le cnam

Bases de données multimédia

Recherche avec contrôle de pertinence
Recherche pluri-modale et trans-modale
Catégorisation de bases d'images

Michel Crucianu (CNAM)
http://cedric.cnam.fr/~crucianm/bdm.html

le cnam

Contenu de la séance

- Recherche itérative avec contrôle de pertinence
  - Critères de sélection et apprentissage actif
  - Évaluation du contrôle de pertinence et exemples
  - Structures d'index pour la recherche itérative
- Recherche pluri-modale et trans-modale
  - Représentation des données textuelles
  - Recherche pluri-modale
  - Recherche trans-modale
- Catégorisation de bases d'images (compléments)
  - Classifications complémentaires
  - Classification semi-supervisée

le cnam

Contrôle de pertinence : motivation

- Fossé sémantique (*semantic gap*) entre
  - descripteurs du contenu visuel extraits automatiquement des images

  et
  - critères de recherche pertinents pour les utilisateurs
- L'association automatique entre les deux est à ce jour possible uniquement dans des domaines très restreints
- Bon moyen d'identifier la cible d'un utilisateur lors d'une session de recherche : inclure l'utilisateur dans la boucle

→ Recherche itérative avec contrôle de pertinence (*relevance feedback*, voir aussi [CFB04])

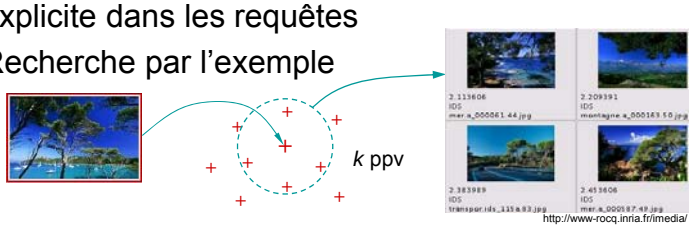
- Peut être utilisé pour la recherche de textes [Sal68], musique, etc. où un fossé sémantique est également présent

le cnam

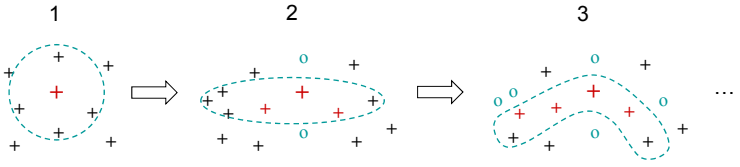
Contrôle de pertinence : objectif

- Objectif : trouver des images sans employer de critère explicite dans les requêtes

1. Recherche par l'exemple



2. Recherche itérative avec contrôle de pertinence



le cnam

Contrôle de pertinence : exemple

Objectif : retrouver des portraits

Base de 7500 images, dont 110 portraits

Disponible : description globale (couleur, texture, forme)

Première page de résultats après 4 itérations →

[FCB04/1]

le cnam

Composantes du mécanisme

1. *Learner* : à partir de l'information disponible (notamment des exemples positifs et/ou négatifs), estimer l'ensemble d'images visé
2. Sélecteur : à partir de l'estimation produite par le *learner*, choisir les images que l'utilisateur doit marquer lors de l'itération suivante
3. Utilisateur : fournir à chaque itération le retour pour les images choisies par le sélecteur
   - Les évaluations sont souvent faites à l'aide d'une vérité terrain, en émulant l'utilisateur

le cnam

Difficultés pour l'apprentissage

- Très peu d'exemples étiquetés : leur nombre est souvent inférieur au nombre de dimensions de l'espace de description !
- Déséquilibre important entre le nombre d'exemples positifs et le nombre d'exemples négatifs
- Forme potentiellement complexe de l'ensemble d'images visé, qui peut même présenter plusieurs modes distants dans l'espace de description
- L'interactivité exige un temps de réponse très court, à la fois pour le *learner* et pour le sélecteur

