DEFI DEFT05 : une approche par classifieur de Bayes

Michel Plantié, Gérard Dray, Jacky Montmain,
Alexandre Meimouni, Pascal Poncelet

LGI2P – Ecole des Mines d'Alès
Parc George Besse, 30000 Nîmes
{michel.plantie, gerard.dray, jacky.montmain}@ema.fr
{alexandre.meimouni, pascal.poncelet}@ema.fr

Mots-clés : Modèles vectoriels, Classifieur de Bayes,
Keywords: Vector models, Bayes classifier,

Résumé Cet article présente notre approche de la problématique soulevée par le défi DEFT05. Cette approche est basée sur une représentation vectorielle des textes, et par l'application de méthodes de classification de «Bayes ».

Abstract This paper presents our approach of the DEFT05 problem. This approach is based on a vector representation of texts, and the application of "Bayes" classifier.

1 Introduction

L'objectif du défi fouille de textes 2005 (DEFT'05) était de supprimer des phrases non pertinentes dans un discours politique. Pour cela un corpus d'apprentissage, constitué de phrases issues de discours de François Mitterrand parsemées dans des discours de Jacques Chirac, a été fourni aux participants. Les participants au défi devaient définir une méthode, automatique (ou semi-automatique) et reproductible, permettant de détecter les phrases de F. Mitterrand. L'évaluation des résultats des méthodes proposées a été réalisée par les organisateurs sur un corpus de test.

Nous avons abordé le défi comme un problème de classification de textes. En choisissant cette approche nous avons donc considéré que les phrases étaient indépendantes les unes des autres, la problématique à traiter se traduisant par :

« Quelle méthode de représentation des textes et quelle méthode de classification permettraient de classer au mieux les phrases de François Mitterrand et de Jaques Chirac? »

2 Méthode expérimentale

2.1 Prétraitement et indexation du corpus

Le corpus d'apprentissage initial était composé de 50096 phrases de J. Chirac et de 7183 phrases de F. Mitterrand. Dans une première étape les phrases du corpus ont été étiquetées et lemmatisées. La seconde étape a consisté à indexer les phrases sous la forme de vecteurs. Trois modèles vectoriels ont été utilisés pour les expérimentations ; chaque coordonnée du vecteur correspondant à un mot du vocabulaire du corpus total (les mots vides étant éliminés).

- Modèle binaire : pour chaque mot de la phrase appartenant au vocabulaire du corpus, la coordonnée du vecteur est mise à 1, les autres coordonnées étant à 0.

- Modèle FT (Fréquence des Termes) : pour chaque mot de la phrase appartenant au vocabulaire, la coordonnée correspondante du vecteur est incrémentée. La valeur d'une coordonnée indiquant le nombre d'occurrences du mot dans la phrase. (Remarque : les phrases étant relativement courtes et les mots vides étant éliminés ce modèle est peu différent du modèle précédent)

2.2 Méthode de classification

Comme nous l'avons indiqué dans l'introduction, nous avons abordé le défi comme un problème de classification de textes. La classification de textes consiste à assigner des catégories prédéfinies à des documents textuels. Soit $C = \left(c_1,...,c_j,...,c_n\right)$ l'ensemble des classes d'appartenance possibles pour un document.

Nous avons choisi, pour des raisons de rapidité de calcul, le classifieur naïf de Bayes. Cette méthode peu complexe, a déjà fourni de bons résultats sur des problèmes de classification de textes. Cette approche suppose l'existence d'un modèle stochastique de génération des documents textuels. En inversant ce modèle, nous pouvons prédire, pour un nouveau document, la probabilité d'appartenir à une classe quelconque. La règle de classification de Bayes consistant à attribuer la classe dont la probabilité est la plus élevée.

Un document est représenté par un vecteur $d_i = \left(ft_{i1}...,ft_{it},...,ft_{i|V|}\right)$ dans lequel V représente l'ensemble des mots du vocabulaire retenu du corpus et ft_{it} représente le nombre d'occurrences du mot m_t dans le document d_i.

