MTCopula: Génération de données synthétiques et complexes basées sur les Copules

Fodil Benali*,**, Damien Bodénès*, Nicolas Labroche **, Cyril de Runz**

*Adwanted Group, Paris
{fbenali,dbodenes}@adwanted.com,
**BDTLN - LIFAT, University of Tours, Blois, France
{nicolas.labroche, cyril.derunz}@univ-tours.fr

Résumé. Cet article est une version courte de Benali et al. (2021) [1]. La plupart des techniques existantes de génération de données ne fonctionnent bien que pour de faibles dimensions et échouent à capturer les dépendances complexes entre les dimensions des données. L'identification de la bonne combinaison de modèles et de leurs paramètres respectifs reste un problème ouvert. Nous présentons MTCopula, une nouvelle approche de génération de données synthétiques complexes, flexible et extensible, qui choisit automatiquement le meilleur modèle de copules et les marginales les mieux ajustées pour capturer la complexité des données en se reposant sur le critère d'information d'Akaike.

1 Introduction

De nos jours, il peut être difficile d'obtenir des données de valeur et de qualité en quantité, du fait des moyens de collecte et des possibles problèmes de confidentialité, comme c'est le cas de la planification publicitaire, notre contexte industriel. Dans ce contexte, seuls de petits volumes données complexes et de haute qualité (multidimensionnelles, multivariées, catégorielles/continues, etc.), représentatifs de l'ensemble des données, sont disponibles pour générer un jeu de données synthétiques large et réaliste. Par conséquent, il existe un réel besoin pour un générateur de données complexes réalistes.

Notre objectif est de générer de nouvelles données qui conservent les mêmes caractéristiques que les données originales, telles que la distribution de leurs attributs et leur interdépendance, afin que tout travail effectué sur les données d'origine puisse être réalisé en utilisant les données synthétiques (Petricioli et al., 2020). Ceci ne peut pas être fait en utilisant la méthode habituelle de génération de données synthétiques unidimensionnelles car, lorsqu'elle est appliquée dans un contexte hautement dimensionnel, elle ne permet pas de modéliser la dépendance entre les variables. Pour résoudre ces problèmes, plusieurs travaux récents se sont concentrés sur des approches d'apprentissage profond comme le Generative Adversarial Network (GAN), mais ces approches nécessitent une grande quantité de données pour l'étape d'apprentissage et ne peuvent donc pas être utilisées pour notre problème.

1. http://ceur-ws.org/Vol-2840/paper8.pdf