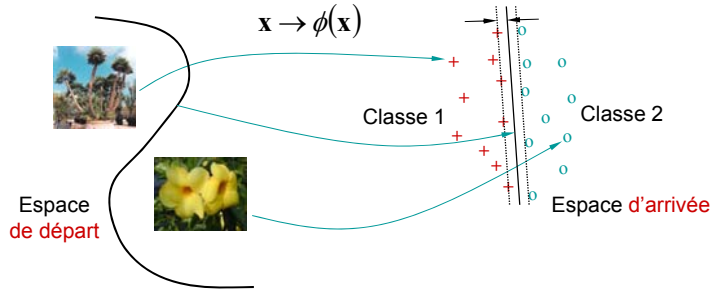
le cnam

Learners employés

- Machines à vecteurs support (*support vector machines*, SVM, [SS02])
  - Avantages dans le contexte du retour de pertinence
    - La fonction de décision associée permet à la fois la définition d'une frontière et le classement des images
    - Avec un large choix des noyaux, les SVM permettent une grande liberté dans la forme des classes (avec un contrôle par la régularisation)
    - D'autres sources d'information (en dehors des exemples) permettent de définir des noyaux appropriés (*kernel engineering*)
    - Apprentissage très rapide avec le nombre relativement limité d'exemples fournis par le contrôle de pertinence
    - Moindre sensibilité au déséquilibre entre exemples positifs et négatifs

le cnam

SVM et discrimination d'images



le cnam

Sélecteur : objectifs et critères

- Objectifs
  1. Retourner un maximum d'images pertinentes à l'utilisateur
  2. Maximiser le transfert d'information utilisateur → système

  → Une même méthode ne pouvant pas répondre aux deux objectifs, on peut se servir de 2 sélecteurs (un par objectif)
- Critères de sélection
  - « *Most positive* » (MP) : retourner les images les plus pertinentes suivant l'estimation actuelle faite par le *learner* (se focalise sur l'objectif 1) – critère classique le plus utilisé
  - « *Most informative* » (MI) : retourner les images qui permettent à l'utilisateur de fournir un maximum d'information sur sa cible (se focalise sur l'objectif 2) → minimiser le nombre d'exemples
  - Des critères mixtes sont possibles

le cnam

Contrôle de pertinence : exemple (2)

Objectif : retrouver des régions représentant des villages

Base avec 24000 régions, dont 87 dans la classe

Disponible : description des régions (couleur, texture, formes à l'intérieur)

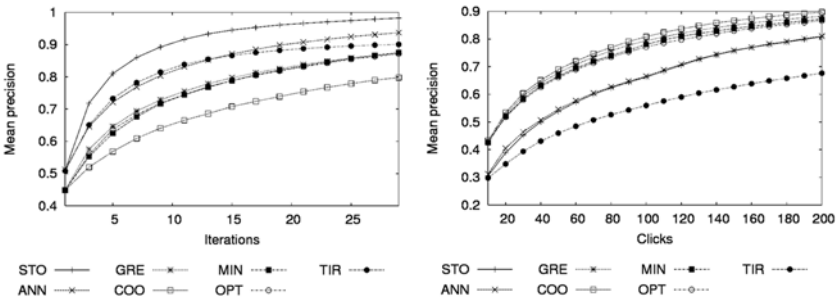
Première page de résultats avec 6 exemples positifs et 28 négatifs

[FCB04/2]

le cnam

Évaluation : rapidité de convergence

- Évolution de la précision moyenne avec les itérations (gauche) ou les *clicks* (droite ; *click* = image marquée)

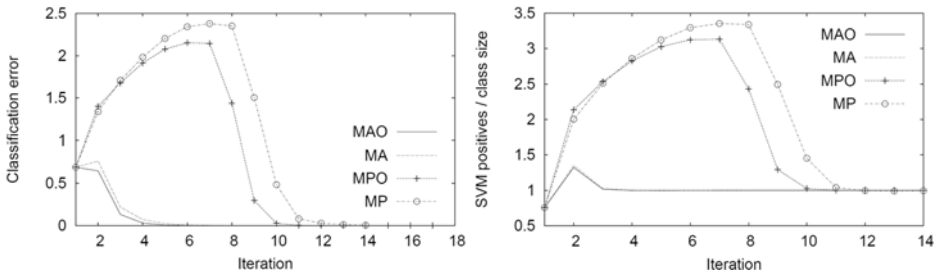


[CTF05]

le cnam

Évaluation : rapidité de convergence

- Évolution de l'erreur de classification (gauche) et de la taille estimée de la classe pertinente (droite)

