T.C.
YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

BÜYÜK VERİDE ÇAPRAZ ENTROPİYE DAYALI KAYIP FONKSİYONU YARDIMIYLA DEĞİŞKEN SEÇİMİ

Hıncal TOPÇUOĞLU

DOKTORA TEZİ
İstatistik Anabilim Dalı
İstatistik Programı

Danışman
Doç. Dr. Atıf Ahmet EVREN

Mart, 2026

T.C.
YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

BÜYÜK VERİDE ÇAPRAZ ENTROPİYE DAYALI KAYIP FONKSİYONU YARDIMIYLA DEĞİŞKEN SEÇİMİ

Hıncal TOPÇUOĞLU tarafından hazırlanan tez çalışması 24.03.2026 tarihinde aşağıdaki jüri tarafından Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü İstatistik Anabilim Dalı İstatistik Programı **DOKTORA TEZİ** olarak kabul edilmiştir.

Doç. Dr. Atıf Ahmet EVREN
Yıldız Teknik Üniversitesi
Danışman

Jüri Üyeleri

Doç. Dr. Atıf Ahmet EVREN, Danışman
Yıldız Teknik Üniversitesi _____

Dr. Öğr. Üyesi Doğan YILDIZ, Üye
Yıldız Teknik Üniversitesi _____

Prof. Dr. Özlem YORULMAZ, Üye
İstanbul Üniversitesi _____

Doç. Dr. Gülder KEMALBAY, Üye
Yıldız Teknik Üniversitesi _____

Doç. Dr. Nihat TAK, Üye
Marmara Üniversitesi _____

Danışmanım Doç. Dr. Atıf Ahmet EVREN sorumluluğunda tarafımca hazırlanan BÜYÜK VERİDE ÇAPRAZ ENTROPİYE DAYALI KAYIP FONKSİYONU YARDIMIYLA DEĞİŞKEN SEÇİMİ başlıklı çalışmada veri toplama ve veri kullanımında gerekli yasal izinleri aldığımı, diğer kaynaklardan aldığım bilgileri ana metin ve referanslarda eksiksiz gösterdiğimi, araştırma verilerine ve sonuçlarına ilişkin çarpıtma ve/veya sahtecilik yapmadığımı, çalışmam süresince bilimsel araştırma ve etik ilkelerine uygun davrandığımı beyan ederim. Beyanımın aksinin ispatı halinde her türlü yasal sonucu kabul ederim.

Hıncal TOPÇUOĞLU

İmza

Canım kızım Ada'ya..

TEŞEKKÜR

Bu tezin hazırlanması sürecinde bilgi ve tecrübeleriyle bana yol gösteren değerli danışmanım Doç. Dr. Atıf Ahmet EVREN'e en içten teşekkürlerimi sunarım. Ayrıca, çalışmalarıma değerli katkılar sunan jüri üyelerime ve bölümümdeki tüm hocalarıma teşekkür ederim.

Hıncal TOPÇUOĞLU

İÇİNDEKİLER

SİMGE LİSTESİ **ix**

KISALTMA LİSTESİ **xi**

ŞEKİL LİSTESİ **xiii**

TABLO LİSTESİ **xiv**

ÖZET **xv**

ABSTRACT **xvii**

1 GİRİŞ **1**

1.1 Yüksek Boyutlu Verinin Zorluğu 1

1.2 Öznitelik Seçimi: Kritik Bir Ön İşleme Adımı 1

1.3 Yeni Bir Yaklaşım: Öznitelik Seçimi İçin Öğrenilebilir Dikkat (LAFS) 3

1.4 Katkılar ve Tez Taslağı 3

2 LİTERATÜR TARAMASI **5**

2.1 Öznitelik Seçiminin Genel Görünümü 5

2.2 Filtre Yöntemleri: İstatistiksel Saflık ve Hız 5

2.3 Sarmalayıcı Yöntemler: Performans Arayışı 6

2.4 Gömülü Yöntemler: Bütünleşik Seçim 7

2.5 Tablosal Veri İçin Derin Öğrenme Devrimi 7

2.6 Derin Öğrenmede Bilgi-Kuramsal İlkeler 7

3 ÖZNİTELİK SEÇİMİ YÖNTEMLERİNİN MATEMATİKSEL TEMELLERİ **9**

3.1 Öznitelik Seçimi Paradigmalarına Giriş 9

3.2 Filtre Yöntemleri: İlgililik ve Gereksizliğin İstatistiksel Ölçüleri . . 9

3.2.1 Karşılıklı Bilgi (Mutual Information - MI) 9

3.2.2 Minimum Gereksizlik Maksimum İlgililik (mRMR) 10

3.3 Sarmalayıcı Yöntemler: Arama Tabanlı bir Yaklaşım 11

3.3.1 Genel Formülasyon 12

3.3.2 Özyinelemeli Öznitelik Eleme (Recursive Feature Elimination - RFE) . . . 12
3.4 Gömülü Yöntemler: Bütünleşik Seçim . . . 13
3.4.1 LASSO (L1 Düzenlileştirmesi) . . . 13

4 ÖZNİTELİK SEÇİMİ İÇİN BİLGİ-KURAMSAL BİR ARAÇ SETİ 15
4.1 Makine Öğrenmesinde Bilgi Kuramının Rolü . . . 15
4.2 Temel: Shannon Entropisi . . . 16
4.2.1 Shannon Entropisinin Tanımı . . . 16
4.2.2 Shannon Entropisinin Özellikleri . . . 16
4.3 Bileşik ve Koşullu Entropi . . . 17
4.3.1 Bileşik Entropi . . . 17
4.3.2 Koşullu Entropi . . . 17
4.4 Göreceli Entropi ve Karşılıklı Bilgi . . . 18
4.4.1 Göreceli Entropi (Kullback-Leibler Iraksaması) . . . 18
4.4.2 Jeffrey Iraksaması . . . 18
4.4.3 Karşılıklı Bilgi (Mutual Information - MI) . . . 19
4.5 Çapraz Entropi . . . 19
4.6 Shannon Entropisinin Genelleştirmeleri . . . 20
4.6.1 Rényi Entropisi . . . 20
4.6.2 Tsallis Entropisi . . . 22
4.7 Makine Öğrenmesinde Bilgi Kuramının Temel Uygulamaları . . . 23
4.7.1 Karar Ağaçlarında Bölme Kriterleri . . . 23
4.7.2 Parametre Tahmini için Amaç Fonksiyonları . . . 24
4.7.3 Bir Düzenlileştirici Olarak Entropi . . . 25
4.7.4 Aktif Öğrenmede Belirsizlik Örneklemesi . . . 26
4.8 Bölüm Sonucu . . . 26

5 LAFS ÇERÇEVESİ: TÜREVLENEBİLİR DİKKAT TABANLI BİR YAKLAŞIM 27
5.1 Gerekçe ve Tasarım Felsefesi . . . 27
5.2 Problemin Resmi Formülasyonu . . . 27
5.3 LAFS Mimarisi . . . 28
5.3.1 Dikkat Puanlayıcı Ağı (AttnNet) . . . 28
5.3.2 Sınıflandırıcı Ağı (ClassNet) . . . 29
5.4 Hibrit Bilgi-Kuramsal Kayıp Fonksiyonu . . . 29
5.4.1 Görev Kaybı (L_{task}) . . . 29
5.4.2 Bilgi-Kuramsal Düzenlileştirici (L_{info}) . . . 30
5.4.3 Tam Öğrenme Hedefi . . . 30
5.5 Global Öznitelik Öneminin Çıkarılması . . . 30

