

Nouveau Modèle de Sélection de Caractéristiques basé sur la Théorie des Ensembles Approximatifs pour les Données Massives

Zaineb Chelly Dagdia^{*,**} Christine Zarges^{*}
Gaël Beck^{***}, Mustapha Lebbah^{***}

^{*}Department of Computer Science, Aberystwyth University, United Kingdom

^{**}LARODEC, Institut Supérieur de Gestion de Tunis, Tunisia

{zaineb.chelly, c.zarges}@aber.ac.uk,

^{***}Computer Science Laboratory (LIPN), University Paris-North-13, Villetaneuse, France

{beck, mustapha.lebbah}@lipn.univ-paris13.fr

1 Modèle de Sélection de Caractéristiques

Notre modèle, Sp-RST, définit le problème d'apprentissage des données massives comme un système d'information T_{RDD} où l'univers $U = \{x_1, \dots, x_N\}$ est l'ensemble des objets, $C = \{c_1, \dots, c_V\}$ est l'ensemble des caractéristiques conditionnelles parmi lesquelles Sp-RST sélectionne les caractéristiques les plus pertinentes et la caractéristique de décision $D = \{d_1, \dots, d_W\}$ correspond à la classe. Afin d'assurer la scalabilité de notre algorithme, Sp-RST partage la T_{RDD} en m blocs de données basés sur des partitions de C . Par conséquent, $T_{RDD} = \bigcup_{i=1}^m (C_r)T_{RDD(i)}$; où $r \in \{1, \dots, V\}$. Chaque $T_{RDD(i)}$ est construit en fonction de r caractéristiques aléatoirement sélectionnées à partir de C ; où $\forall T_{RDD(i)} : \exists \{c_r\} = \bigcap_{i=1}^m T_{RDD(i)}$. De ce fait, au lieu d'appliquer Sp-RST (voir Algorithme 1) sur la T_{RDD} comprenant l'ensemble des attributs C , l'algorithme distribué sera appliqué sur chaque $T_{RDD(i)}$.

Nous appliquons l'algorithme N fois sur les m blocs de la T_{RDD} . Plus précisément, à travers toutes les itérations, l'algorithme générera d'abord les m $T_{RDD(i)}$, ensuite pour chaque partition, les instructions distribuées de Sp-RST (Algorithme 1, lignes 5 à 7) seront exécutées à part la ligne 1 de l'algorithme. Cette tâche est indépendante des m partitions générées vu qu'elle calcule la relation d'indiscernabilité de la caractéristique de la décision $IND(D)$ et elle est non liée aux caractéristiques conditionnelles. En dehors de la boucle, ligne 9, le résultat de chaque partition est soit un seul reduct $RED_{i(D)}(C_r)$ ou un ensemble de reducts $RED_{i(D)}^F(C_r)$. Si Sp-RST génère un seul reduct, pour une partition $T_{RDD(i)}$ alors la sortie de cette phase de sélection de caractéristiques est l'ensemble des caractéristiques de $RED_{i(D)}(C_r)$. Ces caractéristiques sont les plus pertinentes parmi l'ensemble C_r et résultent à un nouveau système d'information $T_{RDD(i)}, T_{RDD(i)}(RED)$, qui préserve quasiment la même qualité des données que $T_{RDD(i)}(C_r)$ qui est basé sur tout l'ensemble des caractéristiques C_r . Si Sp-RST génère une famille de reducts alors l'algorithme choisit aléatoirement un reduct pour représenter la $T_{RDD(i)}$. À ce stade, à chaque bloc de données i correspond un ensemble de caractéristiques sélectionnées $RED_{i(D)}(C_r)$. Cependant, puisque chaque $T_{RDD(i)}$