Apprentissage actif d'émotions dans les dialogues Homme-Machine

Alexis Bondu*, Vincent Lemaire*, Barbara Poulain*

*France Telecom R&D TECH/SUSI/TSI
2 avenue Pierre Marzin 22300 Lannion
prenom.nom@orange-ft.com

Résumé.
La prise en compte des émotions dans les interactions Homme-machine permet de concevoir des systèmes intelligents, capables de s'adapter aux utilisateurs. Les techniques de redirection d'appels dans les centres téléphoniques automatisés se basent sur la détection des émotions dans la parole. Les principales difficultés pour mettre en œuvre de tels systèmes sont l'acquisition et l'étiquetage des données d'apprentissage. Cet article propose l'application de deux stratégies d'apprentissage actif à la détection d'émotions dans des dialogues en interaction homme-machine. L'étude porte sur des données réelles issues de l'utilisation d'un serveur vocal et propose des outils adaptés à la conception de systèmes automatisés de redirection d'appels.

1 Introduction

Grâce à des techniques récentes de traitement de la parole, de nombreux centres d'appels téléphoniques automatisés voient le jour. Ces serveurs vocaux permettent aux utilisateurs d'exécuter diverses tâches en dialoguant avec une machine. Les entreprises cherchent à améliorer la satisfaction de leurs clients en les redirigeant en cas de difficulté vers un opérateur humain. L'aiguillage des utilisateurs mécontents revient à détecter les émotions négatives dans leurs dialogues avec la machine, sous l'hypothèse qu'un problème de dialogue génère un état émotionnel particulier chez le sujet.

La détection d'émotions dans la parole est généralement traitée comme un problème d'apprentissage supervisé. Cela s'explique par le fait que les descripteurs utilisés sont relativement éloignés du concept d'émotion, dans la pratique l'étiquetage d'exemples s'avère nécessaire. La détection d'émotions se limite généralement à une classification binaire, la prise en compte de labels plus fins pose le problème de l'objectivité de l'étiquetage (Liscombe et al., 2005). Dans ce cadre, les données sont coûteuses à acquérir et à étiqueter. L'apprentissage actif peut diminuer ce coût en étiquetant uniquement les exemples jugés informatifs pour le modèle.

Cet article propose une approche d'apprentissage actif pour la redirection automatique d'appels. La première section présente le contexte de l'étude ainsi que les données utilisées. Les différentes stratégies d'apprentissage envisagées ainsi que le modèle utilisé sont traités dans la section 2. La dernière section est consacrée aux résultats obtenus et à leur discussion.

2 Classification d'émotions : caractérisation des données

Cet article se base sur des travaux antérieurs (Poulain, 2006) cherchant à caractériser au mieux des échanges vocaux en vue d'une classification d'émotions. Le but est de réguler le dialogue entre des utilisateurs et un serveur vocal. Cette étude porte plus particulièrement sur la pertinence des variables décrivant les données par rapport à la détection d'émotions.

Les données utilisées sont issues d'une expérience mettant en jeu 32 utilisateurs qui testent un service boursier fonctionnant sur un serveur vocal. Du point de vue de l'utilisateur, le test consiste à gérer un portefeuille fictif d'actions, le but étant de réaliser la plus forte plus value. Les traces vocales obtenues constituent le corpus de cette étude, soit 5496 "tours de parole" échangés avec la machine. Les tours de parole sont caractérisés par 200 variables acoustiques, décrivant notamment la variation du volume sonore, la variation de la hauteur de voix, le rythme d'élocution. Les données sont également caractérisées par 8 variables dialogiques décrivant notamment l'âge du locuteur, le rang du dialogue, la durée du dialogue. Chaque tour de parole est étiqueté manuellement comme étant porteur d'émotions positives ou négatives.