6 DENEYSEL DOĞRULAMA **32**
6.1 Bölüme Giriş ve Deneysel Amaçlar 32
6.2 Deneysel Kurulum 33
6.2.1 Karşılaştırma Algoritmalarının Detayları 33
6.2.2 Değerlendirme Metrikleri 37
6.3 Deney 1: Doğrusal Olmayan Etkileşimler Üzerindeki Performans (XOR Veri Seti) 37
6.3.1 Amaç ve Veri Üretimi 37
6.3.2 Sonuçlar ve Analiz 37
6.4 Deney 2: Yüksek Çoklu Doğrusallık Altındaki Performans (SYNTH Veri Seti) 39
6.4.1 Amaç ve Veri Üretimi 39
6.4.2 Sonuçlar ve Analiz 39
6.5 Deney 3: Gerçek Dünya Veri Setleri Üzerinde Büyük Ölçekli Karşılaştırma 40
6.5.1 Amaç ve Veri Seti Seçimi 40
6.5.2 Değerlendirme Protokolü 40
6.5.3 Sonuçlar ve Analiz 41
6.6 Hesaplamalı Verimlilik Analizi (Çalışma Zamanları) 43
6.6.1 Metodoloji 43
6.6.2 Sonuçlar ve Analiz 43
6.7 Bölüm Özeti ve Temel Bulgular 44
6.8 Hiperparametre Hassasiyeti ve Ayar Analizi 45
6.8.1 Modifikasyon Metodolojisi 45
6.8.2 Ayar Sonuçları ve Analizi 46
6.9 Entropik Düzenleyicinin (λ) Hassasiyet Analizi 48
6.9.1 Metodoloji 48
6.9.2 Sonuçlar ve Analiz 48
6.10 Farklı Ağırlık Başlatma Çalışması: Kaiming (He) Başlatmasının Etkisi 50
6.10.1 Teorik Gerekçe ve Uygulama 50
6.10.2 Karşılaştırmalı Sonuçlar ve Analiz 51

7 SONUÇ **53**
7.1 Tezin Özeti 53
7.2 Başlıca Katkılar ve Bulgular 54
7.3 Sınırlılıklar ve Gelecek Araştırma Alanları 55

KAYNAKÇA **57**

TEZDEN ÜRETİLMİŞ YAYINLAR **61**

SİMGE LİSTESİ

XGBoost	Aşırı Gradyan Artırma (eXtreme Gradient Boosting)
IB	Bilgi Darboğazı (Information Bottleneck)
L_{info}	Bilgi-kuramsal düzenleyici (Entropi) (Information-theoretic regularizer (Entropy))
FRF-fs	Bulanık Rastgele Orman Öznitelik Seçimi (Fuzzy Random Forest Feature Selection)
MLP	Çok Katmanlı Algılayıcı (Multi-Layer Perceptron)
VDR	Değişken Keşif Oranı (Variable Discovery Rate)
SVM	Destek Vektör Makineleri (Support Vector Machines)
AttnNet	Dikkat Puanlayıcı Ağı (Attention Scorer Network)
ReLU	Doğrultulmuş Doğrusal Birim (Rectified Linear Unit)
λ	Entropik düzenlileştirme için hiperparametre (Hyperparameter for entropic regularization)
LASSO	En Küçük Mutlak Büzülme ve Seçim Operatörü (Least Absolute Shrinkage and Selection Operator)
L_{task}	Görev performansı kaybı (Çapraz Entropi) (Task-performance loss (Cross-Entropy))
MI	Karşılıklı Bilgi (Mutual Information)
$D_{KL}(P\|\|Q)$	Kullback-Leibler Iraksaması (Kullback-Leibler Divergence)
MLE	Maksimum Olabilirlik Kestirimi (Maximum Likelihood Estimation)
mRMR	Minimum Gereksizlik Maksimum İlgililik (Minimum Redundancy Maximum Relevance)
MSE	Ortalama Kare Hata (Mean Squared Error)
LAFS	Öznitelik Seçimi için Öğrenilebilir Dikkat (Learnable Attention for Feature Selection)

RFE	Özyinelemeli Öznitelik Eleme (Recursive Feature Elimination)
AUC	ROC Eğrisi Altındaki Alan (Area Under the ROC Curve)
$H(X)$	Shannon Entropisi (Shannon Entropy)
ClassNet	Sınıflandırıcı Ağı (Classifier Network)
FSA	Tavlama ile Öznitelik Seçimi (Feature Selection with Annealing)

KISALTMA LİSTESİ

SBS	Ardışık Geri Seçim (Sequential Backward Selection)
SFS	Ardışık İleri Seçim (Sequential Forward Selection)
XGBoost	Aşırı Gradyan Artırma (eXtreme Gradient Boosting)
IB	Bilgi Darboğazı (Information Bottleneck)
FRF-fs	Bulanık Rastgele Orman Öznitelik Seçimi (Fuzzy Random Forest Feature Selection)
MLP	Çok Katmanlı Algılayıcı (Multi-Layer Perceptron)
VDR	Değişken Keşif Oranı (Variable Discovery Rate)
SVM	Destek Vektör Makineleri (Support Vector Machines)
ReLU	Doğrultulmuş Doğrusal Birim (Rectified Linear Unit)
LASSO	En Küçük Mutlak Büzülme ve Seçim Operatörü (Least Absolute Shrinkage and Selection Operator)
LightGBM	Hafif Gradyan Artırma Makinesi (Light Gradient Boosting Machine)
MI	Karşılıklı Bilgi (Mutual Information)
KL	Kullback-Leibler
MLE	Maksimum Olabilirlik Kestirimi (Maximum Likelihood Estimation)
mRMR	Minimum Gereksizlik Maksimum İlgililik (Minimum Redundancy Maximum Relevance)
MSE	Ortalama Kare Hata (Mean Squared Error)
LAFS	Öznitelik Seçimi için Öğrenilebilir Dikkat (Learnable Attention for Feature Selection)
RFE	Özyinelemeli Öznitelik Eleme (Recursive Feature Elimination)
AUC	ROC Eğrisi Altındaki Alan (Area Under the ROC Curve)

OLS	Sıradan En Küçük Kareler (Ordinary Least Squares)
FSA	Tavlama ile Öznitelik Seçimi (Feature Selection with Annealing)

ŞEKİL LİSTESİ

Şekil 5.1 LAFS mimarisi; girdi öznitelik vektöründen AttnNet ve ClassNet üzerinden nihai hibrit kayıp hesaplamasına olan akışı göstermektedir. 31

Şekil 6.1 Çeşitli veri setleri için ortalama F1 skorunun λ parametresine karşı grafiği. En uygun λ değeri büyük ölçüde veri setine bağlıdır. 49

TABLO LİSTESİ

Tablo 6.1 XOR veri seti için karşılaştırma sonuçları ($n = 100$). 38
Tablo 6.2 XOR veri seti için karşılaştırma sonuçları ($n = 500$). 38
Tablo 6.3 XOR veri seti için karşılaştırma sonuçları ($n = 1000$). 38
Tablo 6.4 XOR veri seti için karşılaştırma sonuçları ($n = 2000$). 39
Tablo 6.5 XOR veri seti için karşılaştırma sonuçları ($n = 5000$). 39
Tablo 6.6 SYNTH veri seti için, farklı k değerlerine göre benchmark sonuçları. 40
Tablo 6.7 Kapsamlı değerlendirmede kullanılan dokuz referans veri setine genel bakış. 40
Tablo 6.8 Tüm referans veri setleri ve seyreklik seviyeleri (k1, k2 ve k3) için her algoritmaya ait detaylı F1-skorları ve AUC skorları. . . . 42
Tablo 6.9 Gerçek veri setleri üzerinde, her bir veri seti için en büyük k değerinde ölçülen karşılaştırmalı çalışma zamanları (saniye cinsinden). 43
Tablo 6.10 LAFS hiperparametre ayarlama sonuçlarının özeti. Her veri seti için en iyi ortalama F1-Skoru ve karşılık gelen en uygun parametre seti gösterilmektedir. 46
Tablo 6.11 Entropik düzenleyici gücünün (λ) bir fonksiyonu olarak LAFS'ın ortalama F1-Skorunu gösteren hassasiyet çalışması sonuçları. Her veri seti için en iyi performans kalın harflerle belirtilmiştir. 48
Tablo 6.12 Standart ve Kaiming (He) başlatması ile LAFS'ın karşılaştırmalı performansı (Ortalama F1 skoru). 51