La règle de classification de Bayes consiste à attribuer à un document d_i la classe dont la probabilité suivante est la plus élevée :

$$P(c_j/d_i)=\frac{P(c_j)P(d_i/c_j)}{P(d_i)},$$

avec c_j une classe et d_i une phrase.

La probabilité $P(d_i)$ étant indépendante des classes, la règle de classification de Bayes peut être exprimée par :

$$\hat{c}(d_i)=\arg\max_j\left(p(c_j)\,p(d_i/c_j)\right),$$

$\hat{c}(d_i)$ représentant la classe estimée pour le document d_i

En considérant l'indépendance des mots :

$$\hat{c}(d_i)=\arg\max_j\left(p(c_j)\prod_{t=1}^{|V|}p(m_t/c_j)^{ft_{ti}}\right)$$

Les probabilités $p(m_t/c_j)$ et $p(c_j)$ peuvent être estimées sur un corpus d'apprentissage par :

$$p(m_t/c_j)=\frac{1+\sum_{d_i\in c_j}ft_{ti}}{|V|+\sum_{t=1}^{|V|}\sum_{d_i\in c_j}ft_{ti}}$$

$p(c_j)=\frac{|c_j|}{\sum_{k=1}^{|C|}|c_k|}$, avec $|c_j|$ représentant le nombre de documents de la classe j.

Cette méthode a été appliquée à la problématique du défi en considérant : deux classes (Mitterrand et Chirac) et les phrases comme des documents.

Dans le cas de la représentation vectorielle binaire, les fréquences des termes sont remplacées par une variable bin_{ti} valant 1 si le mot t apparaît dans la phrase i et 0 sinon.

2.3 Méthodologie d'apprentissage et résultats

2.3.1 Constitution des ensembles d'apprentissage

Le nombre de phrases de F. Mitterrand dans le corpus d'apprentissage étant disproportionné par rapport au nombre de phrases de J. Chirac, nous avons choisi de créer plusieurs ensemble d'apprentissage. Au total 7 ensembles ont été formés, en conservant dans chacun d'entre eux toutes les phrases de F. Mitterrand et en complétant par le même nombre de phrases de J. Chirac par tirage aléatoire.

Sur chacun de ces ensembles un apprentissage par validation croisée à dix ensembles a été réalisé. Pour chaque modèle vectoriel de représentation nous avons donc pu estimer une moyenne des performances envisageables de notre approche sur le jeu de test.

2.3.2 Stratégie de vote

Chaque phrase du jeu de test est analysée par 7 classifieurs et est donc associée à 7 réponses. L'idée de base revient ensuite à considérer chaque classifieur comme un « expert » qui se prononce sur l'appartenance ou non d'une phrase à une classe. Chaque classifieur C_k donne la classe d'appartenance du candidat ainsi que sa probabilité d'appartenance. Pour chaque phrase on construit alors le vecteur à 7 composantes des probabilités d'appartenance à la classe « Chirac ». Pour un classifieur donné, une probabilité supérieure à $\alpha = 0.5$ signifie que le classifieur associe la phrase à la classe « Chirac » ; il s'agit maintenant de combiner les avis des 7 « experts » sur la base de cette règle de choix individuel. Nous nous sommes donc intéressés à la sémantique des règles de choix collectif comme :

- Le vote majoritaire à α % avec $0.5 \leq \alpha \leq 1$. Il fournit des comportements de vote allant de la majorité à l'unanimité.
- Le vote unanime restreint : choisir l'alternative approuvée par la plupart des classifieurs.
- Le veto limité : rejeter une alternative rejetée au moins par q classifieurs.

Dans ces règles, la notion de quantifieur linguistique est implicite. Un quantifieur linguistique modélise une proportion floue et permet d'expliciter des propositions telles que : « q classifieurs parmi n préfèrent la solution S_j ». Le quantifieur Q « au moins q classifieurs parmi n » peut également être représenté par un ensemble flou où $\mu_Q(n)=1$ et $\mu_Q(j) \leq \mu_Q(j+1),\ j=0,n-1$ (voir).