Le sous-ensemble des variables les plus informatives vis-à-vis de la détection d'émotions est déterminé grâce à un sélecteur bayésien naïf (Boullé, 2006). Au début de ce procédé, l'ensemble des attributs est vide, à chaque itération on ajoute l'attribut qui améliore au plus la qualité prédictive du modèle. L'algorithme s'arrête lorsque l'ajout d'attributs n'améliore plus la qualité du modèle. Finalement, 20 variables ont été sélectionnées pour caractériser les échanges vocaux[1]. Dans le cadre de cet article, les données utilisées sont issues du même corpus et de cette étude antérieure. Chaque tour de parole est donc caractérisé par 20 variables.

Étant donnés :

- $\mathcal{M}$ un modèle prédictif muni d'un algorithme d'apprentissage $\mathcal{L}$
- Les ensembles U_x et L_x d'exemples non étiquetés et étiquetés
- N le nombre d'exemples d'apprentissage souhaités.
- L'ensemble d'apprentissage T constitué de couples *"instance, étiquette"* notés $(u, f(u))$.
- La fonction $\mathcal{U}tile : \mathbb{X} \times \mathbb{M} \rightarrow \Re$ qui estime l'utilité d'une instance pour l'apprentissage.

Condition initiale : $\|T\| < N$

Répéter

(A) Entraîner le modèle $\mathcal{M}$ grâce à $\mathcal{L}$ et T(et éventuellement U_x).
(B) Rechercher l'instance $q = argmax_{u \in U_x} \mathcal{U}tile(u, \mathcal{M})$
(C) Retirer q de U_x et demander l'étiquette $f(q)$ à l'oracle.
(D) Ajouter q à L_x et ajouter $(q, f(q))$ à T

Tant que $\|T\| < N$

Algorithme 1: échantillonnage sélectif, Muslea (Muslea, 2002)

[1]Il est important de noter que parmi les 20 variables retenues, certaines sont obtenues de manière non automatique. On suppose dans la suite de l'article qu'on dispose d'un moyen pour les évaluer toutes.

3 Classification active d'émotions

3.1 Introduction

La mise en œuvre d'un système automatique de détection d'émotions requiert généralement l'entraînement d'un classifieur. Ici, le modèle qui va classifier les émotions est réalisé grâce à un processus d'apprentissage actif. A la différence de l'apprentissage passif, qui utilise un ensemble de données déjà étiquetées, l'apprentissage actif permet au modèle de construire lui-même son ensemble d'apprentissage au cours de son entraînement. Parmi les stratégies d'apprentissage actif existantes (Castro et Nowak, 2005), on se place dans le cadre de l'échantillonnage sélectif où le modèle dispose d'un "sac" d'instances non étiquetées dont il peut demander les labels.

Muslea (Muslea, 2002) a formalisé de manière générique l'échantillonnage sélectif à travers l'Algorithme (1). Celui-ci met en jeu la fonction $\mathcal{U}tile(u, \mathcal{M})$ qui estime l'intérêt d'une instance $u \in U_x$ pour l'apprentissage du modèle $\mathcal{M}$. La problématique centrale de l'apprentissage actif est de *"préjuger efficacement"* de l'intérêt des exemples avant de les étiqueter.

3.2 Le choix du modèle

La grande variété des modèles capables de résoudre des problèmes de classification et parfois le grand nombre de paramètres nécessaires à leur utilisation rend souvent l'apport d'une stratégie d'apprentissage difficile à mesurer. On choisit d'utiliser une fenêtre de Parzen à noyau gaussien et de norme L2 (Parzen, 1962) car ce modèle prédictif n'utilise qu'un seul paramètre et est capable de fonctionner naturellement avec peu d'exemples. La "sortie" de ce modèle est une estimation de la probabilité d'observer l'étiquette y_i conditionnellement à l'instance u_n :

$$\hat{P}(y_i|u_n) = \frac{\sum_{j=1}^{N} \mathbb{1}_{\{f(l_j)=y_i\}} K(u_n, l_j)}{\sum_{j=1}^{N} K(u_n, l_j)} \qquad avec\, l_j \in L_x \; et \; u_n \in U_x \cup L_x \qquad (1)$$

$$K(u_n, l_j) = e^{\frac{||u_n - l_j||^2}{2\sigma^2}}$$

La valeur optimale (σ^2=0.24) du paramètre du noyau a été déterminée grâce à une cross-validation sur l'erreur quadratique moyenne (Chappelle, 2005). Cette valeur est utilisée par la suite pour fixer le paramètre de la fenêtre de parzen.