ÖZET

BÜYÜK VERİDE ÇAPRAZ ENTROPİYE DAYALI KAYIP FONKSİYONU YARDIMIYLA DEĞİŞKEN SEÇİMİ

Hıncal TOPÇUOĞLU

İstatistik Anabilim Dalı
Doktora Tezi

Danışman: Doç. Dr. Atıf Ahmet EVREN

Özellik seçimi, makine öğreniminde çok boyutluluk sorunu hafifletmek için kritik bir ön işleme adımıdır. Mevcut yöntemler tam bir denge sunmamaktadır: filtre yöntemleri hızlıdır ancak genellikle optimum değildir, sarmalayıcı yöntemler ise güçlüdür ancak hesaplama açısından çok maliyetlidir. Bu tez, sarmalayıcı yöntemlerin performansını daha basit modellerin hızında elde eden yeni, uçtan uca türevlenebilir bir çerçeve olan LAFS'yi (Öğrenilebilir Dikkat Mekanizması ile Özellik Seçimi) önermektedir. LAFS, tek bir ileri geçişte tüm özellikler için bağlam duyarlı önem puanlarını eş zamanlı olarak öğrenmek için bir sinirsel dikkat mekanizması kullanır. Seyrek ve gereksiz olmayan bir özellik alt kümesinin seçilmesini teşvik etmek için, standart sınıflandırma amacını dikkat ağırlıkları üzerinde bilgi kuramsal bir entropik düzenleyici ile birleştiren yeni bir hibrit kayıp fonksiyonu sunulmaktadır. Karşılaştırma veri kümeleri üzerinde kapsamlı ampirik doğrulama yoluyla, LAFS'nin karmaşık özellik etkileşimlerini başarıyla tanımladığını ve çoklu doğrusallığı ele aldığını, en gelişmiş yöntemlerle karşılaştırılabilir sonuçlar elde ettiğini göstermektedir. Çalışmamız, doğruluk-verimlilik sınırında yeni bir nokta belirleyerek, dikkat mekanizmasına dayalı mimarilerin özellik seçimi problemine güçlü ve uyumlu bir çözüm sağladığını ortaya koymaktadır.

Anahtar Kelimeler: özellik seçimi, dikkat mekanizması, bilgi teorisi, derin öğrenme, tablo verileri, boyut indirgeme

YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

ABSTRACT

VARIABLE SELECTION USING A LOSS FUNCTION BASED ON CROSS ENTROPY IN BIG DATA

Hıncal TOPÇUOĞLU

Department of Statistics
Doctor of Philosophy Thesis

Supervisor: Assoc. Prof. Atıf Ahmet EVREN

Feature selection is a critical preprocessing step for mitigating the curse of dimensionality in machine learning. Existing methods present a difficult trade-off: filter methods are fast but often suboptimal, while wrapper methods are powerful but computationally prohibitive. This thesis proposes LAFS (Learnable Attention for Feature Selection), a novel, end-to-end differentiable framework that achieves the performance of wrapper methods at the speed of simpler models. LAFS employs a neural attention mechanism to learn context-aware importance scores for all features simultaneously in a single forward pass. To encourage the selection of a sparse and non-redundant feature subset, we introduce a novel hybrid loss function that combines the standard classification objective with an information-theoretic entropic regularizer on the attention weights. Through comprehensive empirical validation on benchmark datasets, we demonstrate that LAFS successfully identifies complex feature interactions and handles multicollinearity, achieving results comparable to state-of-the-art methods. Our work establishes a new point on the accuracy-efficiency frontier, demonstrating that attention-based architectures provide a powerful and compatible solution to the feature selection problem.

Keywords: feature selection, attention mechanism, information theory, deep learning, tabular data, dimensionality reduction

YILDIZ TECHNICAL UNIVERSITY
GRADUATE SCHOOL OF SCIENCE AND ENGINEERING

1
GİRİŞ

1.1 Yüksek Boyutlu Verinin Zorluğu

Dijital çağın başlangıcı, çok sayıda bilimsel ve endüstriyel alanda benzeri görülmemiş bir veri üretimi ve toplanması patlamasına yol açmıştır. Genomik ve tıbbi görüntülemeden finansal modelleme ve doğal dil işlemeye kadar uzanan alanlar, artık devasa ölçek ve karmaşıklığa sahip veri setleriyle karakterize edilmektedir. Bu veri zengini ortam, hem keşif için önemli bir fırsat hem de makine öğrenmesi için zorlu bir meydan okuma sunmaktadır. Bu alandaki temel engellerden biri, Bellman tarafından yüksek boyutlu uzaylarda veri analiz ederken ortaya çıkan sezgilere aykırı zorlukları tanımlamak için kullanılan bir terim olan "boyutluluk laneti (curse of dimensionality)" olarak bilinen olgudur.

Bir veri setindeki özniteliklerin (veya boyutların, p) sayısı arttıkça, veri uzayının hacmi katlanarak büyür ve bu da mevcut veri noktalarının giderek daha seyrek hale gelmesine neden olur. Bu seyreklik, makine öğrenmesi modellerinin anlamlı örüntüleri ayırt etmesini son derece zorlaştırarak bir dizi olumsuz sonuca yol açar. Çok sayıda ilgisiz veya gereksiz özniteliğin dahil edilmesi, gürültü ekleyerek modellerin tahmin performansını ciddi şekilde düşürebilir, aşırı öğrenme (overfitting) riskini artırabilir ve hesaplama ile depolama maliyetlerini başa çıkılamaz seviyelere çıkarabilir. Ayrıca, çok sayıda öznitelik üzerine kurulan modeller genellikle karmaşık ve şeffaf olmadığından, birçok uygulamada kritik bir gereklilik olan sonuçlarının yorumlanabilirliğini engeller. Sonuç olarak, boyutluluğu azaltmak için sağlam ve verimli yöntemlerin geliştirilmesi, modern makine öğrenmesi araştırmalarında merkezi ve vazgeçilmez bir zorluk haline gelmiştir.

1.2 Öznitelik Seçimi: Kritik Bir Ön İşleme Adımı

Öznitelik seçimi, çok boyutluluk sorunun hafifletmek için temel bir stratejidir. Başlıca amacı, ilgisiz veya gereksiz olanları atarken, orijinal kümeden en

bilgilendirici özniteliklerin minimal bir alt kümesini belirlemek ve korumaktır. Etkili bir öznitelik seçimi süreci birden fazla fayda sağlar: tahmin modellerinin genelleme yeteneğini artırır, eğitim sürelerini kısaltır ve en etkili değişkenlere odaklanarak modelin yorumlanabilirliğini geliştirir.

Mevcut öznitelik seçimi algoritmaları genel olarak üç ana aileye ayrılabilir: filtreler, sarmalayıcılar ve gömülü yöntemler.

- *Filtre Yöntemleri:* Bu yöntemler, herhangi bir spesifik öğrenme algoritmasından bağımsız olarak, özniteliklerin ilgililiğini kendi içsel istatistiksel özelliklerine göre değerlendirir. Karşılıklı Bilgi (Mutual Information - MI) veya Ki-kare testleri gibi teknikler, model eğitimi içermedikleri için hesaplama açısından verimli ve hızlıdır. Ancak, temel sınırlılıkları "miyop" doğalarıdır; öznitelikleri tek başına değerlendirerek, özellikle karmaşık ve çok değişkenli etkileşimlerin varlığında, kolektif olarak güçlü olan öznitelik alt kümelerini tespit etmekte genellikle başarısız olurlar.
- *Sarmalayıcı Yöntemler (Wrapper Methods):* Buna karşılık, Özyinelemeli Öznitelik Eleme (Recursive Feature Elimination - RFE) gibi sarmalayıcı yöntemler, öznitelik alt kümelerini puanlamak ve seçmek için belirli bir tahmin modeli kullanır. Bu yaklaşım, doğrudan seçilen modelin performansı için optimizasyon yapar ve genellikle alanındaki en iyi sonuçları verir. Ancak bu güç, caydırıcı derecede yüksek bir hesaplama maliyetiyle birlikte gelir. Farklı öznitelik alt kümeleri üzerinde bir modelin yüzlerce veya binlerce kez eğitilmesini içerebilen yinelemeli arama süreci, yüksek boyutlu veri setleri için uygulanamaz hale gelir.
- *Gömülü Yöntemler (Embedded Methods):* Bu yöntemler, öznitelik seçimi sürecini doğrudan model eğitimi aşamasına entegre ederek bir uzlaşma sunar. L1 (Lasso) regülarizasyonu gibi teknikler, büyük katsayıları cezalandırarak seyreklik sağlar ve böylece model uydurmanın bir parçası olarak etkili bir şekilde öznitelik seçimi yapar. Sarmalayıcılardan daha verimli olmalarına rağmen, birçok gömülü yöntem doğası gereği doğrusaldır ve doğrusal olmayan öznitelik bağımlılıklarını yakalamakta zorlanabilir.

