Résultats récents en classification supervisée de documents

Jean Beney
Département Informatique
LCI - INSA de Lyon
jean.beney@insa-lyon.fr

EPO

1)Filtrage :

400 demandes de brevets par jour

à envoyer aux équipes concernées

(519 équipes/44 directoires/13 clusters)

radar/hautes fréquences/électronique

nécessité de personnes

universellement compétentes

=> 81,2% de réussite au niveau du directoire

en utilisant les graphiques

EPO, suite

2) Classification pour faciliter la recherche :
624 ou 13 000 classes

Travaux effectués avec Pr. C.H.A. Koster
Université Radbout, Nimègue (NL)

Jeu d'essai EPO2

- 2000 documents/directoire
  - bien classifiés
  - peu de fautes de frappe
  - 5000 mots en moyenne
    - résumés 100-150 mots

- + 1000 documents inconnus/directoire

Méthode Winnow

Méthode apparentée au perceptron :

apprentissage par l'erreur

Winnow symétrique :

2 poids par terme et par classe $w^{+}_{t,c}$, $w^{-}_{t,c}$ modifiés conjointement

Score linéaire :

$$S(d,c)=\Sigma_{t\in d}(w^{+}_{t,c}-w^{-}_{t,c})s(t,c)$$

où $s(t,c)$ représente l'importance d'un terme dans un document

Apprentissage

Double seuil : $\theta^- < \theta^+$

$S(d,c) < \theta^-$ et pertinent : promotion

$w^+_{t,c} * \alpha$ $w^-_{t,c} * \beta$

$S(d,c) > \theta^+$ et non pertinent : rétrogradation

$w^+_{t,c} * \beta$ $w^-_{t,c} * \alpha$

$\alpha > 1$ $\beta < 1$

Étalonnage

Dans la littérature :

$\alpha = 1.1 \quad \beta = 0.9 \quad \theta^- = 0.9 \quad \theta^+ = 1.1$

10 itérations

Après expérimentation sur EPO :

$\alpha < 1.01 \qquad \beta = 1/\alpha \qquad \theta^- = 0.5 \qquad \theta^+ = 2$

10 itérations

=> F(1)=80% sur documents vus

=64% sur documents cachés

interprétation des résultats

Beaucoup plus d'exemples que dans d'autres problèmes de classification :

-> apprendre plus lentement

pour varier les exemples

Double seuil :

-> assez large pour renforcer les documents moyennement classés

-> évite le sur-apprentissage sur les documents difficilement classables (*outliers)*

Évaluation d'un classifieur

- Table des résultats

	Pertinent	Non pertinent	Total
Sélectionné	X	Y	S
Non sélectionné	Z	W	N-S
Total	Re	N-Re	N

- Précision et Rappel

$$\mathbf{P} = \frac{\mathbf{X}}{\mathbf{S}} = \frac{\mathbf{X}}{\mathbf{X}+\mathbf{Y}} \qquad \mathbf{R} = \frac{\mathbf{X}}{\mathbf{R_e}} = \frac{\mathbf{X}}{\mathbf{X}+\mathbf{Z}}$$

Différents usages

- Filtrage chez EPO :
  max. la précision du premier choix
  pour éviter les envois inutiles
- Classification chez EPO :
  max. le rappel
  pour faciliter de travail des examinateurs

Évaluation moyenne

- Le point d'égalité (BEP) $P = R$
- La fonction $F(\beta) = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$

compromis $\beta = 1$

$$F(1) = \frac{2PR}{P+R} = \frac{2X}{R_e + S}$$

Obtenir l'égalité

- $$\mathbf{P = \frac{X}{S} = \frac{X}{X+Y} = R = \frac{X}{R_e} = \frac{X}{X+Z}}$$

- soit $X = 0$ pas de pertinents sélectionnés

  soit $S = R_e \Leftrightarrow Y = Z$ sélectionner autant de documents qu'il existe de pertinents (ceux avec les plus grands scores)

- Présenté à SFC'06

Maximum de F(1)

- Méthode habituelle :

  Ordonner les documents suivant leurs scores
  parcourir la liste, calculer F à chaque pas

Max F <>Point d'égalité

- Pas de relation formelle utile
- Comparaison expérimentale
  sur l'ensemble d'apprentissage

$\delta\theta < 2\,\%$ entre F(1)max et BEP

donc rechercher le maximum de F(1)
aux environs du point d'égalité

$\delta F = F(1)max - F(1)au\ BEP < 7\,\%$

Résultats sur un ensemble de test

- F(1) au point d'égalité précédent est parfois meilleur qu'au maximum de F(1) précédent.
- Donc, on peut éviter de chercher le maximum de F(1) et prendre le point d'égalité
  (gain : 60% de la recherche, 20mn)
- Car un ensemble de documents d'apprentissage n'est pas exactement représentatif de l'ensemble réel.

Travaux en cours-1

- Comparaison Winnow-SVM

- base théorique existante : borne supérieure de l'erreur de généralisation

- pratique :
  - comparer les performances
  - comparer les plans de séparation trouvés

Winnow-SVM

- Hypothèse :
  les solutions sont très proches
- Conséquence :
  Winnow permet de sélectionner
  des supports potentiels
  -> Accélération de SVM
- Selon certains auteurs :
  Winnow est mieux adapté à ce problème

Travaux en cours-2

- Analyse grammaticale
  Actuellement : terme=forme de mot
- Possibilités d'amélioration :
  - terme=lemme ou racine
    Mais : lunette≠lunettes
  - terme=composition de mots
    [pièce DE monnaie]
    [éliminer OBJ impuretés]
    Mais : en très grand nombre

AGFL

- Grammaires Affixes
  efficacité malgré
  ambiguïtés
  énorme lexique

- http://www.agfl.cs.ru.nl/
  Licence GPL

Lexique

- Pour un meilleur lexique :
  croisement ABU-OOo
  amélioration des deux
- Retombées :
  ajout des catégories lexicales à OOo
  -> vérifications grammaticales
- Formes génériques :
  .*ment adverbe ou nom masc.
  .*tion nom fém.

Easy

- Campagne dévaluation d'analyseurs
  - découpage en groupes élémentaires
  - relations entre ces groupes
- en attente des données

Conclusion

Il semble possible
d'égaler les experts,

mais il manque une idée.

EPO : filtrage, recherche d'information
étalonnage
mesure de la qualité, optimisation
comparaison Winnow-SVM
analyse de la langue, AGFL, lexique