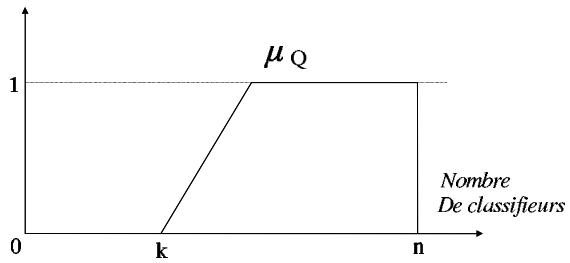

Figure 1

Les règles de majorité peuvent être implémentées en utilisant la notion de Moyenne Pondérée Ordonnée (OWA) introduite par (YAGER, 1988). L'idée est d'utiliser un ensemble de poids $p_1,..,p_n$ qui ne sont pas assignés aux classifieurs mais fonction de l'ordre partiel des scores (les probabilités d'appartenance ici) attribués par ceux-ci : les poids les plus élevés sont assignés aux classifieurs qui expriment les meilleurs scores partiels. Prenons σ une permutation sur (1,2,..,n) telle que $x^{C_{\sigma(1)}} \geq x^{C_{\sigma(2)}} \geq ... \geq x^{C_{\sigma(n)}}$ où $x^{C_{\sigma(k)}}$ est le score associé à une

phrase candidate par le classifieur $C_{\sigma(k)}$. La combinaison convexe est alors définie par :

$$\varphi(x^{C_1}, x^{C_2}, .., x^{C_n}, p_1, p_2, .., p_n) = \sum_{j=1}^{n} p_j x^{C_{\sigma(j)}}$$

La moyenne arithmétique correspond au cas $p_j = 1/n, \forall j$. On a $\varphi = \max$ quand $p_1 = 1, p_j = 0, j \geq 2$; $\varphi = \min$ quand $p_n = 1, p_j = 0, j \leq n-1$. La règle de majorité « q parmi n » est obtenue quand $p_1 = p_2 = .. = p_q = 1/q$ *and* $p_{q+1} = p_{q+2} = .. = p_n = 0$.

D'autres modèles de règles de majorité sont proposés dans (ZADEH, 1983 ; KACPRIZYK, 1987).

Les règles d'unanimité restreinte contrairement aux règles de majorité restreinte, n'autorisent pas de compensation par les classifieurs Ck. C'est une altération de la règle du minimum qui stipule qu'une alternative est approuvée collectivement lorsque chaque classifieur C_k séparément approuve cette alternative : $\varphi(x^{C_1}, x^{C_2}, .., x^{C_n}, p_1, p_2, .., p_n) = \min_{j=1,n} \max(p_j, x^{C_{\sigma(j)}})$, où les poids satisfont la condition $\min_{j=1,n} p_j = 0$. L'approbation requise par q classifieurs parmi n est accomplie quand $p_1 = p_2 = .. = p_q = 0$ *et* $p_{q+1} = p_{q+2} = .. = p_n = 1$. Il est à noter que l'approbation partielle ($x^k = q/n$) par tous les Ck n'est pas identique à l'approbation complète par q classifieurs alors que ces approbations sont équivalentes dans le modèle OWA. $\varphi = \max$ quand $p_1 = 0, p_j = 1, \forall j \geq 2$ (KONING, 1990).

Quand $0 = p_1 \leq p_2 \leq .. \leq p_n$, la situation où q est mal défini peut être assimilée à l'approbation « la plupart des classifieurs C_k». « La plupart des » est vu ici comme une proportion absolue décrite par un ensemble flou Q où $\mu_Q(j) = p_j$ (KONING, 1990).

Dans ce qui suit le modèle d'agrégation des classifieurs est assimilé à une règle d'unanimité restreinte « q classifieurs parmi n » : $\varphi(x^{C_1}, x^{C_2}, .., x^{C_n}, p_1, p_2, .., p_n) = \min_{j=1,n} \max(p_j, x^{C_{\sigma(j)}})$ avec $p_1 = p_2 = .. = p_q = 0$ et $p_{q+1} = p_{q+2} = .. = p_n = 1$. Cela correspond à l'idée que si au moins q classifieurs ont donné une réponse favorable (p>0.5) pour l'attribution d'une phrase à la classe « Chirac », alors la phrase sera collectivement classée « Chirac ». L'avis des (n-q) autres classifieurs n'a pas d'importance selon cette règle qui n'autorise pas la compensation.

