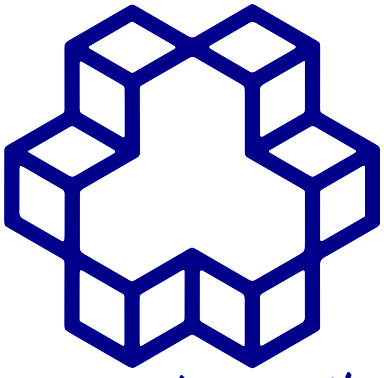


طرح پیشنهادی پروژه درس یادگیری ماشین

نام و نام خانوادگی	علی مهرابی
شمارهٔ دانشجویی	۴۰۲۲۳۸۵۴
استاد درس	دکتر علیاری
تاریخ	خرداد ۱۴۰۳

۱ عنوان پروژه

توسعه یک مدل پیش بینی خرید مشتری، بر پایه GAN و Clustering

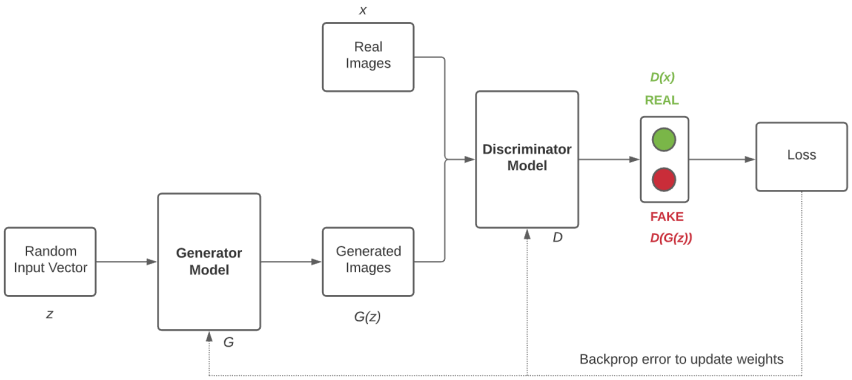
۲ شرح مختصر و نوآوری

ریزش مشتریان [۱] در شرایط غیر قراردادی [۲] به عنوان یکی از مشخصه های میزان سلامت یک کسب و کار است. تخمین اینکه از بین مشتری های حال حاضر یک کسب و کار، کدامیک در ماه آینده نیز مشتری باقی خواهد ماند یکی از اهدافی است که به کمک الگوریتم های هوش مصنوعی قابل دستیابی است.

مسئله ای که در تمامی مدل ها در این زمینه با آن مواجه هستند، نامتعادل [۳] بودن داده است. به طوری که بسته به دیتاست در دسترس، نرخ نامتعادل بودن می تواند به ۱ درصد نیز برسد و این میزان نامتعادلی، کار پیش بینی را با چالش های جدی مواجه می کند.

الگوریتم های oversampling مانند SMOTE یکی از اولین راه حل هایی است که در این موارد استفاده می شود. اما استفاده از این تکنیک برای مقابله با عدم توازن داده ها به دلیل عدم درک صحیح از ساختار های زمانی و عدم شفافیت در تولید داده، ممکن است موجب تولید داده مصنوعی نامعتبر و کاهش عملکرد مدل گردد.

شبکه های مولد متخاصم یا GANs یک تکنیک مبتنی بر شبکه های عصبی عمیق [۴] است که در سال ۲۰۱۴ توسط Ian Goodfellow معرفی شد و توجهات زیادی را تا به امروز به خود جذب کرده. ساختمان کلی این الگوریتم در شکل ۱ توضیح داده شده است.



شکل ۱: ساختمان کلی مدل GAN

هدف ما در این پروژه توسعه و ادغام قابلیت داده سازی GANs در کنار خوشه بندی [۵] مشتریان و مقایسه نتایج با دیگر الگوریتم های upsampling مانند SMOTE، ADASYN، Borderline SMOTE خواهد بود. در این پروژه از دیتاست مربوط به ریزش مشتریان که یک دیتاست imbalanced است استفاده می شود.

