Üniversite : T.C. İstanbul Kültür Üniversitesi
Enstitü : Lisansüstü Eğitim Enstitüsü
Anabilim Dalı : Bilgisayar Mühendisliği
Program : Bilgisayar Mühendisliği
Tez Danışmanı : Doç. Dr. Fatma Patlar AKBULUT
Tez Türü ve Tarihi : Yüksek Lisans – Eylül 2023

ÖZET

Veri hacimleri katlanarak artmasından dolayı, yüksek boyutlu veri setleri, analiz ve modellemenin yanı sıra görselleştirme ve yorumlama adımlarındaki performanslarında zorluklar ortaya çıkmaya başladı. Bu zorluğun bir sonucu olarak, onu tanımlamak için "boyutsallığın laneti" türetilmiştir. Çeşitli öğrenme yöntemleri, içsel bağlantılarını korurken veri setlerinin boyutlarını azaltarak boyutsallığın lanetini hafifletmeyi amaçladıkları için bu çalışmanın konusudur. Çoklu öğrenme algoritmalarının boyutsallığı düşürmedeki başarısına ilişkin çalışmalar cesaret verici olsa da bu çalışmanın büyük çoğunluğu ya hayali ya da sayısal veri setleriyle yürütülmüştür. Sonuç olarak, bu algoritmaların kategorik verilerle kullanım için uyarlanmasına yönelik araştırmalarda çok büyük bir boşluk vardır. Bu boşluğu doldurmak için, bu tezde önce kategorik veri kümeleri üzerinde öne çıkan manifold öğrenme algoritmalarını seçilmiş ve değerlendirilmiş, ardından sonuçlar derinlemesine analiz edilerek sunulmuştur. Manifold öğrenme teknikleri Çekirdek Temel Birleşen Analizi, İzomap, Lokal Doğrusal Gömme (LLE) (diğer 3 varyantı ile), t-Dağıtılmış Stokastik Komşu Gömme (t-SNE) ve Standart Manifold Yaklaşımı ve Projeksiyonu (UMAP) bu tez için seçilmiştir. Araştırmacılar, bir dizi farklı alandaki başarıları nedeniyle bu algoritmalara çok dikkat etmişlerdir. Bu çalışma, kategorik veriler bağlamında faydalarını keşfederek, boyut indirgeme yöntemlerine ilişkin artan bilgi birikimine katkıda bulunmaktadır. Bu çeşitli öğrenme algoritmalarının, kategorik veri setlerindeki temel yapıyı ve ilişkileri ne kadar iyi yakaladığını ve koruduğunu incelemek, bu yöntemleri değerlendirmenin önemli bir parçasıdır. Performansı değerlendirmek için çeşitli değerlendirme ölçüm teknikleri kullanılmıştır. Bulgular, kategorik verilere uygulandığında manifold öğrenme algoritmalarının güçlü yanlarını ve sınırlamalarını ortaya çıkararak, çeşitli veri türleriyle uğraşan araştırmacılar için yararlı bilgiler sağlamayı hedefler. Bu çalışmadan elde edilen bulgular, boyutluluk indirgeme yöntemlerinin kullanımını geliştirerek, yüksek boyutlu kategorik veri kümelerinin daha doğru modellenmesine ve yorumlanmasına yol açacaktır.

Anahtar Kelimeler: Boyut İndirgeme, Manifold Öğrenme, Çekirdek Temel Birleşen Analizi, İzomap, Lokal Doğrusal Gömme, t-Dağıtılmış Stokastik komşu Gömme, Standart Manifold Yaklaşımı ve Projeksiyonu

University : İstanbul Kültür University
Institute : Institute of Graduate Studies
Department : Computer Engineering
Program : Computer Engineering
Supervisor : Assoc. Prof. Fatma Patlar AKBULUT
Degree Awarded and Date : Master's degree – September 2023

ABSTRACT

High-dimensional datasets have presented difficulties in terms of performance during analysis and modeling, as well as visualization and interpretation, as data volumes have increased at an exponential rate. As a result of this difficulty, the curse of dimensionality has been coined to describe it. Manifold learning methods are the topic of this study since they seek to alleviate the curse of dimensionality by decreasing the data sets' dimensions while maintaining their inherent linkages. While studies on the effectiveness of manifold learning algorithms in lowering dimensionality have been encouraging, most of this work has been conducted with either imagine or numerical datasets. As a result, there is an enormous gap in the research on adapting these algorithms for use with categorical data. To fill this void, this thesis first selects and evaluates prominent manifold learning algorithms on categorical datasets, then analyzes and discusses these results in depth. Manifold learning techniques Kernel PCA, Isomap, Locally Linear Embedding (LLE) (and its 3 variants) -, t-Distribution Stochastic Neighbor Embedding (t-SNE), and Uniform manifold approximation and projection (UMAP) were selected for this investigation. Researchers have paid close attention to these algorithms because of their success in several different fields. This study contributes to the growing body of knowledge regarding dimensionality reduction methods by exploring their utility in the context of categorical data. Examining how well these manifold learning algorithms capture and retain the underlying structure and relationships in categorical datasets is an essential part of evaluating these methods. Different evaluation measurement techniques are utilized to evaluate performance. The findings provide useful insights for researchers dealing with varied data types by revealing the strengths and limitations of manifold learning algorithms when applied to categorical data. The findings from this study will enhance the use of dimensionality reduction methods, leading to more accurate modeling and interpretation of high-dimensional categorical datasets.

Keywords: Dimensionality Reduction, Manifold Learning, Kernel PCA, Isomap, Locally Linear Embedding (LLE), t-Distribution Stochastic Neighbor Embedding (t-SNE), Uniform manifold approximation and projection (UMAP)