Pour que le modèle puisse affecter une étiquette $\hat{f}(u_n)$ à l'instance u_n, un seuil de décision noté $\mathcal{S}euil(L_x)$ est calculé. Ce seuil minimise l'erreur de prédiction[2] sur l'ensemble d'apprentissage. L'étiquette attribuée est $\hat{f}(u_n) = 1$ si $\{\hat{P}(y_1|u_n) > \mathcal{S}euil(L_x)\}$, et $\hat{f}(u_n) = 0$ sinon. Puisque le seul paramètre de la fenêtre de Parzen est fixé, l'apprentissage du modèle se réduit au "comptage" des instances (au sens du noyau gaussien). Cela permet de comparer uniquement les stratégies de sélection d'exemples sans être influencé par l'apprentissage du modèle.

3.3 Deux stratégies d'apprentissage actif

La première stratégie d'apprentissage actif proposée a pour but de réduire l'erreur de généralisation du modèle, cette erreur peut être estimée par le risque empirique (Zhu et al., 2003).

[2] La mesure d'erreur utilisée est le BER, voir section 4.1

Ici, le risque $R(\mathcal{M})$est définit comme étant la somme des probabilités que le modèle prenne une mauvaise décision sur l'ensemble d'apprentissage. On note $P(y_i|l_n)$ la probabilité réelle d'observer la classe y_i pour l'instance $l_n \in L_x$. Le risque empirique s'écrit alors selon l'équation 2, avec $\mathbb{1}$ la fonction indicatrice égale à 1 si $f(l_n) \neq y_i$ et égale à 0 sinon. La fenêtre de parzen estime $P(y_i|l_n)$, on peut donc approximer le risque empirique en adoptant un apriori uniforme sur les $P(l_n)$ (voir équation 3). Le but de cette stratégie est de sélectionner l'instance non étiquetée $u_i \in U_x$ qui minimisera le risque à l'itération $t+1$. On estime $R(\mathcal{M}^{+u_n})$ le risque *"attendu"* après l'étiquetage de l'instance u_n. Pour se faire, on se base sur les données étiquetées dont on dispose et on suppose que $f(u_n) = y_1$ ***[resp*** $f(u_n) = y_0$***]*** pour estimer $\hat{R}(\mathcal{M}^{+(u_n,y_1)})$ ***[resp*** $\hat{R}(\mathcal{M}^{+(u_n,y_0)})$***]***. L'équation 4 montre comment agréger les estimations de risque selon les probabilités d'observer chacune des classes. Pour exprimer la stratégie de réduction du risque sous forme algorithmique, il suffit de remplacer l'étape (B) de l'algorithme 1 par : *"Rechercher l'instance* $q = argmin_{u \in U_x} \hat{R}(\mathcal{M}^{+u_n})$*"*.

$$R(\mathcal{M}) = \sum_{n=1}^{N} \sum_{y_i=0,1} \mathbb{1}_{\{f(l_n) \neq y_i\}} P(y_i|l_n) P(l_n) \qquad avec\, l_n \in L_x \tag{2}$$

$$\hat{R}(\mathcal{M}) = \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}_{\{f(l_n) \neq y_i\}} \hat{P}(y_i|l_n) \tag{3}$$

$$\hat{R}(\mathcal{M}^{+u_n}) = \hat{P}(y_1|u_n)\hat{R}(\mathcal{M}^{+(u_n,y_1)}) + \hat{P}(y_0|u_n)\hat{R}(\mathcal{M}^{+(u_n,y_0)}) \qquad avec\, u_n \in U_x \tag{4}$$