[1] Customer Churn
[2] Non-Contractual Setting
[3] Imbalanced
[4] Deep Neural Network
[5] Clustering

۳ کارهای انجام شده در این زمینه

با توجه به جدید بودن مبحث GAN و همینطور توجه به این موضوع که کاربرد اصلی این شبکه در حوزه پردازش تصویر و موارد مربوط به این حوزه تلقی می شود، تمامی مراجع مربوط به ادغام مدل پیش بینی خرید مشتری و GAN مربوط به چندسال اخیر می شوند.

در [۱] که مربوط به سال ۲۰۲۳ است، با هدف پیش بینی خروج دانشجویان از collage دو الگوریتم SMOTE و GAN برای بالانس کردن داده مقایسه شده اند که در نهایت مشخص می شود GAN توانسته بهبود هایی را در Precision، Recall و F1-score بدهد.

در [۲] نیز، با هدف بهبود متریک های متداول در حوزه پیش بینی و با تمرکز بر روی داده های حوزه پزشکی، عملکرد دو تکنیک SMOTE و متد های مبتنی بر GAN (به طور دقیق تر Conditional GAN و Conditional Tabular GAN) مقایسه می شود. دیتاست مورد استفاده در این مقاله شامل داده های categorical و numerical است. علاوه بر این این مقاله از متد های SVM، LR، RFو MLP به عنوان الگوریتم های طبقه بند استفاده می کند و در نتیجه آمده است که مدل های مبتنی بر GAN نسبت به سایر مدل های متداول بهتر عمل می کند.

در مرجع [۳] به اثرات استفاده از GAN در بهبود عملکرد مدل طبقه بند در دیتاست balanced و imbalanced پرداخته می شود. سه دیتاست استفاده شده در این مقاله شامل دیتاست ارزیابی ماشین که highly imbalanced تلقی شده، دیتاست شناخت فعالیت انسان که not highly imbalanced تلقی شده و دیتاست مربوط به یک بانتک در پرتغال که یک دیتاست بالانس است می شود. در این مقاله از Conditional GAN استفاده شده است و ارزیابی روی مدل DT گزارش شده که در هر سه مورد موفق به بهبود دقت مدل شده است.

در [۴] یک مدل هایبرید پیشنهاد شده است که تکنیک WGAN و یک لایه undersampler را پشت هم قرار می دهد تا مشکل imbalance بودن داده را رفع کند. در نهایت با پیاده سازی این الگوریتم روی دو دیتاست که یکی ساختگی و دیگری مربوط به مشتریان واقعی است، بهبود هایی در متریک های ذکر شده مشاهده شده است.

در [۵] به کمک تلفیق GAN و Clustering سعی شده که مشکل پیش بینی به کمک داده های imbalanced رفع گردد. در این مقاله به مشکل class overlap هم پرداخته می شود. در الگوریتم معرفی شده در این مقاله ابتدا داده ها خوشه بندی می شوند، سپس با توجه به خوشه های مختلف عمل undersampling روی داده ها انجام می شود و سپس یک الگوریتم TWGAN-GP پیاده شده تا داده های مصنوعی روی کلاس اقلیت تولید شود. این الگوریتم روی ۲۲ دیتاست پیاده شده و نشان داده می شود که باعث بهبود AUC و F1-score می شود.

۴ منابع

[1] Park, J., Kwon, S., Jeong, S.-P. (2023). A study on improving turnover intention forecasting by solving imbalanced data problems: Focusing on SMOTE and generative adversarial networks. Journal of Big Data, 10(36).

[2] Eom, G., Byeon, H. (2023). Searching for optimal oversampling to process imbalanced data: Generative adversarial networks and synthetic minority over-sampling technique. Mathematics, 11(3605)

[3] Ayoub, S., Gulzar, Y., Rustamov, J., Jabbari, A., Reegu, F. A., Turaev, S. (2023). Adversarial approaches to tackle imbalanced data in machine learning. Sustainability, 15(7097).

[4] Zhu, B., Pan, X., vanden Broucke, S., Xiao, J. (2022). A GAN-based hybrid sampling method for imbalanced customer classification. Information Sciences, 609, 1397-1411.

[5] Ding, H., Cui, X. (2023). A clustering and generative adversarial networks-based hybrid approach for imbalanced data classification. Journal of Ambient Intelligence and Humanized Computing, 14(8003–8018)