Bu durum, hesaplama verimliliği ile tahmin gücü arasında zor bir dengeyi ortaya koymaktadır. Bu nedenle araştırma topluluğu şu zorlayıcı soruyla karşı karşıyadır: Sarmalayıcı yöntemlerin yüksek performansına, daha basit modellerle karşılaştırılabilir bir hesaplama maliyetiyle ulaşmak mümkün müdür?

1.3 Yeni Bir Yaklaşım: Öznitelik Seçimi İçin Öğrenilebilir Dikkat (LAFS)

Bu tez, filtre yöntemlerinin hızı ile sarmalayıcı yöntemlerin gücü arasındaki boşluğu doldurmak için tasarlanmış, yeni, uçtan uca türevlenebilir bir çerçeve olan *LAFS'ı (Öznitelik Seçimi için Öğrenilebilir Dikkat - Learnable Attention for Feature Selection)* tanıtmaktadır. LAFS, yinelemeli bir aramaya dayanmak yerine, tüm öznitelikleri paralel olarak bütüncül bir şekilde değerlendirmek ve tek bir ileri geçişte bağlama duyarlı bir önem puanı atamak için bir sinirsel (neural network) dikkat mekanizması (attention mechanism) kullanır.

Bu çalışmanın temel katkısı, modeli seyrek ancak oldukça bilgilendirici bir öznitelik alt kümesi seçmeye yönlendiren hibrit, bilgi-kuramsal bir öğrenme hedefidir. Standart göreve dayalı sınıflandırma kaybını, dikkat dağılımı üzerinde etkili olan bir entropik düzenlileştirici ile güçlendirmektedir. Modeli dikkat ağırlıklarını kullanarak açıkça düşük entropi üretmeye teşvik ederek, öğrenme sürecini yalnızca birkaç özniteliğin yüksek önem aldığı çözümlere doğru yönlendiriyoruz. Bu yaklaşım, yüksek tahmin doğruluğu ve öznitelik seyrekliği hedeflerini tek bir, türevlenebilir eğitim süreci içinde birleştirmektedir.

1.4 Katkılar ve Tez Taslağı

Bu tez, öznitelik seçimi alanına aşağıdaki temel katkıları yapmaktadır:

1. *Özgün ve Türevlenebilir bir Mimari:* Geleneksel yöntemlere güçlü ve ölçeklenebilir bir alternatif sunan, öznitelik seçimi görevi için özel olarak tasarlanmış, tamamen türevlenebilir, dikkat tabanlı bir model olan LAFS'ı önermektedir.

2. *Bilgi-Kuramsal Bir Öğrenme Hedefi:* Minimal ve önemi yüksek olan öznitelik alt kümelerinin seçimini teşvik eden, görev performansı kaybını entropik bir düzenlileştirici ile birleştiren hibrit bir kayıp fonksiyonu sunmaktır.

3. *Kapsamlı Deneysel Doğrulama:* Sentetik ve gerçek dünya referans veri setleri üzerinde yapılan kapsamlı deneylerle LAFS'ın etkinliğini gösterme ve sarmalayıcı yöntemlerden önemli ölçüde daha verimli olurken, alanındaki en iyi algoritmalarla karşılaştırılabilir sonuçlar elde ettiğini ortaya koymaktır.

Bu tezin geri kalanı aşağıdaki şekilde organize edilmiştir. *Bölüm 2*, klasik öznitelik seçimi yöntemleri ve derin öğrenme tabanlı yaklaşımların yükselişi

üzerine kapsamlı bir literatür taraması sunmaktadır. *Bölüm 3*, öznitelik seçimi yöntemlerinin matematiksel temellerini göstermektedir. *Bölüm 4*, probleme bilgi-kuramsal yaklaşımı açıklamaktadır. *Bölüm 5*, mimarisi ve hibrit kayıp fonksiyonunun formülasyonu da dahil olmak üzere LAFS çerçevesinin ayrıntılı metodolojisini sunmaktadır. Son olarak *Bölüm 6*, deneysel kurguyu tanımlamakta ve ampirik doğrulamamızın sonuçlarını sunmakta, farklı türdeki veri setleri üzerindeki deneyleri, bulgularımızın çıkarımlarını tartışmakta, mevcut çalışmanın sınırlılıklarını belirtmekte ve gelecekteki araştırmalar için umut vaat eden yollar önermektedir.

2
LİTERATÜR TARAMASI

2.1 Öznitelik Seçiminin Genel Görünümü

Öznitelik seçimi, on yıllardır makine öğrenmesi ve istatistik alanlarında merkezi bir tema olmuş ve bu da geniş ve çeşitli bir literatür birikimine yol açmıştır. Temel amaç, model karmaşıklığı ile tahmin performansı arasında en uygun dengeyi bulmaktır; bu zorluk literatürde kapsamlı bir şekilde incelenmiştir [1]. Kapsamlı bir derleme, yöntemlerin genel görünümünü dört ana aileye ayırmaktadır: filtreler, sarmalayıcılar, gömülü ve hibrit yöntemler [2]. Bu taksonomi, alanın evrimini anlamak için yapılandırılmış bir çerçeve sunmaktadır. Bu bölüm, LAFS çerçevesinin doğrudan temelini oluşturan daha güncel derin öğrenme ve bilgi-kuramsal kavramlara girmeden önce, bu klasik paradigmaları inceleyecek ve kilit yazarların özel katkılarını vurgulayacaktır.

2.2 Filtre Yöntemleri: İstatistiksel Saflık ve Hız

Herhangi bir öğrenme algoritmasından bağımsız bir ön işleme adımı olarak çalışan filtre yöntemleri, öznitelik seçimine yönelik en erken yaklaşımı temsil etmektedir. Temelleri klasik istatistiğe dayanmaktadır. Örneğin, ilk olarak 1900'de önerilen Ki-kare testi, kategorik özniteliklerin hedef sınıftan bağımsızlığını değerlendirmek için kullanılır [3]. Benzer şekilde, diskriminant analizinden türetilen Fisher Skoru [4], sınıflar arası varyansın sınıf içi varyansa oranını ölçerek öznitelikleri değerlendirir.

Birçok filtre yönteminin temel taşı, bilgi kuramı üzerine yapılan temel çalışmalara dayanan Karşılıklı Bilgi (Mutual Information - MI) kavramıdır [5]. MI, iki değişken arasındaki istatistiksel bağımlılığı nicelleştirerek, güçlü ve doğrusal olmayan bir ilgililik (association) ölçüsü sunar. Ancak, özniteliklerin basit bir MI tabanlı sıralaması kritik bir kusuru vardır: birbirleriyle yüksek oranda ilişkili olan öznitelikleri seçme eğilimindedir, bu da gereksiz bir öznitelik setine yol açar.