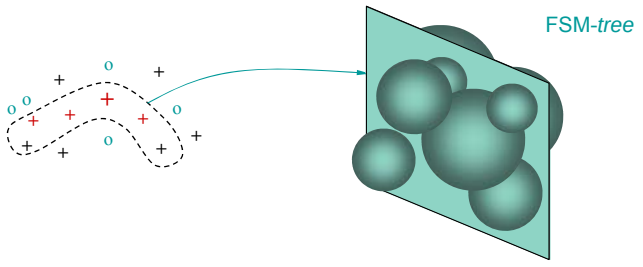


[FCB04/1]

le cnam

Passage à l'échelle : FSM-*tree* et requête hyperplan

- Principe de la méthode de [CEOT08] :
  - Tenir compte des exemples positifs et des exemples négatifs
  - Modéliser la discrimination par un 2-*class* SVM
  - Construire un M-*tree* dans l'espace d'arrivée (FSM-*tree*)
  - Retourner les images les plus proches de la frontière, trouvées par une requête *k*NN dans le FSM-*tree* avec un hyperplan

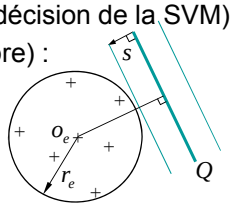


le cnam

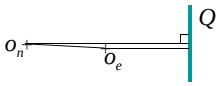
FSM-*tree* et requête hyperplan (2)

- Distance à l'hyperplan : $d(Q,o_e)=\frac{\left|\sum_i \alpha_i\, y_i\, K(o_e,x_i)+b\right|}{\sqrt{\sum_{i,j}\alpha_i\,\alpha_j\, y_i\, y_j\, K(x_i,x_j)}}$

  ($f(o_e)=\sum_i \alpha_i\, y_i\, K(o_e,x_i)+b$ étant la fonction de décision de la SVM)
- Principe d'élagage (test de rejet d'un sous-arbre) :
  - Si $d(Q,o_e)>r_e+s$ alors le nœud n'est pas conservé pour exploration ultérieure



- Comment éviter plus de calculs de distances :



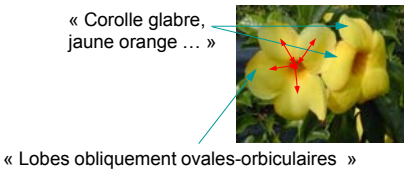
$d(Q,o_e)+d(o_e,o_n)\geq d(Q,o_n)$ mais $d(Q,o_e)+d(Q,o_n)\not\geq d(o_e,o_n)$

$\Rightarrow d(Q,o_e)\geq d(Q,o_n)-d(o_e,o_n)$ ($d(Q,o_e)\not\geq |d(Q,o_n)-d(o_e,o_n)|$), donc si $d(Q,o_n)-d(o_e,o_n)>r_e+s$ alors $d(Q,o_e)>r_e+s$

le cnam

Pourquoi plusieurs modalités ?

- Le contenu d'une base multimédia peut être caractérisé par des données de nature différente
  - Descripteurs du contenu visuel
  - Méta-données textuelles structurées (exemple : nom photographe, date, paramètres de la prise de vue)
  - Méta-données textuelles non structurées ou semi-structurées : liste de mots-clés, textes descriptifs
  - Graphes décrivant des relations spatiales entre composantes



le cnam

Pluri-modale ou trans-modale ?

- Pluri-modale : exploiter conjointement les informations provenant de plusieurs modalités pour améliorer la recherche
  - Contenu visuel et contenu textuel sont en partie redondants et en partie complémentaires
  - Nécessaire de décrire chaque type de contenu et d'en tenir compte pour sélectionner ou classer les résultats, par exemple à travers
    - Fusion précoce : concaténation des deux descriptions dans un vecteur unique
    - Fusion tardive : classer les résultats par rapport à chaque type de description, ensuite fusionner ces classements
- *Trans*-modale : requête dans une modalité, résultats dans l'autre !
  - Image → texte : annotation d'images (mots-clés), génération de description textuelle (phrases bien formées)
  - Texte → image : illustration de texte ou (d'ensembles) de mots-clés
  - Passe en général par la recherche d'une représentation « commune » entre modalités

le cnam

Types de mots-clés

1. Désignent des objets identifiables dans l'image ou des caractéristiques visuelles de ces objets
   - La détection et/ou la reconnaissance d'objets permet dans certains cas de les obtenir
2. Désignent des caractéristiques visuelles de la scène
   - La reconnaissance de scènes permet dans certains cas de les obtenir
3. Concernent l'interprétation de la scène (qu'un humain peut inférer)
   - Très difficiles à obtenir automatiquement à partir des images
   - Complètent le contenu visuel des images
4. Mettent l'image dans un contexte qui ne peut pas être inféré par un non spécialiste à partir du contenu visuel de l'image
   - Ne peuvent donc pas être obtenus automatiquement
   - Complètent le contenu visuel de l'image

