Apprentissage déséquilibré pour la détection automatique des signaux de l'implication durable dans les conversations en parfumerie

Yizhe Wang, D Nouvel, L Marguerite, P Gaël

Apprentissage déséquilibré pour la détection automatique des signaux de l'implication durable dans les conversations en parfumerie

W. Yizhe, D. Nouvel, L. Marguerite, P. Gaël

XiKO

ERTIM - Équipe de Recherche Textes, Informatique, Multilinguisme

Problématique

Notion de l'implication durable

- Un des concepts clés dans les études en comportement du consommateur
- État stable du consommateur auprès d'un produit
- Référence à la fois à l'expérience ou la connaissance antérieure du produit et aux valeurs intérieures des individus

Méthodologie :

- *Corpus* : Avis des consommateurs en français
- *Annotation* : Selon la présence des signaux existent ou pas
- *Classification automatique (Scikit-learn)*
- *Comparaison* : Smote, Adasyn, Tomek links, Smote-TL, Modification du poids de la classe

Corpus

Collecte

- Crawl de site d'avis(beauté-test) spécialisé dans les produits de beauté

Guides d'annotation

- l'expression de l'intention d"une utilisation prolongée
- l'expression du rachat
- l'expression d'attachement très forte au niveau de l'adoption

Expressions de l'intention d'une utilisation prolongée	Expressions de rachat	Expressions de l'adoptation
ne change plus	*acheterai encore*	*adopté*
ai toujours un flacon	*rachèterais*	*adaptation*
reviens toujours	*acheter à nouveau*	*adopter*
c'est mon 4eme flacon	*reprendrai*	
plusieurs flacons	*y retourner*	

TABLE 1 – Exemples d'expressions contenant les signaux demandés

Volumétrie

Positive	Négative	Total
907	8273	9180

TABLE 2 – Statistiques sur le corpus

	Originel	Adasyn	TL	Smote	Smote+TL
nombre d'échantillons dans la classe minoritaire	907	6672	727	6559	6557
nombre total d'échantillons	9180	13231	7243	13118	13114

TABLE 3 – Changement du nombre d'échantillons

Conclusions et perspectives

Conclusions

- Complexité concernant la pré-définition des règles d'annotation (notion abstraite)
- Déséquilibre du corpus
- Évaluation sur 5 algorithmes souvent utilisés en classification asymétrique

Perspectives

- Mise en place d'autres méthodes dédiées à la classification déséquilibrée
- Évaluation sur d'autres modèles d'apprentissage comme réseaux de neurones
- Conception d'une mesure d'évaluation adaptée au besoin du client
- Essai d'autres méthodes d'extraction de caractéristiques

Méthodes

Approches au niveau des données

- **Smote** : Cette méthode de sur-échantillonnage se concentre sur la classe minoritaire, qui est augmentée en créant des exemples «synthétiques». La donnée générée n'est jamais un double exact de l'un de ses parents.
- **Adasyn** (Adaptive synthetic sampling) : L'approche améliore l'apprentissage par rapport aux distributions de données de deux façons : elle réduit le biais introduit par le déséquilibre des classes et déplace de façon adaptative la limite de classification à l'égard des exemples difficiles à apprendre.
- **Tomek links** Les données qui vont être les plus problématiques pour la plupart des algorithmes de classification seront supprimées
- **Smote-TL** : L'approche Smote-TL est la combinaison des algorithmes Smote et Tomek Links. Elle a d'abord été utilisée pour améliorer la classification des exemples sur le problème de l'annotation des protéines en bioinformatique.

Approche algorithmique

L'apprentissage sensible aux coûts tient compte des coûts associés aux exemples mal classés selon leur proportion dans les données. Cette méthode cible le problème en utilisant des matrices de coûts qui décrivent les coûts de classification erronée.

Classification

Démarche

- Classes
  - Nombre de classes : 2 classes
- Paramètres d'apprentissage
  - Nombres de mots pour chaque commentaire (1 - 2514)
  - Tokenization et Suppression des mots vides
  - Classifieur : SVM
  - Noyau : Linéaire
  - Optimisation du modèle sont faites en condition de la f-mesure
  - Poids de classe pour la méthode algorithmique : 2
- Parmi toutes les méthodes meilleurs résultats avec apprentissage sensible aux coûts

Méthode d'évaluation

- Précision, rappel et f-mesure
- Dans l'application métier, l'objective métier ou la demande du client qui décide du choix de modèle

Résultats

	Précision	Rappel	F-mesure
SVM	59.75%	62.50%	61.09%
SVM+opt	58.56%	67.76%	63%
SVM+Smote	50.41%	81.85%	62.31%
SVM+S-TL	53.42%	76.79%	63.07%
SVM+Adasyn	39.02%	**84.21%**	53.33%
SVM+TL	**66.67%**	60.53%	63.45%
SVM+couts	63.20%	67.76%	65.40%
SVM+couts+opt	60.20%	77.63%	**67.82%**

TABLE 4 – Tableau de résultats

Discussion

- Un cas du déséquilibre relatif
- Inconvénient essentiel du sur-échantillonnage : le problème de sur-apprentissage risque d'apparaître et l'augmentation du temps d'apprentissage
- Inconvénient du sous-échantillonnage : l'élimination des données potentiellement utiles
- Performance du modèle dépend du corpus