Nous avons étudié dans le cadre du défi DEFT05 les deux stratégies d'agrégation des votes des classifieurs : l'unanimité restreinte précédemment décrite et la plus classique moyenne arithmétique.

2.3.3 Paramètres des stratégies de vote

Concernant la moyenne arithmétique, un seul paramètre a été réglé : c'est le seuil qui détermine l'appartenance d'un candidat à une classe pour chaque classifieur Ck. Dans notre cas nous avons pris la valeur de 0.5. Ce seuil signifie que chaque classifieur stipulera que si son vote est supérieur à 0.5 le candidat correspondant appartient à la classe « Chirac ». Cette valeur de 0.5 s'explique par le fait que le classifieur lui-même calcule une probabilité d'appartenance d'un candidat à une classe et dès que la valeur de 0.5 est franchie la réponse

est considérée positive pour cette classe. Remettre en cause cette valeur change les données fondamentales du classifieur.

Concernant la règle d'agrégation d'unanimité « q classifieurs parmi n », deux paramètres ont été réglés : le seuil qui détermine l'appartenance d'un candidat à une classe pour chaque classifieur Ck, et le paramètre q. Nous avons choisis comme précédemment dans nos expérimentations un seuil de 0.5 pour tous les classifieurs. Concernant le paramètre q, nous avons choisi la valeur de 4 qui correspond au fait que 4 votes favorables à une classe sur 7 sont suffisants pour associer ce candidat à cette classe. Nous avons effectué quelques essais également avec la valeur 7 pour q, qui correspond à l'unanimité complète, mais la différence n'est pas déterminante.

2.3.4 Résultats

Nous avons appliqués les deux stratégies de vote uniquement sur les classifieurs à base de vecteurs FT. Concernant les vecteurs binaires seule la stratégie d'agrégation d'unanimité de « 4 classifieurs parmi 7 » a été appliquée.

Les performances de classification obtenues (mesure F-score) par validation croisée pour le défi tâche 1 sont :

Pour l'exécution 1, nous avons utilisés les vecteurs binaires, et nous avons appliqué la moyenne arithmétique des classifieurs, ce qui donne les résultats suivants :

Classe « Chirac » : précision 0.807, rappel 0.767, F-Score : 0.786

Classe « Mitterand » : précision 0.789, rappel 0.826, F-Score : 0.807

Pour l'exécution 2, nous avons utilisés les vecteurs FT, nous avons appliqué la moyenne arithmétique des classifieurs, ce qui donne les résultats suivants :

Classe « Chirac » : précision 0.808, rappel 0.770, F-Score : 0.789

Classe « Mitterand » : précision 0.781, rappel 0.817, F-Score : 0.799

Pour l'exécution 3, nous avons utilisés une moyenne arithmétique des 14 classifieurs constitués par les deux jeux d'essais précédents.

Les performances de classification obtenues (mesure F-score) par validation croisée pour le défi tâche 2 sont :

Pour l'exécution 1, nous avons utilisés les vecteurs binaires, nous avons appliqué la moyenne arithmétique des classifieurs, ce qui donne les résultats suivants :

Classe « Chirac » : précision 0.820, rappel 0.761, F-Score : 0.788

Classe « Mitterand » : précision 0.793, rappel 0.841, F-Score : 0.816

Pour l'exécution 2, nous avons utilisés les vecteurs FT, nous avons appliqué la moyenne arithmétique des classifieurs, ce qui donne les résultats suivants :

Classe « Chirac » : précision 0.821, rappel 0.768, F-Score : 0.773

Classe « Mitterand » : précision 0.782, rappel 0.785, F-Score : 0.783

Pour l'exécution 3, nous avons utilisés les vecteurs FT, et nous avons appliqué la stratégie d'unanimité restreinte des classifieurs, ce qui donne les résultats suivants :

Classe « Chirac » : précision 0.768, rappel 0.767, F-Score : 0.789

Classe « Mitterand » : précision 0.796, rappel 0.853, F-Score : 0.824

Les performances de classification obtenues (mesure F-score) par validation croisée pour le défi tâche 3 sont :