le cnam

Image et mots-clés : exemple

© http://la-revolution-des-oeillets.ifrance.com/

1. Char (/ tank), voiture, immeuble, personnes
2. Extérieur, jour, ville
3. Fraternisation
4. Lisbonne, révolution des œillets, 25 avril 1974 , Caetano, Salazar

le cnam

Prise en compte des mots

- Objectif : mesurer la similarité entre les ensembles de mots associés à 2 images
- Sources
  1. Ensembles de mots-clés : mots pertinents mais peu nombreux
     - Exemple : « *sunset*, *seascape*, *sailboat* »
  2. Textes : nombreux mots, lesquels sont pertinents ?
     - Exemple : « Tôt le 25 avril 1974, au Portugal, des capitaines en rupture avec le système de Salazar se révoltent et prennent le pouvoir. La voix calme d'un mystérieux « Commandement du Mouvement des Forces armées» transmise par les radios de Lisbonne, Renascenta et Radio Clube donnant le signal de la révolte aux capitaines mutins, exhorte les gens à rester chez eux et à garder leur calme. »

le cnam

Prise en compte des mots (2)

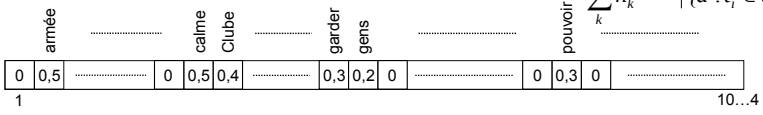
- Sous quelle forme les prendre en compte ?
  1. Mots : forme directement disponible, mais relation complexe avec la signification (problèmes : homonymie, synonymie, etc.)
  2. Concepts : signification claire, mais le plus souvent forme non directement disponible (mot → concept : désambiguïsation…)
- Mesurer la similarité entre 2 ensembles de mots :
  1. Compter les cooccurrences de mots entre les 2 ensembles ; simple, mais traite les mots de façon égalitaire et ne tient pas compte d'éventuelles relations entre mots différents
  2. Relations issues d'analyse statistique : pondération différenciée des mots selon leur potentiel discriminatif, prise en compte des cooccurrences dans l'ensemble du corpus pour identifier une sémantique « latente » (⇒ désambiguïsation implicite)
  3. Relations conceptuelles : partie ou ensemble des relations issues d'une ontologie (*sort_of*, *part_of*, proximité conceptuelle, etc.)

le cnam

Modèle vectoriel d'un texte

- Extraction des termes par l'élimination des mots peu spécifiques (conjonctions, prépositions, verbes auxiliaires, etc.) et lemmatisation

  « Tôt le 25 avril 1974, au Portugal, des capitaines en rupture avec le système de Salazar se révoltent et prennent le pouvoir. La voix calme d'un mystérieux « Commandement du Mouvement des Forces armées» transmise par les radios de Lisbonne, Renascenta et Radio Clube donnant le signal de la révolte aux capitaines mutins, exhorte les gens à rester chez eux et à garder leur calme. »
- Pour un ensemble donné de textes, association d'une dimension à chaque terme et représentation de chaque texte par le vecteur (de très grande dimension) correspondant aux termes qu'il contient [Sal68]
- Pondération des termes : nombreuses propositions, comme *tf*idf* (*term frequency * inverse document frequency* [SMG83]) : $\frac{n_i}{\sum_k n_k} \log \frac{|D|}{|\{d : t_i \in d\}|}$



le cnam

Modèle vectoriel d'un texte (2)

- Mesurer la dissimilarité : cosinus de l'angle, distance L1 ou L2,…
- Quelques insuffisances du modèle vectoriel de base
  - Les relations entre mots (synonymie, antonymie, hyperonymie) ne sont pas représentées
    - Par exemple, si les termes « voiture », « automobile » et « véhicule » sont rencontrés dans les textes, 1 dimension différente est affectée à chacun
  - L'ambiguïté (pas uniquement due à l'homonymie) n'est pas traitée
  - Grand nombre de mots peu fréquents, qui ne sont pas pertinents pour autant → présence de « bruit »
- L'impact de ces insuffisances augmente quand le nombre de termes par texte diminue, or les ensembles de mots clés associés aux images sont en général de faible cardinalité !