La deuxième stratégie d'apprentissage consiste à choisir l'instance pour laquelle la prédiction du modèle est la plus incertaine possible. On considère que *l'incertitude* d'une prédiction est maximale quand la probabilité de sortie du modèle se rapproche du seuil de décision (voir équation 5). L'algorithme correspondant à cette stratégie s'obtient en remplaçant l'étape (B) de l'algorithme 1 par : *"Rechercher l'instance* $q = argmax_{u \in U_x} \mathcal{I}ncertain(u_n)$*"*.

$$\mathcal{I}ncertain(u_n) = \sum_{y_i=0,1} \mathbb{1}_{\{\hat{f}(u_n)=y_i\}} \frac{1}{|\hat{P}(y_i|u_n) - \mathcal{S}euil(L_x)|} \tag{5}$$

En dehors de ces deux stratégies actives, une approche "stochastique" sélectionne uniformément les exemples selon leur distribution de probabilité. Cette dernière approche est notre juge de paix et tient lieu de référence pour mesurer l'apport des stratégies actives.

4 Résultats

4.1 Résultats et Discussion

Les résultats présentés sont issus d'expériences répétées cinq fois pour chaque stratégie d'apprentissage[3]. Au début de l'expérience, l'ensemble d'apprentissage ne possède que deux exemples (un positif et un négatif) choisis de manière aléatoire. A chaque itération, dix

[3]les moustaches sur les courbes de la figure 1 correspondent à 4 fois la variance des résultats.

exemples sont choisis pour être étiquetés et ajoutés à l'ensemble d'apprentissage. Le problème de classification traité est déséquilibré au profit d'une des deux classes. On possède 92% d'émotions "positives ou neutres" et 8% d'émotions "négatives". Afin de réprésenter correctement les gains en terme de classification l'évaluation du modèle est réalisée en utilisant le "Balanced Error Rate" (BER) (Yi-Wei et Chih-Jen, 2006) sur un ensemble de test[4].

Selon les résultats de la figure 1, la "réduction du risque" est la stratégie qui maximise la qualité du modèle. En étiquetant 60 exemples grâce à cette approche, la performance du modèle est très proche du BER asymptotique[5] (l'écart n'est que de 0.04). La stratégie de "maximisation de l'incertitude" n'est pas performante au début de l'apprentissage. Il faut étiqueter 180 exemples pour que cette stratégie donne de meilleurs résultats que l'approche "stochastique". Cette approche à l'avantage d'être rapide, pour 300 exemples étiquetés on observe une performance comparable à la "réduction du risque", avec un temps de calcul 5 fois moins important.

La figure 2 compare les performances d'un modèle passif entraîné sur la totalité de l'ensemble d'apprentissage et de deux modèles actifs entraînés sur 100 exemples. Les résultats sont présentés sous la forme de courbes de lift réalisées sur l'ensemble de test. Ces courbes montrent la proportion d'émotions "négatives" détectées par le modèle, en considérant une certaine proportion de la population totale. Par exemple, le modèle entraîné grâce à la minimisation du risque détecte 74% des émotions négatives en utilisant 20% de la population totale (voir point "A" de la figure 2). La "maximisation de l'incertitude" permet de détecter plus efficacement les émotions négatives que la "réduction du risque" (ce résultat ne prend pas en compte le taux de fausses alertes). Les deux modèles actifs offrent des performances proches de celle du modèle passif, en utilisant 37 fois moins d'exemples d'apprentissage.