Pour l'exécution 1, nous avons utilisés les vecteurs binaires, et nous avons appliqué la stratégie d'unanimité restreinte des classifieurs, ce qui donne les résultats suivants :

Classe « Chirac » : précision 0.842, rappel 0.778, F-Score : 0.809

Classe « Mitterand » : précision 0.783, rappel 0.854, F-Score : 0.823

Pour l'exécution 2, nous avons utilisés les vecteurs FT, nous avons appliqué la moyenne arithmétique des classifieurs, ce qui donne les résultats suivants :

Classe « Chirac » : précision 0.812, rappel 0.752, F-Score : 0.781

Classe « Mitterand » : précision 0.781, rappel 0.835, F-Score : 0.807

Pour l'exécution 3, nous avons utilisés les vecteurs FT, et nous avons appliqué la stratégie d'unanimité restreinte des classifieurs, ce qui donne les résultats suivants :

Classe « Chirac » : précision 0.836, rappel 0.774, F-Score : 0.804

Classe « Mitterand » : précision 0.752, rappel 0.819, F-Score : 0.784

3 Conclusion et perspectives

Les résultats obtenus sur l'ensemble d'apprentissage montrent que les représentations vectorielles (binaire et FT) ne sont pas influentes pour cette problématique. Le modèle binaire donnant même de meilleurs résultats.

Les résultats obtenus par validation croisée, nous laissaient espérer des performances semblables sur le jeu de test. Malheureusement ce ne fut pas le cas. Pour l'instant nous n'avons pas d'explication à cette baisse importante des performances de notre approche sur le corpus de test.

D'autres expérimentations sont actuellement en cours pour essayer de comprendre cette contre performance et améliorer notre approche. En particulier, nous travaillons sur des méthodes permettant de sélectionner les mots les plus discriminants et ainsi diminuer la taille

des vecteurs de représentation. Ainsi, d'autres méthodes de classification pourraient être appliquées (arbre de décision, clustering flou, …), combinées et évaluées.

Références

ANDREW MCCALLUM, KAMAL NIGAM (1998), A Comparison of Event Models for Naive Bayes Text Classification, AAAI-98 Workshop on Learning for Text Categorization

KACPRZYK, J. (1987). TOWARDS "HUMAN-CONSISTENT" DECISION SUPPORT SYSTEMS THROUGH COMMONSENSE KNOWLEDGE-BASED DECISION MAKING AND CONTROL MODELS: A FUZZY APPROACH. COMPUTERS AND ARTIFICIAL INTELLIGENCE, 6 (2), 97-122.

KONING, J-L. (1990). UN MÉCANISME DE GESTION DE RÈGLES DE DÉCISION ANTAGONISTES POUR LES SYSTÈMES À BASE DE CONNAISSANCES. THÈSE DE L'UNIVERSITÉ PAUL SABATIER DE TOULOUSE.

(SALTON ET AL., 1983) SALTON, G., INTRODUCTION TO MODERN INFORMATION RETRIEVAL », MCGRAW-HILL BOOK COMPANY, NEW YORK 1983

KARL-MICHAEL SCHNEIDER (2004) : A NEW FEATURE SELECTION SCORE FOR NAIVE BAYES TEXT CLASSIFICATION BASED ON KL-DIVERGENCE. OMPANION VOLUME TO THE PROCEEDINGS OF THE 42ND MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (ACL 2004) <HTTP://WWW.ACL2004.ORG/>, BARCELONA, SPAIN, PP. 186-189, 2004

YAGER, R. (1988). ON ORDERED WEIGHTED AVERAGING AGGREGATION OPERATORS IN MULTICRITERIA DECISION-MAKING. IEEE TRANSACTION SYSTEMS, MAN AND CYBERNETICS, 18, 183-190.

ZADEH, L. (1983). A COMPUTATIONAL APPROACH TO FUZZY QUANTIFIERS IN NATURAL LANGUAGES. COMPUTERS AND MATHEMATICS WITH APPLICATIONS, 9, 149-184