le cnam

Relations statistiques

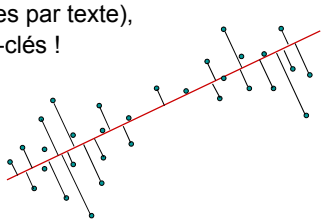
(analyse sémantique latente, *latent semantic analysis*, LSA, [DDF90])

- Analyse en composantes principales (ACP) sur l'ensemble des vecteurs qui représentent les textes d'un corpus (modèle vectoriel), projection sur les composantes de plus forte variance

→ Éliminer le « bruit » qui domine les composantes de faible variance

→ Mettre en évidence des « composantes sémantiques cachées » (résultant de cooccurrences) qui représentent des relations statistiquement déterminées entre mots et à traiter partiellement l'ambiguïté

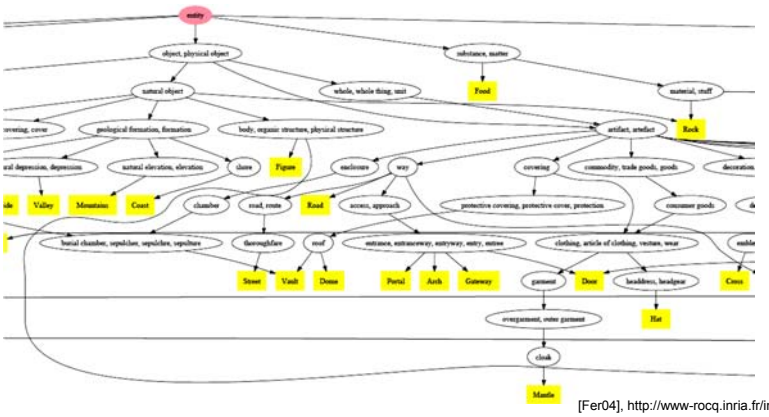
- Condition : les caractéristiques pertinentes doivent être bien représentées (→ beaucoup de textes, beaucoup de termes par texte), peu probable avec des ensembles de mots-clés !



le cnam

Relations conceptuelles

- Relations issues d'une ontologie (*a priori*, non liées à un corpus)
- Fragment de WordNet

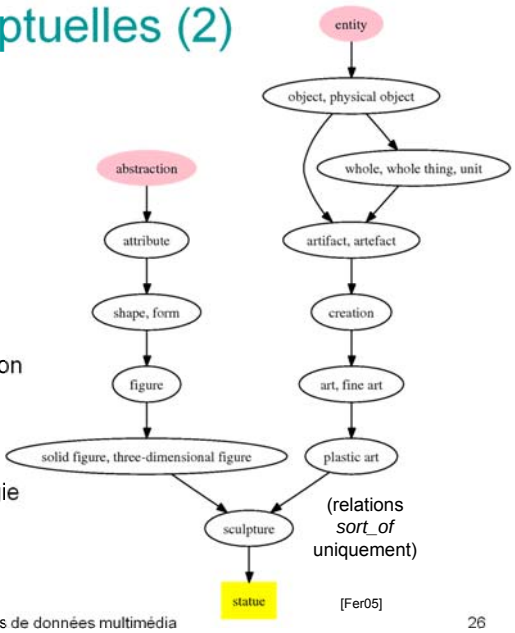


[Fer04], http://www-rocq.inria.fr/imedia

le cnam

Relations conceptuelles (2)

- Avantages
  - Fiables quel que soit le volume du corpus
  - Relative complétude par rapport aux relations possibles
- Désavantages
  - Étape de désambiguïsation préalable
  - Non prise en compte des spécificités du corpus
  - Insuffisances de l'ontologie



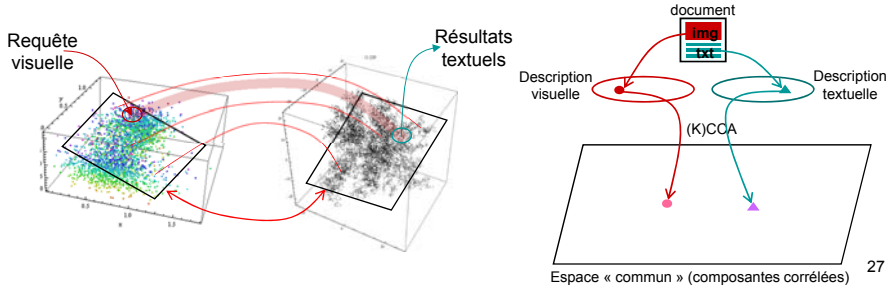
(relations *sort_of* uniquement)