Bu sorunu çözmek için daha sofistike filtreler geliştirilmiştir. Bu alandaki dönüm noktası niteliğindeki bir katkı, Minimum Gereksizlik Maksimum İlgililik (Minimum Redundancy Maximum Relevance - mRMR) algoritmasıdır [6]. Bu yöntem, hedefi aynı anda hedef sınıfla karşılıklı bilgiyi maksimize etmek (Maks-İlgililik) ve seçilen öznitelikler arasındaki karşılıklı bilgiyi minimize etmek (Min-Gereksizlik) olarak formüle eder. Gereksizliği aktif olarak cezalandırma ilkesi, bizim LAFS modelimizde önerilen entropik düzenlileştirici ile derin bir felsefi motivasyonu paylaşmaktadır. Dikkate değer bir diğer filtre yöntemi ise orijinal Relief algoritmasının bir uzantısı olan ReliefF'tir [7]. ReliefF, en yakın komşu örneklere olan mesafeleri analiz ederek öznitelik kalitesini tahmin eder ve alt kümeleri ayrıntılı bir şekilde aramadan öznitelik etkileşimlerini tespit etmesine olanak tanır. Bu ilerlemelere rağmen, tüm filtre yöntemleri, belirli bir tahmin modelinin performansını doğrudan optimize edememe temel sınırlılığıyla kısıtlı kalmaktadır.

2.3 Sarmalayıcı Yöntemler: Performans Arayışı

Sarmalayıcı yöntemler, öznitelik seçimi sürecini bir arama problemi olarak ele alarak ve bir tahmin modelinin performansını amaç fonksiyonu olarak kullanarak filtrelerin temel sınırlılığını giderir. Bu paradigma 1990'ların başlarında popüler hale gelmiştir [8] [9]. Arama uzayı, tüm özniteliklerin kuvvet kümesidir ve amaç, model performansını maksimize eden alt kümeyi bulmaktır.

Bu konuda önemli bir çalışma, bu yaklaşımın gücünü, başlangıçta Destek Vektör Makineleri (Support Vector Machines - SVM) kullanılarak kanser sınıflandırmasında gen seçimi için geliştirilen Özyinelemeli Öznitelik Eleme (Recursive Feature Elimination - RFE) ile göstermiştir [10]. RFE, yinelemeli olarak bir model eğiterek, öznitelik önemlerini hesaplayarak ve istenen alt küme boyutuna ulaşılana kadar en önemsiz özniteliği eleyerek çalışır. Bu yöntem, özellikle Hafif Gradyan Artırma Makinesi (Light Gradient Boosting Machine - LightGBM) gibi güçlü modern sınıflandırıcılarla eşleştirildiğinde, genellikle alanındaki en iyi sonuçları verir [11].

Ancak, sarmalayıcıların gücü aynı zamanda en büyük zayıflıklarıdır. 2^p olası öznitelik alt kümesi boyunca yapılan kapsamlı arama, klasik bir NP-zor problemdir. Ardışık İleri Seçim (Sequential Forward Selection - SFS) [12] ve Ardışık Geri Seçim (Sequential Backward Selection - SBS) [13] gibi arama stratejileri uygulanabilir bir alternatif sunsa da bir modelin binlerce kez eğitilmesini gerektirirler, bu da onları yüksek boyutlu veriler için hesaplama zorluğu içine sokar.

2.4 Gömülü Yöntemler: Bütünleşik Seçim

Gömülü yöntemler, öznitelik seçimini doğrudan model eğitimi sürecine entegre ederek iyi bir uzlaşma olanağı sunar. Bu, onları sarmalayıcılardan daha hesaplama verimli ve filtrelerden daha modele duyarlı hale getirir.

Çarpıcı örnek olarak, daha az bilgilendirici özniteliklerin katsayılarını sıfıra zorlamak için doğrusal bir modelin kayıp fonksiyonuna L1-norm cezası ekleyen LASSO'dur (Least Absolute Shrinkage and Selection Operator) [14]. Doğrusal modellerin ötesinde, ağaç tabanlı topluluklar doğası gereği öznitelik seçimi yapar. Rastgele Ormanlar (Random Forests) [15] ve Gradyan Artırma Makineleri (Gradient Boosting Machines) [16] gibi topluluk yöntemleri, tüm ağaçlardaki öznitelik önemlerini bir araya getirerek güçlü ve yaygın olarak kullanılan bir gömülü seçim mekanizması sunar. Daha yakın zamanlarda, Tavlama ile Öznitelik Seçimi (Feature Selection with Annealing - FSA) yöntemi, kısıtlı bir optimizasyon çerçevesine deterministik bir tavlama programı entegre ederek, tek bir eğitim süreci sırasında öznitelikleri kademeli olarak eler [17]. Güçlü olmalarına rağmen, bu yöntemler genellikle belirli model mimarilerine bağlıdır.

2.5 Tablosal Veri İçin Derin Öğrenme Devrimi

Derin öğrenme, yapılandırılmamış veri alanlarında fiili standart haline gelmiş olsa da, tablosal verilere uygulanması daha çok tartışılmıştır. Ancak son zamanlarda, Transformer modelinden [18] ilham alan dikkat mekanizmalarını kullanan mimariler büyük umut vaat etmiştir.

Bu alandaki öncü bir çalışma, örnek bazında öznitelik seçimi için sıralı, çok adımlı bir dikkat mekanizması kullanan TabNet olmuştur [19]. Bunu takiben, FT-Transformer [20] ve SAINT [21] gibi modeller, Transformer mimarisinin tablosal veriler üzerinde denetimli öğrenme için nasıl uyarlanabileceğini göstermiştir. Bu modeller güçlü öznitelik temsilleri öğrense de, birincil amaçları tahmin doğruluğudur, açıkça öznitelik seçimi değildir. Dikkat ağırlıkları, yorumlama için bir yan ürünken, bizim çözümümüz için nihai, seyrek bir öznitelik setinin oluşturulması birincil hedeftir.

2.6 Derin Öğrenmede Bilgi-Kuramsal İlkeler

Bilgi Darboğazı (Information Bottleneck - IB) ilkesi, en uygun modelin, hedef hakkında hala maksimum düzeyde bilgilendirici olan, girdinin maksimum düzeyde sıkıştırılmış bir temsilini öğrenmesi gerektiğini varsayan güçlü bir teorik çerçeve sunar [22].

Kullanılan algoritmada (LAFS) [23] dikkat dağılımı, bu darboğaz olarak görülebilir. Shannon entropisini en aza indirmek, girdi temsilini sıkıştırmak için doğrudan bir vekil (proxy) görevi görür. Bu ilke, IB hedefini modern sinir ağları için işlenebilir hale getiren Derin Varyasyonel Bilgi Darboğazı (Deep Variational Information Bottleneck) aracılığıyla derin öğrenmeye genişletilmiştir [24]. Bununla yakından ilişkili olan, girdinin farklı bölümleri arasındaki karşılıklı bilgiyi maksimize ederek temsiller öğrenmeye odaklanan Deep InfoMax çalışmasıdır [25]. Ayrıca, Concrete Autoencoders [26] gibi yakın tarihli çalışmalar, uçtan uca, türevlenebilir öznitelik seçimi yapmak için stokastik geçitleri ve gevşetme tekniklerini kullanmayı açıkça araştırmış ve LAFS ile aynı hedefe yönelik bir eğilim göstermiştir.

3
ÖZNİTELİK SEÇİMİ YÖNTEMLERİNİN MATEMATİKSEL TEMELLERİ

3.1 Öznitelik Seçimi Paradigmalarına Giriş

Literatür taramasında da belirtildiği gibi, öznitelik seçimi yöntemleri genel olarak filtreler, sarmalayıcılar ve gömülü yöntemler olarak sınıflandırılabilir. Boyutluluğu azaltma ortak hedefini paylaşsalar da, altında yatan matematiksel ilkeler önemli ölçüde farklılık gösterir. Bu bölüm, her kategoriden en etkili algoritmaların resmi bir matematiksel tanımını sunarak, bir sonraki bölümde tanıtılacak olan LAFS çerçevesinin değerlendirilmesi ve karşılaştırılması için bir temel oluşturmaktadır.

3.2 Filtre Yöntemleri: İlgililik ve Gereksizliğin İstatistiksel Ölçüleri

Filtre yöntemleri, öznitelikleri verinin kendisinden hesaplanan istatistiksel kriterlere göre sıralar. Bu yöntemler modelden bağımsızdır ve bilgi kuramı ile istatistik ilkelerine dayanır.

