\omega_i)$
- $g_S^* : -\log(f_S^*(\omega_i))$
- $h_S^* : \sum_{i=1}^{n} \frac{1}{n} g_S^*(f_S^*(\omega_i))$

$$H_P^*(\lambda|a_j) = \sum_{i=1}^{n} \frac{1}{n} \left(-\frac{\log\left(mindsr\left(\omega_i\right)\right)}{mindsr\left(\omega_i\right)}\right)$$

- $f_P^* : mindsr(\omega_i)$
- $g_P^* : -\frac{\log f_P^*(\omega_i)}{f_P^*(\omega_i)}$
- $h_P^* : \sum_{i=1}^{n} \frac{1}{n} g_P^*(f_P^*(\omega_i))$

Construction d'arbres de décision monotones Marsala & Petturiti (2015)

- Classifieur RDMT(H) paramétré par :
  - Une mesure de discrimination H
  - 3 critères de pré-élagage : partitionnement, arrêt, étiquetage
- Critère de partitionnement (splitting rule)
  - Chaque attribut a_j est binarisé : $\forall x \in X_j$,

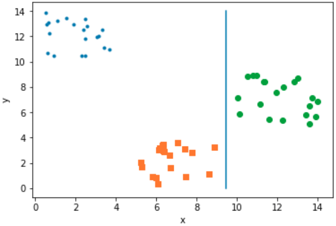
$$a_j^x(\omega_i) = \begin{cases} 0 & \text{si } a_j(\omega_i) \leq x \\ 1 & \text{sinon} \end{cases}$$

  - Trouver $a_*^{x_*}$ minimisant $H(\lambda | a_j^x)$
    - ⋆ a_* est l'attribut utilisé pour le partitionnement
    - ⋆ x_* est la valeur seuil trouvée par discrétisation :

$$x_* = arg\ min\{H(\lambda|a_j^x)|j = 1, ..., m, x \in X_j\}$$

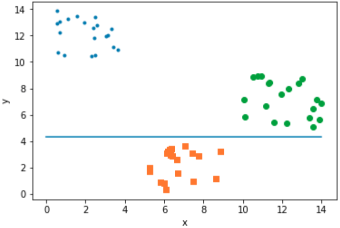
Critère de partitionnement

Seuil de coupure généré par H_S^* :



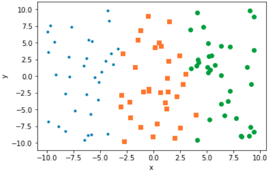
$$H_S^*(\lambda|x) = 0.19 \leq H_S^*(\lambda|y) = 0.53$$

Seuil de coupure généré par H_S :

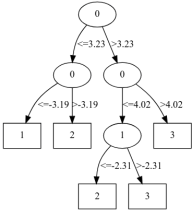


$$H_S(\lambda|x) = H_S(\lambda|y) = 0.67$$

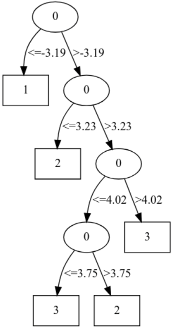
Arbres de décision générés par H_S^* et H_S



