要旨

本研究は、プライバシー規制や共有制限により実データの活用が難しい状況に対し、意味的一貫性と下流有用性を両立する表形式合成データ生成手法 KGSynX を提案する。既存の GAN・拡散・LLM 生成は周辺分布や低次統計の一致に偏重し、列間相互作用や領域固有の論理制約を保持しにくい。KGSynX は、レコードをエンティティ、属性値を属性ノードとする知識グラフを構築し、Node2Vec で抽出した構造埋め込みを LLM のプロンプトへ注入する。知識グラフはスキーマ水準（型・値域・階層）とインスタンス水準（外部キー、条件付き依存、禁止組合せ）を同時に表現し、生成過程の制約として機能する。これにより低頻度カテゴリや外れ値の過剰・過少生成を抑え、局所的なモード崩壊を軽減する。さらに、実データ、合成データで学習した分類器の SHAP 重要度差を意味的整合性の指標とし、乖離の大きい特徴に着目した指示を自動生成するフィードバックでプロンプトを反復改良する。評価では、分布距離や相関構造の保存度も併せて検証した。UCI 心臓病、企業請求書履歴、通信顧客解約の 3 データセットを用い、TSTR（合成データで学習、実データで評価）で検証した結果、精度・F1・AUC が CTGAN、TabDDPM、LLM、LLM+KG を上回り、SHAP 帰属ギャップも縮小した。品質評価は三軸：(1)各特徴量の周辺分布差を KL で定量化、(2)PCA 可視化でクラスタ構造などを確認、(3)実データ・合成データで学習した分類器の SHAP 重要度一致性で意思決定ロジックの保持を検証。以上より、知識グラフと説明可能なフィードバックの統合は、統計的類似性と有用性を同時に高める有効な手法であることが分かった。本手法の新規性は、構造的事前知識の顕在化と説明可能性駆動の閉ループ最適化を単一枠組で結合し、合成データの「使える品質」を定量的に高めた点にある。