3.2.1 Karşılıklı Bilgi (Mutual Information - MI)

Karşılıklı Bilgi, bir öznitelik ile hedef değişken arasındaki istatistiksel bağımlılığı nicelleştirir. Ayrık bir X özniteliği ve ayrık bir Y hedef sınıfı için, karşılıklı bilgi $I(X;Y)$ şu şekilde tanımlanır:

$$I(X;Y) = \sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right) \tag{3.1}$$

burada $p(x,y)$, X ve Y'nin ortak olasılık kütle fonksiyonu, $p(x)$ ve $p(y)$ ise marjinal olasılık kütle fonksiyonlarıdır. Daha yüksek bir $I(X;Y)$ değeri, öznitelik ile hedef arasında daha güçlü bir ilişki olduğunu ima eder.

Basit bir MI tabanlı filtre yönteminin prosedürel adımları algoritma 3.1'de formüle edilmiştir. Bu yaklaşım, her bir öznitelik için MI skorunu bağımsız olarak hesaplamayı ve ardından bu sıralamaya göre en iyi k özniteliği seçmeyi içerir.

Algoritma 3.1 Karşılıklı bilgi ile öznitelik seçimine ilişkin algoritma

Require: D: n örnek içeren bir veri seti. $\mathcal{F} = \{F_1, F_2, ..., F_d\}$: d adet öznitelik kümesi. Y: Hedef sınıf değişkeni. k: Seçilecek istenen öznitelik sayısı.

Ensure: S: En iyi k özniteliği içeren $\mathcal{F}$'nin bir alt kümesi.

1: (öznitelik, skor) çiftlerini saklamak için boş bir `scores` listesi başlat.
2: **for** $\mathcal{F}$'deki her bir F_j özniteliği için **do**
3: F_j özniteliği ile Y hedefi arasındaki Karşılıklı Bilgiyi, $I(F_j; Y)$, hesapla. ▷ Sürekli öznitelikler için genellikle önce ayrıklaştırma (binning) uygulanır.
4: $(F_j, I(F_j; Y))$ çiftini `scores` listesine ekle.
5: **end for**
6: `scores` listesini MI skorlarına göre azalan sırada sırala.
7: S, sıralanmış `scores` listesindeki ilk k özniteliğin kümesi olsun.
8: **return** S

Algoritma 3.1'de detaylandırılan yöntemin temel gücü, hesaplama verimliliğidir. Ancak, sözde kod aynı zamanda ana sınırlılığını da açıkça ortaya koymaktadır: her öznitelik tamamen izole bir şekilde değerlendirilir (Satır 3), bu da algoritmayı "miyop" yapar ve çok değişkenli öznitelik etkileşimlerini veya gereksizliği hesaba katmasını engeller. Bu sınırlılık, doğrudan mRMR gibi daha gelişmiş yöntemlerin geliştirilmesine yol açmıştır.

3.2.2 Minimum Gereksizlik Maksimum İlgililik (mRMR)

mRMR algoritması, öznitelik ilgililiğini ve öznitelikler arası gereksizliği eş zamanlı olarak göz önünde bulundurarak MI yaklaşımını geliştirir. Amaç, "Maks-İlgililik" kriteri D'yi maksimize eden ve "Min-Gereksizlik" kriteri R'yi minimize eden bir S öznitelik alt kümesini bulmaktır. Bunlar şu şekilde tanımlanır:

$$\text{Maks-İlgililik (D):} \quad D = \frac{1}{|S|} \sum_{X_i \in S} I(X_i; Y) \tag{3.2}$$

$$\text{Min-Gereksizlik (R):} \quad R = \frac{1}{|S|^2} \sum_{X_i, X_j \in S} I(X_i; X_j) \tag{3.3}$$

mRMR algoritması, algoritma 3.2'de formüle edildiği gibi, tipik olarak açgözlü bir ileri seçim stratejisi kullanır.

Algoritma 3.2 Minimum Gereksizlik Maksimum İlgililik (mRMR) öznitelik seçimi

Require: D: n örnek içeren bir veri seti. $\mathcal{F} = \{F_1, F_2, ..., F_d\}$: d adet öznitelik kümesi. Y: Hedef sınıf değişkeni. k: Seçilecek istenen öznitelik sayısı.

Ensure: S: En iyi k mRMR özniteliğini içeren $\mathcal{F}$'nin bir alt kümesi.

1: Seçilen öznitelik kümesini $S \leftarrow \emptyset$ olarak başlat.
2: Seçilmemiş özniteliklerin kümesi $\mathcal{U} \leftarrow \mathcal{F}$ olsun.
3: ▷ Adım 1: İlk özniteliği yalnızca maksimum ilgililiğe göre seç.
4: $F_{first} \leftarrow \arg\max_{F_j \in \mathcal{U}} I(F_j; Y)$.
5: $S \leftarrow S \cup \{F_{first}\}$.
6: $\mathcal{U} \leftarrow \mathcal{U} \setminus \{F_{first}\}$.
7: ▷ Adım 2: Kalan $k-1$ özniteliği açgözlü bir şekilde seç.
8: **for** $i = 2$ to k **do**
9: $best_score \leftarrow -\infty$ ve $best_feature \leftarrow$ null olarak başlat.
10: **for** $\mathcal{U}$'daki her bir aday öznitelik F_c için **do**
11: $relevance \leftarrow I(F_c; Y)$.
12: $redundancy \leftarrow \frac{1}{|S|} \sum_{F_s \in S} I(F_c; F_s)$.
13: $current_score \leftarrow relevance - redundancy$.
14: **if** $current_score > best_score$ **then**
15: $best_score \leftarrow current_score$.
16: $best_feature \leftarrow F_c$.
17: **end if**
18: **end for**
19: $S \leftarrow S \cup \{best_feature\}$.
20: $\mathcal{U} \leftarrow \mathcal{U} \setminus \{best_feature\}$.
21: **end for**
22: **return** S

Algoritma 3.2'de detaylandırıldığı gibi, mRMR prosedürü basit bir MI sıralamasından önemli ölçüde daha fazla hesaplama yoğundur. Seçilecek k özniteliğin her biri için, bir aday öznitelik ile daha önce seçilmiş tüm öznitelikler arasındaki karşılıklı bilgiyi hesaplaması gerekir (Satır 13). Bu artan maliyete rağmen, seçim kriteri tamamen veriden türetilen istatistiksel özelliklere dayandığı ve bir tahmin modelinin eğitimini içermediği için bir filtre yöntemi olarak kalır.

3.3 Sarmalayıcı Yöntemler: Arama Tabanlı bir Yaklaşım

Sarmalayıcı yöntemler, öznitelik seçimi problemini, amaç fonksiyonunun belirli bir makine öğrenmesi modelinin performansı olduğu bir arama problemi olarak ele alır.

3.3.1 Genel Formülasyon

$\mathcal{A}$ seçilmiş bir öğrenme algoritması (örneğin, LightGBM, SVM) ve $\mathcal{L}(\cdot)$ bir doğrulama seti üzerindeki model hatasını temsil eden bir değerlendirme metriği olsun. Bir sarmalayıcı yöntemin amacı, bu hatayı en aza indiren k boyutlu en uygun S^* öznitelik alt kümesini bulmaktır:

$$S^* = \arg \min_{S \subset \mathcal{F}, |S|=k} \mathcal{L}(\mathcal{A}(\text{data}_S)) \tag{3.4}$$

Tüm $\binom{d}{k}$ alt küme arasında arama yapmak hesaplama açısından mümkün olmadığından, açgözlü arama stratejileri kullanılır.

3.3.2 Özyinelemeli Öznitelik Eleme (Recursive Feature Elimination - RFE)

Özyinelemeli Öznitelik Eleme (RFE), algoritma 3.3'te formüle edildiği gibi, geri eleme stratejisi kullanan klasik ve güçlü bir sarmalayıcı yöntemdir.

Algoritma 3.3 Özyinelemeli öznitelik eleme (Recursive Feature Elimination – RFE)

Require: D: $\mathcal{F}$ öznitelikleri ve Y hedefini içeren bir veri seti. $\mathcal{A}$: Her özniteliğe bir önem skoru atayabilen bir öğrenme algoritması. k: İstenen nihai öznitelik sayısı.

