GRADUATE SCHOOL OF ENGINEERING & SCIENCE
AGÜ

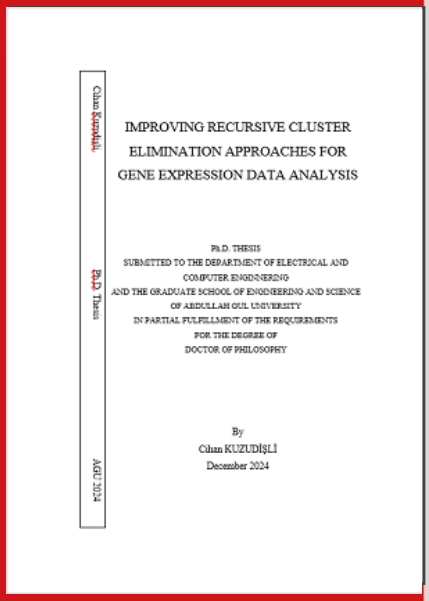
Electrical and Computer Engineering

A Ph.D. Thesis

Cihan KUZUDİŞLİ

cihan.kuzudisli@agu.edu.tr

0000-0003-4774-152X



Thesis Advisor

Assoc. Prof. Burcu BAKIR GÜNGÖR

burcu.gungor@agu.edu.tr

16.01.2025

Improving recursive cluster elimination approaches for gene expression data analysis

abstract The computational and interpretational difficulties caused by the ever-increasing dimensionality of biological data generated by new technologies pose a significant challenge. Feature selection (FS) methods aim to reduce the dimension, and feature grouping has emerged as a foundation for FS techniques that seek to detect strong correlations among features and identify irrelevant features. In this thesis, methods that utilize feature grouping in a supervised context were developed. We initially tested the effects of different clustering algorithms on SVM-RCE and observed the best performance with K-means. In the first developed method, Recursive Cluster Elimination with Intra-cluster Feature Elimination (RCE-IFE), both cluster and intra-cluster elimination is performed recursively in each cluster reduction step. Our experimental findings imply that RCE-IFE provides robust classifier performance and significantly reduces feature size while maintaining feature relevance and consistency. In the second developed Grouping – Scoring – Model (G-S-M) based study, G-S-M_Rep, we use prior knowledge to form disease groups, and select top features as representative of each group. These representative features are learnt by the model in a cumulative manner. Results show that G-S-M_Rep attains satisfactory model performance with a small number of features. Consequently, this thesis presents methods based on feature grouping and focuses on improving feature reduction capability, classification performance, feature relevancy, and feature consistency.

keywords Feature Grouping, Machine Learning, Recursive Cluster Elimination, Intra-cluster Feature Elimination

özet Yeni teknolojilerle üretilen biyolojik verilerin giderek artan boyutluluğunun neden olduğu hesaplama ve yorumlama güçlükleri önemli bir zorluk oluşturmaktadır. Özellik seçimi (FS) yöntemleri boyutu azaltmayı amaçlar ve özellik gruplaması, özellikler arasında güçlü korelasyonları tespit etmeyi ve ilgisiz özellikleri belirlemeyi amaçlayan FS teknikleri için bir temel olarak ortaya çıkmıştır. Bu tezde, gözetimli bir bağlamda özellik gruplandırmasını kullanan yöntemler geliştirilmiştir. Başlangıçta farklı kümeleme algoritmalarının SVM-RCE üzerindeki etkilerini test ettik ve K-means ile en iyi performansı gözlemledik. Geliştirilen ilk yöntem olan Öbek İçi Özellik Eleme ile Yinelemeli Öbek Eleme (RCE-IFE) yönteminde, her öbek azaltma adımında hem öbek hem de öbek içi eleme yinelemeli olarak gerçekleştirilir. Deneysel bulgularımız, RCE-IFE'nin güçlü bir sınıflandırıcı performansı sağladığını ve özellik ilgisini ve tutarlılığını korurken özellik boyutunu önemli ölçüde azalttığını göstermektedir. İkinci geliştirilen Gruplama – Puanlama – Model (G-S-M) tabanlı çalışma olan G-S-M_Rep'de, hastalık gruplarını oluşturmak için ön bilgileri kullanıyoruz ve her grubu temsil edecek en iyi özellikleri seçiyoruz. Bu temsili özellikler model tarafından kümülatif bir şekilde öğrenilir. Sonuçlar G-S-M_Rep'in az sayıda özellikle tatmin edici bir model performansına ulaştığını göstermektedir. Sonuç olarak, bu tez özellik gruplandırmaya dayalı yöntemleri sunmakta ve özellik azaltma yeteneğini, sınıflandırma performansını, özellik alaka düzeyini ve özellik tutarlılığını iyileştirmeye odaklanmaktadır.

anahtar kelime Özellik Gruplandırma, Makine Öğrenimi, Yinelemeli Öbek Eliminasyonu, Öbek İçi Özellik Eleme