[Fer05]

le cnam

Recherche trans-modale

- Représentation « commune » obtenue par analyse canoniques des corrélations (CCA) ou sa version à noyaux (KCCA [HSS04])
  1. Images représentées par descripteurs visuels
  2. Textes (ou tags) représentés par descripteurs textuels
  3. (K)CCA trouve les sous-espaces les mieux corrélés → espace « commun » (voir par ex. [CCD14]

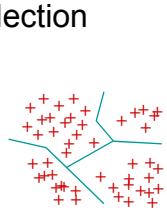


1/29/2016 M. Crucianu 27

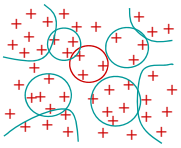
le cnam

Catégorisation du contenu des bases

- Rendre explicite une structure qui peut être mise en rapport avec des critères de recherche des utilisateurs
- Partitionnement de la collection
  - Résumés globaux, *a priori*
    - Quelle granularité ?
    - Quelle pertinence ?
    - Niveau sémantique ?



  - Vues locales, contextuelles (problème peu abordé)
    - Structurer les alternatives à chaque étape de la recherche
    - Niveau sémantique ?



29 janvier 2016 Bases de données multimédia 28

le cnam

Classifications complémentaires

- BD relationnelles
  - 1 table (relation)
    - définie par un ensemble d'attributs (attribut = ensemble de valeurs)
    - contient des enregistrements (*n*uplets)
  - 1 enregistrement : 1 valeur pour chaque attribut

  → La structure est explicite, permettant de définir des critères d'interrogation (valeurs d'un ou plusieurs attributs)
- Base d'images (hors possibles méta-données structurées) = ensemble d'images

  → Comment rendre explicite une structure pertinente ?

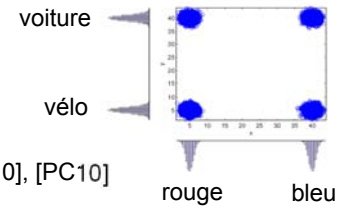
  → Peut-on extraire automatiquement des variables catégorielles comme équivalent des attributs dans les BD relationnelles ?

le cnam

Classifications complémentaires (2)

- Exemple imaginaire : images de voitures bleues ou rouges et de vélos bleus ou rouges, décrite chacune par un descripteur vectoriel composite

- Pouvons-nous trouver automatiquement deux classifications complémentaires des images (une par type d'objet et une par couleur) ?



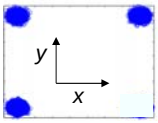
- Méthodes : [CFD07], [JMD08], [DB10], [PC10]

le cnam

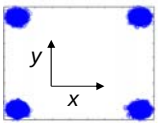
Une méthode : TCA + classification

- *Tree-Component* Analysis (TCA, analyse en sous-espaces indépendants) + classification automatique (voir [PC10])
- TCA : étant donné un ensemble de données dans un espace vectoriel, on cherche des sous-espaces tels que
  - Dans un même sous-espace, les variables sont dépendantes
  - Entre deux sous-espaces, les variables sont indépendantes
  - Exemples

Variables x,y dépendantes



Variables x,y indépendantes



- Adaptations : calcul de l'information mutuelle entre classifications, ajout d'un critère de qualité de la classification
  - → Pas nécessaire de connaître le nombre de sous-espaces ou de groupes dans chaque sous-espace
  - → Pour N vecteurs de dimension d, complexité $O(d^3 \times N)$

29 janvier 2016 Bases de données avancées 2 31

le cnam

Exemple 2 : les données

- 21 classes extraites de COIL100 (1 classe = 100 photos d'un même objet, chaque photo prise sous un angle différent)

