Learning classification rules for telecom customer call data under concept drift

M. Black, R. Hickey

Le cas d’application est de garder à jour le profil des clients s’enregistrant à un service.

En traitant le timestamp comme un attribut à deux modalités, on va voir l’importance de celui-ci dans les règles établies des arbres de décisions. Une famille de d’algorithmes est présentée avec CD3, 4 et 5. Le principe de base de ces algorithmes est que l’on dispose d’une base d’entrainement et d’un arbre valide. On va ensuite recevoir de nouvelles observations, on leurs ajoutent la variable timestamp avec comme modalité « nouveau ». Les observations d’entrainements ont comme modalité « courant ». On entraine l’arbre. En fonction du rôle de la variable qu’a le timestamp on actualise les règles décisions actuelles ou non. La différence entre CD3, 4 et 5 est dans comment l’on actualise la base d’apprentissage. Pour CD3, il faut supprimer les observations anciennes qui ont mené à créer des règles qui ont été invalidées. CD4 n’a pas de processus de suppression d’observations non représentatives. CD5 n’utilise plus le timestamp comme attribut binaire mais comme un attribut numérique unique à chaque batch. CD4 et 5 apportent une faible amélioration des résultats à CD3.