Ensure: S: En iyi k özniteliği içeren $\mathcal{F}$'nin bir alt kümesi.

1: Mevcut öznitelik kümesini $S_c \leftarrow \mathcal{F}$ olarak başlat.
2: **while** $|S_c| > k$ **do**
3: Öğrenme algoritması $\mathcal{A}$'yı, sadece S_c'deki öznitelikleri kullanarak veri seti üzerinde eğit.
4: S_c'deki her bir F_j özniteliği için bir w_j önem skoru hesapla. ▷ Skorlar model katsayıları, Gini önemi (importance) vb. olabilir.
5: En önemsiz özniteliği bul: $F^- \leftarrow \arg\min_{F_j \in S_c} w_j$.
6: En önemsiz özniteliği ele: $S_c \leftarrow S_c \setminus \{F^-\}$.
7: **end while**
8: **return** S_c

Algoritma 3.3'ün gösterdiği gibi, RFE'nin gücü, modelin her adımda yeniden eğitilmesidir (Satır 3), bu da karmaşık öznitelik etkileşimlerini hesaba katmasına olanak tanır. Ancak, bu, bir modelin $d - k$ kez eğitilmesini gerektirir, bu da maliyeti yüksek boyutlu veri setleri için artırır. Bu ödünleşim (trade-off), LAFS gibi yöntemlerin çözmeyi amaçladığı merkezi bir problemdir.

3.4 Gömülü Yöntemler: Bütünleşik Seçim

Gömülü yöntemler, öznitelik seçimini, genellikle kayıp fonksiyonuna bir düzenlileştirme terimi ekleyerek, model eğitimi sürecinin bir parçası olarak dahil eder.

3.4.1 LASSO (L1 Düzenlileştirmesi)

LASSO (Least Absolute Shrinkage and Selection Operator), doğrusal modeller için gömülü bir yöntemin en iyi örneğidir. Standart en küçük kareler (ordinary least squares - OLS) hedefini, model katsayılarının mutlak değerlerinin toplamıyla orantılı bir ceza terimi ekleyerek değiştirir:

$$\hat{\boldsymbol{\beta}}_{\text{LASSO}} = \arg\min_{\boldsymbol{\beta}} \left(\frac{1}{2n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \right) \tag{3.5}$$

L1-normunun kritik özelliği, bazı katsayıları tam olarak sıfır olmaya zorlamasıdır. Bu amaç fonksiyonu, algoritma 3.4'da detaylandırıldığı gibi, tipik olarak Koordinat İnişi (Coordinate Descent) gibi algoritmalarla çözülür [27–29]. Bu süreç, katsayıları sıfıra çeken mekanizma olan Yumuşak-Eşikleme Operatörüne (Soft-Thresholding Operator), $S_\lambda(z) = \text{sign}(z)\max(|z| - \lambda, 0)$, dayanır.

Algoritma 3.4 Koordinat inişi ile LASSO algoritması

Require: Veri matrisi $\mathbf{X}$, hedef vektör $\mathbf{y}$ ve düzenlileştirme parametresi λ.
Ensure: Bazı katsayıları sıfır olabilen katsayı vektörü $\boldsymbol{\beta}$.

$\boldsymbol{\beta} \leftarrow \mathbf{0}$ olarak başlat.
repeat
  $\boldsymbol{\beta}_{old} \leftarrow \boldsymbol{\beta}$.
  for $j = 1$ to d **do**
    Kısmi artıkları hesapla: $\mathbf{r}_j \leftarrow \mathbf{y} - \mathbf{X}\boldsymbol{\beta} + \beta_j X_j$.
    Korelasyonu hesapla: $z_j \leftarrow \frac{1}{n} X_j^T \mathbf{r}_j$.
    Yumuşak-eşikleme uygula: $\beta_j \leftarrow S_\lambda(z_j)$.
  end for
until yakınsama sağlanana kadar
return $\boldsymbol{\beta}$
▷ Seçilen öznitelik kümesi dolaylı olarak $\{F_j \mid \beta_j \neq 0\}$'dır.

Algoritma 3.4'te gösterildiği gibi, öznitelik seçimi optimizasyon döngüsü içinde gerçekleşir. Bu, RFE'nin açık "bir-öznitelik-çıkar" adımından temel olarak farklıdır ve bir gömülü yöntemin tanımlayıcı karakteristiğidir.

Bu bölüm, LAFS'ın karşılaştırıldığı yöntemler için gerekli matematiksel bağlamı sağlamıştır. Bu yöntemlerin, bir sonraki bölümde detaylandırılacak olan LAFS çerçevesinin tamamen türevlenebilir, dikkat tabanlı hedefine bir karşıtlık noktası olarak hizmet edecek ceza tabanlı doğasını vurgulamaktadır.

4
ÖZNİTELİK SEÇİMİ İÇİN BİLGİ-KURAMSAL BİR ARAÇ SETİ

4.1 Makine Öğrenmesinde Bilgi Kuramının Rolü

İlk olarak 1948 tarihli çığır açan "A Mathematical Theory of Communication" çalışmasında geliştirilen bilgi kuramı, belirsizliği, bilgiyi ve istatistiksel uzaklığı nicelleştirmek için güçlü bir matematiksel çerçeve sunar. Kökenleri sinyal iletimine dayansa da, ilkelerinin istatistik için temel olduğu hızla fark edilmiş ve bu bağlantı Kullback'in "Information Theory and Statistics" [30] adlı temel metniyle pekiştirilmiştir. Bugün, alanın temel kavramları, en başta Cover ve Thomas'ın "Elements of Information Theory" [31] eseri olmak üzere, Ash'in "Information Theory" [32] gibi diğer önemli çalışmalarla modern ve kesin ders kitaplarında kapsamlı bir şekilde detaylandırılmıştır.

Bilgi kuramının ilkeleri, veri bilimi ve makine öğrenmesi alanlarında vazgeçilmez hale gelmiş ve ölçümleri, Agresti [33] tarafından ele alınan kategorik veriler için tahmin modelleri oluşturmaktan, Pardo [34] tarafından detaylandırılan ıraksama ölçülerine dayalı gelişmiş istatistiksel çıkarım yöntemleri geliştirmeye kadar geniş bir uygulama yelpazesinde aktif olarak kullanılmaktadır. Bu bağlamda, Evren vd. yazarları, maksimum entropi altında nitel değişkenlik ölçülerinin özelliklerini araştırmış [35] ve ekonomik hipotezlerdeki eğriselliği test etmek için Rényi ve Tsallis gibi genelleştirilmiş karşılıklı bilgi ölçülerini uygulamışlardır [36]. Ayrıca, gelir eşitsizliğinin araştırılmasında, çok kullanılan Dirichlet dağılımının parametrelerinin belirlenmesinde entropiye dayalı tahmin yöntemlerini vurgulayan çalışmalar da mevcuttur [37]. Ayrıca Türk akademik literatüründe, entropiye dayalı parametre tahmini konusunda Şamilov [38] gibi araştırmacıların çalışmalarına bakılabilir.

Bu bölüm, bu alandaki temel kavramlara resmi ve ayrıntılı bir giriş sunmaktadır. Shannon'un temel entropi tanımıyla başlanarak ve Kullback-Leibler ıraksaması

ile Jeffreys [39] tarafından önerilen simetrik formülasyon da dahil olmak üzere türevlerini tartışılacaktır. Ardından, kapsamımızı Rényi entropisi [40] ve Tsallis entropisi [41] gibi önemli genelleştirilmiş entropi formülasyonlarını içerecek şekilde genişletecektir. Bu teorik araç seti, yalnızca akademik bir derleme değildir; bir sonraki bölümde detaylandırılacak olan LAFS çerçevesinin merkezinde yer alan özgün hibrit kayıp fonksiyonu için matematiksel gerekçeyi sunmaktadır.

