Approche fondée sur les motifs disjonctifs pour l'apprentissage des règles de classification

Amel Hidouri*, Said Jabbour **, Badran Raddaoui***, Ahmed Samet*

*ICube CNRS UMR 7357, Strasbourg, France
**CRIL, Université d'Artois & CNRS, Lens, France
***SAMOVAR, Télécom SudParis, Institut Polytechnique de Paris, France

1 Génération de règles de classification basées sur les k-GDI

Dans ce travail, nous explorons l'apprentissage de règles de décision à partir de données à l'aide de techniques de fouille de motifs à travers le modèle k-GDI (motif disjonctif généralisé) pour dériver des règles de classification optimales à partir de données d'entrée. L'objectif de ce processus d'apprentissage est de créer une caractérisation qui produit un ensemble de k règles. Ces règles doivent couvrir efficacement la majorité des instances dans la classe cible tout en étant minimales dans la deuxième classe. Pour cela, il est important de noter que chaque ensemble de données binaires peut être représenté comme une base de données transactionnelle en modélisant chaque attribut booléenne x avec deux items x pour $x = 1$ et $\overline{x}$ pour $x = 0$. D'un point de vue de fouille de motifs, nous montrons que le calcul des règles de classification est équivalent à trouver un ensemble de motifs disjoints généralisés qui couvrent presque toutes les transactions de la classe positive, tout en étant peu fréquents dans l'autre classe.

Avant d'introduire notre formulation basée sur SAT pour le calcul des règles de classification à partir des données d'entrée en utilisant la fouille de k-GDI, nous proposons deux propriétés du k-GDI : *irréductibilité* et *générateur minimal*.

Trouver un ensemble de règles de classification optimales de taille au plus k à partir des données d'entrée est équivalent à identifier les k-GDI dans la base de données transactionnelle correspondante qui minimise la fonction objectif suivante :

$$\sum_{Y \in [X]} |Y| + \lambda(|\mathcal{D}^+ \setminus \texttt{Cover}([X], \mathcal{D}^+)| + |\texttt{Cover}([X], \mathcal{D}^-)|) \quad (1)$$

L'ensemble des k-GDI qui satisfait la fonction objectif 1 présente certaines caractéristiques. En fait, pour satisfaire la fonction 1, le k-GDI doit être petit et irréductible (c'est-à-dire que $\sum_{Y \in [X]} |Y|$ est aussi précis que possible). Nous appellerons ce motif un k-GDI **optimal**. Pour qu'un k-GDI donné $[X]$ soit optimal, chaque motif Y dans $[X]$ doit correspondre de manière unique à au moins une transaction dans la sous-base de données $\mathcal{D}^+$, c'est-à-dire être le seul à satisfaire un exemple de la classe positive. Plus formellement :

Proposition 1. *Étant donné une base de données $\mathcal{D}$, $[X]$ est un k-GDI optimal dans $\mathcal{D}$ si et seulement si $[X]$ est un k-GDI irréductible dans $\mathcal{D}^+$.*

La deuxième condition est la propriété de *générateur minimal*, qui est essentielle pour répondre à l'exigence d'optimalité du k-GDI.

Proposition 2. *Si $[X]$ est un k-GDI optimal dans $\mathcal{D}$, alors pour tout $Y \in [X]$, Y est un générateur minimal dans $\mathcal{D}^-$.*

Étant donné une donnée d'entrée $\mathcal{D}$, le calcul de l'ensemble des règles de classification optimales de taille k à partir de $\mathcal{D}$ nécessite le calcul de tous les k-GDI optimaux dans $\mathcal{D}$. Pour ce faire, nous proposons d'encoder le problème de fouille des k-GDI comme un problème de MaxSAT pondéré. Pour cela, nous définissons deux ensembles de variables Booléennes : $p_{a,i}$ pour indiquer que l'item a est considéré dans le i^{th} motif. De plus, pour encoder la couverture du i^{th} motif, nous utilisons un ensemble de variables $q_{i,j}$ indiquant que le i^{th} motif est couvert par la j^{th} transaction. À l'aide de ces variables, nous introduisons ensuite un ensemble de contraintes logiques pour trouver l'ensemble des k-GDIs optimaux dans un ensemble de données donné comme montré dans la Figure 1

$$\bigwedge_{i=1}^{k} \bigwedge_{a,\bar{a}\in\Omega} (\neg p_{a,i} \vee \neg p_{\bar{a},i}) \quad (2) \qquad \bigwedge_{T_i\in\mathcal{D}} (\bigwedge_{j=1}^{k} (\neg q_{i,j} \leftrightarrow (\bigvee_{a\notin T_i} p_{a,j}) \wedge \bigvee_{a\in\Omega} p_{a,j})) \quad (3)$$

$$\bigwedge_{j=1}^{k} (\bigvee_{T_i\in\mathcal{D}^+} (q_{i,j} \wedge \bigwedge_{l=1|l\neq j}^{k} \neg q_{i,l})) \quad (4) \qquad \bigwedge_{j=1}^{k} (\sum_{T_i\in\mathcal{D}^+} (q_{i,j} \wedge \bigwedge_{l=1|l\neq j}^{k} \neg q_{i,l}) \geq \alpha) \quad (5)$$

$$\bigwedge_{j=1}^{k} (q_{i,j} \implies \neg q_{j,k}) \quad (6) \qquad \sum_{j=1}^{k} (q_{i,j}) \leq k \quad (7)$$

FIG. 1 – *Encodage basé sur SAT pour la découverte des K-GDI*

Les expérimentations sont présentées dans l'article original, où une comparaison avec MLIC Malioutov et Meel (2018), IMLI Ghosh et Meel (2019), ainsi qu'avec d'autres classifieurs, a démontré l'efficacité, l'efficience et l'évolutivité de notre approche par rapport à l'état de l'art sur divers jeux de données.

Remerciements

Ce travail est partiellement financé de l'Agence Nationale de la Recherche ANR au titre de la subvention HYCI : 'ANR-22-CE55-0010' et du projet 'ANR-22-CE92-0007-02'.

Références

Ghosh, B. et K. S. Meel (2019). Imli : An incremental framework for maxsat-based learning of interpretable classification rules. In *AAAI/ACM Conference on AI*, pp. 203–210.

Malioutov, D. et K. S. Meel (2018). MLIC : A maxsat-based framework for learning interpretable classification rules. In *CP*, pp. 312–327.