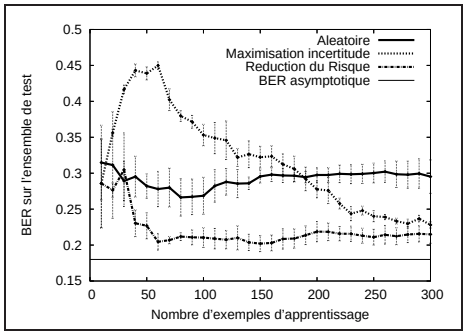

FIG. 1 – *BER en fonction du nombre d'exemples d'apprentissage utilisés.*

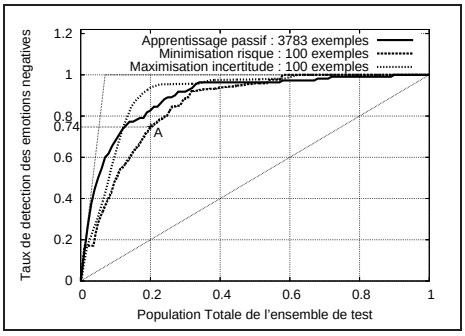

FIG. 2 – *Courbes de lift : Apprentissage Actif / Passif (moyenne sur 5 expériences)*

4.2 Conclusion et perspectives

Cet article montre l'intérêt de l'apprentissage actif pour un domaine où l'acquisition et l'étiquetage des données sont particulièrement coûteux. Au vu des résultats obtenus lors de nos expériences, l'apprentissage actif est pertinent pour la détection d'émotions dans la parole.

[4]L'ensemble de test comporte 1613 exemples et l'ensemble d'apprentissage 3783 exemples

[5]Le BER asymptotique est calculé en considérant les 3783 exemples d'apprentissage.

D'une manière générale, les stratégies d'apprentissage actif permettent d'estimer l'utilité des exemples d'apprentissage. Ces mêmes critères pourraient être utilisés dans le cadre d'un apprentissage en ligne. L'ensemble d'apprentissage serait constitué des N exemples les plus "utiles" vus jusqu'à présent (avec N fixé). Cette approche permettrait de considérer des problèmes d'apprentissage non stationnaires.

La détection d'émotions pourait être traitée par une double stratégie diminuant le coût d'acquisition des données : (i) une sélection de variables permettant de conserver uniquement les caractéristiques nécessaires et suffisantes à la classification ; (ii) une sélection d'exemples permettant de ne conserver que les instances utiles à l'apprentissage. Ceci sera exploré dans des travaux futurs.

Références

Boullé, M. (2006). An enhanced selective naive bayes method with optimal discretization. In I. Guyon, S. Gunn, M. Nikravesh, et L. Zadeh (Eds.), *Feature extraction, foundations and Application*, pp. 499–507. Springer.

Castro, R. andWillett, R. et R. Nowak (2005). Faster rate in regression via active learning. In *NIPS (Neural Information Processing Systems)*, Vancouver.

Chappelle, O. (2005). Active learning for parzen windows classifier. In *AI & Statistics*, Barbados, pp. 49–56.

Liscombe, J., G. Riccardi, et D. Hakkani-Tür (2005). Using context to improve emotion detection in spoken dialog systems. In *InterSpeech*, Lisbon.

Muslea, I. (2002). *Active Learning With Multiple View*. Phd thesis, University of southern california.

Parzen, E. (1962). On estimation of a probability density function and mode. *Annals of Mathematical Statistics 33*, 1065–1076.

Poulain, B. (2006). Sélection de variables et modélisation d'expressions d'émotions dans des dialogues hommes-machine. In *EGC (Extraction et Gestion de Connaissance)*, Lille. + Note technique http ://perso.rd.francetelecom.fr/lemaire.

Yi-Wei, C. et Chih-Jen (2006). Combining SVMs with various feature selection strategies. In I. Guyon, S. Gunn, M. Nikravesh, et L. Zadeh (Eds.), *Feature extraction, foundations and Application*, pp. 315–323. Springer.

Zhu, X., J. Lafferty, et Z. Ghahramani (2003). Combining active learning and semi-supervised learning using gaussian fields and harmonic functions. In *ICML (International Conference on Machine Learning)*, Washington.

Summary

Taking into account emotions in Human-machine interactions can be helpfull for intelligent systems designing. The main difficulty, for the conception of calls center's automatic shunting system, is the cost of data labelling. This paper propose to reduce this cost thanks to active learning strategies. The study is based on real data resulting from the use of a vocal stock exchange server.