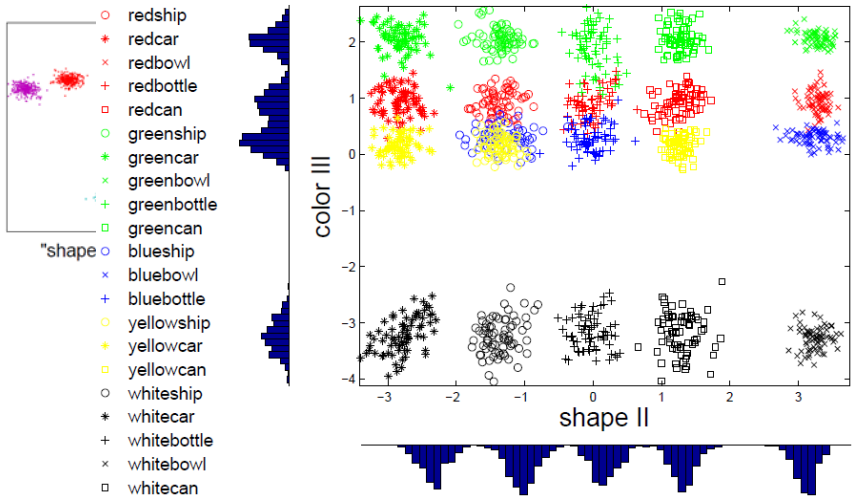
Variable « forme » (5 « valeurs »)

Variable « couleur » (5 « valeurs »)

29 janvier 2016 Bases de données avancées 2 32

le cnam

Exemple 2 : les résultats



29 janvier 2016 Bases de données avancées 2 33

le cnam

Classification semi-supervisée

1. Similarités issues des données décrivant les images (descripteurs visuels, mots-clés, relations spatiales…)
   - Directement calculables là où les données correspondantes sont disponibles
   - Nombreuses et bon marché
2. Similarités fournies par les utilisateurs
   - Explicitement ou à travers des corrélations de recherche, ...
   - Étiquettes de classe, contraintes, …
   - Caractère approximatif, domaine partiel de définition
   - Plus rares et coûteuses

→ Combiner ces sources d'information

29 janvier 2016 Bases de données avancées 2 34

le cnam

Classification semi-supervisée (2)

- La supervision (information concernant la cible : étiquettes de classe, contraintes, …) n'est disponible que pour une (faible) partie des données

→ apprentissage semi-supervisé
  - Question : les différentes sources sont-elles cohérentes ?

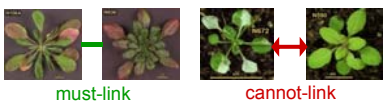
- Coût élevé de
  - l'acquisition de la supervision (exige en général de l'interaction)
  - l'utilisation des données (complexité algorithmique)

→ apprentissage actif : sélection par l'algorithme des données pour lesquelles la supervision est demandée
  - → Amélioration maximale des résultats avec un coût minimal
  - Question : dans quelles conditions cela mène à une déception ?

le cnam

Exemple : similarités plus contraintes

- Situation visée (catégorisation de bases d'images) :
  - Grande base inconnue (ou peu connue)
  - Classification simple : résultats médiocres
    - → Supervision nécessaire !
  - Classes d'images inconnues *a priori*
    - → L'utilisateur ne peut pas donner d'étiquettes mais est capable de dire si 2 images devraient être dans une même classe (contrainte *must-link*) ou dans des classes différentes (contrainte *cannot-link*)
    - → Étant donnée la taille de la base, la quantité de supervision (contraintes) doit être minimale

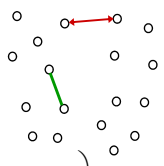


le cnam

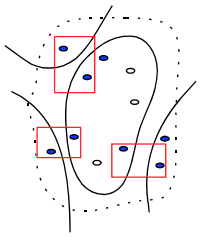
AFCC : principe de la méthode [GCB08]

- Aspect semi-supervisé
  - Combiner deux sources d'information :
    1. Similarités entre descripteurs d'images
    2. Contraintes binaires disponibles
  - → Nouvelle fonction à minimiser (basée sur CA) :

$$J(\mathbf{M}) = J_{\mathrm{CA}}(\mathbf{M}) + \alpha \left(\sum_{(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{M}} \sum_{p=1}^{k} \sum_{l=1, l \neq p}^{k} u_{ip}\, u_{jl} \; + \sum_{(\mathbf{x}_i, \mathbf{x}_j) \in C} \sum_{p=1}^{k} u_{ip}\, u_{jp} \right)$$



- Aspect actif
  - Minimiser le nombre de contraintes nécessaires ⇐ maximiser le transfert d'information utilisateur → système
  - → 2 critères de sélection complémentaires :
    1. Contraintes informatives : images ambiguës des groupes les moins bien définis
    2. Faible redondance entre les contraintes