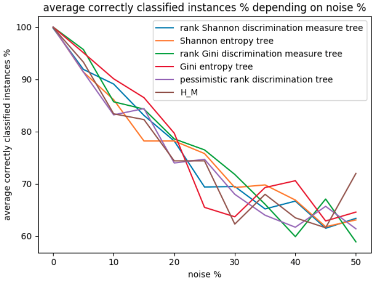
Arbre généré par H_S^*

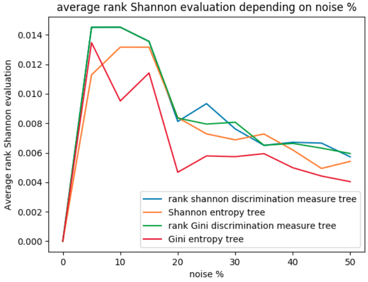


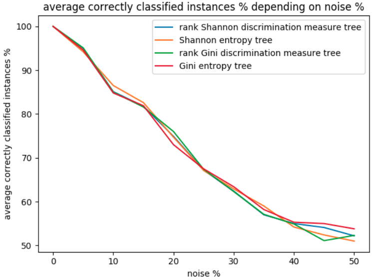
Arbre généré par H_S

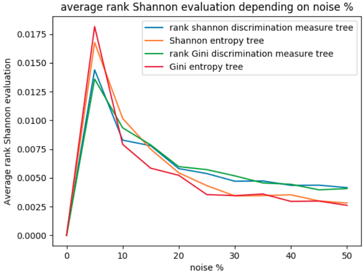


Expérimentations sur des datasets jouets

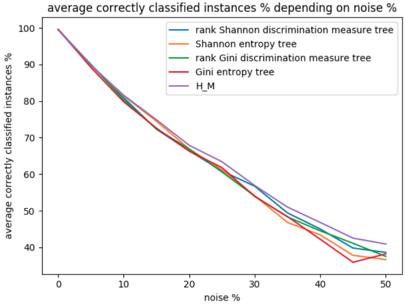


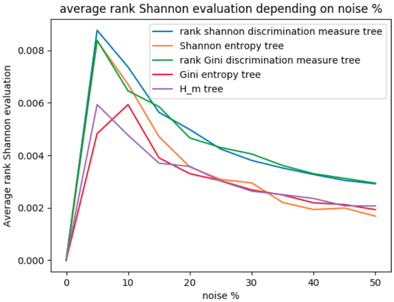




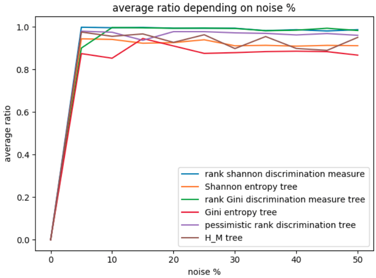


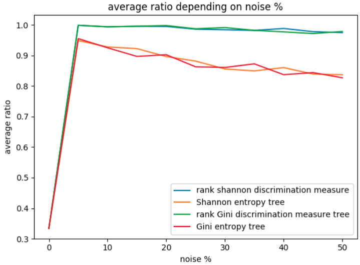
Expérimentations sur des datasets jouets

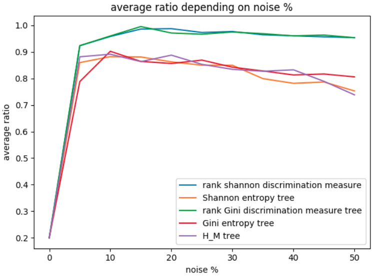




Expérimentations sur des datasets jouets : ratio du nombre de paires d'éléments non monotones







Expérimentations sur des datasets réels

- CPU (classification) : 8 classes, 9 attributs, 209 instances

	H_S^*	H_S	H_Q^*
CCI	68.40 % $\pm$ 0.04 %	65.06 % $\pm$ 0.03 %	67.93 % $\pm$ 0.03 %
avgDepth	10.50 $\pm$ 0.50	9.00 $\pm$ 0.00	11.00 $\pm$ 1.00
avgLeaves	35.00 $\pm$ 2.00	27.50 $\pm$ 3.50	40.50 $\pm$ 3.50
avgRatio	77.99 % $\pm$ 0.11 %	86.00 % $\pm$ 2.83 %	50.64 % $\pm$ 16.03 %
avgNbPairs	9.50 $\pm$ 0.50	7.50 $\pm$ 1.50	12.50 $\pm$ 1.50
avgEval	0.02 $\pm$ 0.00	0.04 $\pm$ 0.02	0.01 $\pm$ 0.00
avgPairRatio	0.28 $\pm$ 0.00	0.28 $\pm$ 0.02	0.32 $\pm$ 0.01
avgNbExamples	27.00 $\pm$ 0.00	28.00 $\pm$ 3.00	55.50 $\pm$ 16.50

- Airfoil Self-Noise (régression) : 5 attributs, 1503 instances, discrétisation en 5 classes

	H_S^*	H_S	H_G^*	H_G
CCI	39.25% $\pm$ 0.10 %	33.00% $\pm$ 0.12%	35.14 % $\pm$ 0.07%	32.33 % $\pm$ 0.11
avgDepth	20.5 $\pm$ 1.66	14.75$\pm$1.09	18.25$\pm$1.48	15.5 $\pm$2.69
avgLeaves	158.75 $\pm$ 5.12	140.5 $\pm$ 1.66	164.25 $\pm$ 3.83	143.75 $\pm$ 5.72
avgRatio	74.82 % $\pm$ 0.03%	80.74 % $\pm$ 0.04%	75.60 % $\pm$ 0.04%	79.11 % $\pm$ 0.04%
avgNbPairs	44.5 $\pm$ 2.60	48.25 $\pm$ 1.92	45.75 $\pm$ 3.03	47.25 $\pm$ 3.63

Conclusion

Hu, Q., Guo, M., Yu, D. & Liu, J. (2010). Information entropy for ordinal classification. In *Science China Information Sciences* 53.6, p. 1188-1200.

Marsala, C. & Petturiti, D. (2015). Rank discrimination measures for enforcing monotonicity in decision tree induction. In *Information Sciences* 291, p. 143-171.

Pazzani, M. J., Mani, S. & Shankle, W. R. (2001). Acceptance of rules generated by machine learning among medical experts. In *Methods of information in medicine* 40.05, p. 380-385.

Potharst, R. & Feelders, A. J. (2002). Classification trees for problems with monotonicity constraints. In *ACM SIGKDD Explorations Newsletter* 4.1, p. 1-10.

Apprentissage et classification monotone

Laura Nguyen

LFI

19 juillet 2018

Ensembles dominants : exemple

Soit le dataset suivant :

Ω	a_1	a_2	λ
ω_1	1	5	1
ω_2	1	7	1
ω_3	2	8	1
ω_4	5	9	2
ω_5	5	7	2
ω_6	4	6	2

- L'ensemble dominant généré par a_1 pour ω_3 est :

$$[\omega_3]^{\leq}_{a_1} = \{\omega_3, \omega_4, \omega_5, \omega_6\}$$

- L'ensemble dominant généré par λ pour ω_3 est :

$$[\omega_3]^{\leq}_{\lambda} = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6\}$$