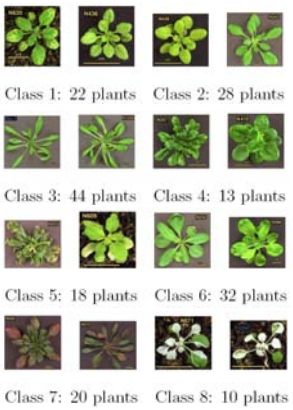


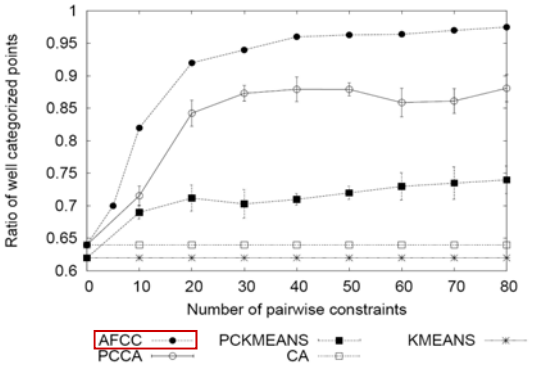
29 janvier 2016 Bases de données avancées 2 37

le cnam

AFCC : résultats illustratifs

- Base *Arabidopsis*





Images fournies par NASC (http://arabidopsis.info), vérité terrain INRA (http://www.inra.fr)

29 janvier 2016 Bases de données avancées 2 38

le cnam

Références

[BH01] A. Budanitsky, G. Hirst. Semantic distances in WordNet: an experimental, application-oriented evaluation of five measures. In *Proceedings of NAACL* 2001.

[CEOT08] Crucianu, M., Estevez, D., Oria, V., Tarel, J.-Ph. Speeding Up Active Relevance Feedback with Approximate kNN Retrieval for Hyperplane Queries, *International Journal of Imaging Systems and Technology*, Vol. 18 (2-3): 150-159, 2008.

[CFB04] Crucianu, M., Ferecatu, M., Boujemaa, N. Relevance feedback for image retrieval: a short survey, juin 2004, 20 p., dans *State of the Art in Audiovisual Content-Based Retrieval, Information Universal Access and Interaction, Including Datamodels and Languages*, rapport du Réseau d'Excellence DELOS2 (6ePCRD).

[CTF05] Crucianu, M., Tarel, J.-Ph., Ferecatu, M. A Comparison of User Strategies in Image Retrieval with Relevance Feedback, *7th International Workshop on Audio-Visual Content and Information Visualization in Digital Libraries* (AVIVDiLib'05), pp. 121-130, Cortona, Italie, 2005.

[CCD14] J. Costa Pereira, E. Coviello, G. Doyle, N. Rasiwasia,G. Lanckriet, R. Levy, and N. Vasconcelos. On the role of correlation and abstraction in cross-modal multimedia retrieval. *Trans. on PAMI*, 36(3):521-535, 2014.

[FBC05] M. Ferecatu, N. Boujemaa, M. Crucianu. Hybrid Visual and Conceptual Image Representation in an Active Relevance Feedback Context. *Proceedings of the 6th ACM SIGMM International Workshop on Multimedia Information Retrieval*, Singapour, 2005.

[Fer05] Ferecatu, M. Image retrieval with active relevance feedback using both visual and keyword-based descriptors, *Thèse de doctorat*, Université de Versailles Saint-Quentin-en-Yvelines, 2005.

le cnam

Références

[GCB08] N. Grira, M. Crucianu, N. Boujemaa. Active Semi-Supervised Fuzzy Clustering, *Pattern Recognition*, Vol. 41, No 5, pp. 1851-1861.

[HSS04] D. R. Hardoon, S. R. Szedmak, and J. R. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. Neural Comput., 16(12):2639-2664, 2004.

[JMF99] A. Jain, M. Murty, P. Flynn. Data clustering: a review. ACM Computing Surveys, 31(3):264-323, 1999.

[Sal68] G. Salton. *Automatic Information Organization and Retrieval*. McGraw-Hill, 1968.

[SMG83] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.

[SS03] B. Schölkopf, A. Smola. *Learning with Kernels*. MIT Press, 2002.

[TK00] S. Tong, D. Koller. Support vector machine active learning with applications to text classification. In *Proceedings of ICML-00, 17th International Conference on Machine Learning*, pp. 999–1006. Morgan Kaufmann, 2000.