4.2 Temel: Shannon Entropisi

Özünde Shannon entropisi, bir rassal değişkenin olası sonuçlarında bulunan ortalama belirsizliğin bir ölçüsüdür. Temel sezgi, yüksek olasılıklı bir olayın gerçekleşmesinin çok şaşırtıcı olmadığı ve dolayısıyla az bilgi taşıdığı, oysa nadir bir olayın gerçekleşmesinin oldukça şaşırtıcı olduğu ve önemli bilgi taşıdığıdır. $p(x)$ olasılığına sahip tek bir x sonucunun bilgi içeriği veya "şaşırtıcılığı", resmi olarak şu şekilde tanımlanır:

$$I(x) = -\log(p(x)) \tag{4.1}$$

Logaritma tabanının seçimi, bilgi birimlerini belirler; bilgi kuramında taban 2 gelenekseldir ve "bit" birimlerine yol açar. Logaritma, bağımsız olaylardan gelen bilginin toplanabilir olmasını sağlar ki bu arzu edilen bir özelliktir. Eksi işareti ise olasılıklar $[0, 1]$ aralığında olduğu için sonucu negatif olmayan bir değere dönüştürür.

4.2.1 Shannon Entropisinin Tanımı

$H(X)$ ile gösterilen Shannon entropisi, ayrık bir X rassal değişkeninin tüm olası sonuçları üzerinden bilgi içeriğinin beklenen değeridir:

$$H(X) = E[I(X)] = \sum_{x \in \mathcal{X}} p(x)I(x) = -\sum_{x \in \mathcal{X}} p(x)\log(p(x)) \tag{4.2}$$

Kabul gereği, $0 \log 0 = 0$ olarak alınır.

4.2.2 Shannon Entropisinin Özellikleri

Shannon entropisinin birkaç temel özelliği vardır:

- *Negatif olmama:* $H(X) \geq 0$. Entropi, yalnızca ve yalnızca bir sonucun kesin olması durumunda (bazı i için $p(x_i) = 1$) sıfırdır, bu da tam bir belirsizlik yokluğunu temsil eder.

- *Maksimum Entropi:* N olası sonuca sahip bir rassal değişken için entropi, dağılım tekdüze olduğunda, yani tüm i için $p(x_i) = 1/N$ olduğunda maksimize edilir. Bu durumda, $H(X) = \log(N)$ olur. Bu, maksimum belirsizlik durumunu temsil eder.

- *Sürekli durum için,* toplama işlemi bir integral ile değiştirilir ve bu da *diferansiyel entropiyi* tanımlar:

$$h(X) = -\int_{-\infty}^{\infty} f(x)\log(f(x))\,dx \tag{4.3}$$

  burada $f(x)$, X'in olasılık yoğunluk fonksiyonudur. Ayrık entropinin aksine, diferansiyel entropi negatif olabilir.

4.3 Bileşik ve Koşullu Entropi

4.3.1 Bileşik Entropi

Bir çift ayrık rassal değişken (X, Y)'nin bileşik entropisi $H(X, Y)$, bu çiftle ilişkili toplam belirsizliği ölçer:

$$H(X,Y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log(p(x,y)) \tag{4.4}$$

burada $p(x, y)$ bileşik olasılık dağılımıdır.

4.3.2 Koşullu Entropi

Koşullu entropi $H(Y|X)$, bir başka rassal değişken X'in değeri bilindiğinde, bir Y rassal değişkenindeki kalan belirsizliği ölçer. Şu şekilde tanımlanır:

$$H(Y|X) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log(p(y|x)) \tag{4.5}$$

Bu, *entropi için zincir kuralına* yol açar:

$$H(X,Y) = H(X) + H(Y|X) \tag{4.6}$$

Bu kural sezgisel olarak, bir değişken çiftinin toplam belirsizliğinin, birincinin belirsizliği artı, birinci bilindiğinde ikincinin belirsizliği olduğunu belirtir. Önemli bir özellik, bilginin asla belirsizliği artırmamasıdır: $H(Y|X) \leq H(Y)$.

4.4 Göreceli Entropi ve Karşılıklı Bilgi

4.4.1 Göreceli Entropi (Kullback-Leibler Iraksaması)

Kullback-Leibler (KL) ıraksaması olarak da bilinen göreceli entropi, bir P olasılık dağılımının ikinci bir referans olasılık dağılımı Q'dan ne kadar saptığının temel bir ölçüsüdür. Q için optimize edilmiş bir kod kullanıldığında P'den gelen örnekleri kodlamak için gereken ek bilgiyi (bit cinsinden) nicelleştirir. Ayrık P ve Q dağılımları için KL ıraksaması şu şekilde tanımlanır:

$$D_{KL}(P\|Q) = \sum_{x\in\mathcal{X}} p(x) \log\left(\frac{p(x)}{q(x)}\right) \tag{4.7}$$

KL ıraksaması negatif değildir (Gibbs eşitsizliği olarak bilinen bir sonuç), $D_{KL}(P\|Q) \geq 0$, ve eşitlik yalnızca ve yalnızca $P = Q$ durumunda geçerlidir. KL ıraksamasının kritik bir özelliği asimetrik olmasıdır, yani $D_{KL}(P\|Q) \neq D_{KL}(Q\|P)$. Bu, onun gerçek bir uzaklık metriği olmadığı anlamına gelir, çünkü P'den Q'ya olan "uzaklık", Q'dan P'ye olan "uzaklık" ile aynı değildir. Bu asimetri, bir dağılımın "gerçek" bir referans (P) ve diğerinin bir yaklaşım (Q) olduğu bağlamlarda bir hata değil, bir özelliktir; ancak simetrik bir karşılaştırma gerektiğinde bir sınırlamadır.

4.4.2 Jeffrey Iraksaması

KL ıraksamasının asimetrisinin üstesinden gelmek için çeşitli simetrikleştirilmiş varyantlar önerilmiştir. En yaygın olanlardan biri, iki KL ıraksamasının toplamı olarak tanımlanan Jeffrey Iraksaması'dır (J-ıraksaması olarak da bilinir):

$$J(P,Q) = D_{KL}(P\|Q) + D_{KL}(Q\|P) \tag{4.8}$$

KL ıraksamasının tanımını yerine koyarak, Jeffrey Iraksaması'nı açıkça yazabiliriz:

$$J(P,Q) = \sum_{x\in\mathcal{X}} p(x) \log\left(\frac{p(x)}{q(x)}\right) + \sum_{x\in\mathcal{X}} q(x) \log\left(\frac{q(x)}{p(x)}\right) \tag{4.9}$$

Bu ifade daha kompakt bir forma dönüştürülebilir:

$$J(P,Q) = \sum_{x\in\mathcal{X}} (p(x) - q(x)) \log\left(\frac{p(x)}{q(x)}\right) \tag{4.10}$$

Jeffrey Iraksaması'nın aşağıdaki temel özellikleri vardır:

- *Simetri:* Tanımı gereği, $J(P,Q) = J(Q,P)$ olduğu açıktır.

- *Negatif olmama:* D_{KL} her zaman negatif olmadığından, toplamları olan $J(P,Q)$ da negatif değildir ve $J(P,Q) = 0$ ancak ve ancak $P = Q$ ise geçerlidir.

Simetrik olmasına rağmen, Jeffrey Iraksaması, KL ıraksaması gibi üçgen eşitsizliğini sağlamaz ve bu nedenle gerçek bir uzaklık metriği değildir. Ancak, simetrik yapısı, hiçbir dağılımın "temel gerçek" olarak hizmet etmediği ve dengeli bir benzemezlik ölçüsüne ihtiyaç duyulan uygulamalarda onu değerli bir araç haline getirir.

4.4.3 Karşılıklı Bilgi (Mutual Information - MI)

Karşılıklı Bilgi, iki rassal değişken arasındaki karşılıklı bağımlılığı ölçer. Bir rassal değişkeni gözlemleyerek diğeri hakkında elde edilen "bilgi miktarını" nicelleştirir. İki eşdeğer şekilde tanımlanabilir:

1. Belirsizlikteki azalma olarak:

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \tag{4.11}$$

2. Bileşik dağılım ile marjinallerin çarpımı arasındaki KL ıraksaması cinsinden:

$$I(X;Y) = D_{KL}(p(x,y)\|p(x)p(y)) \tag{4.12}$$

Bu ikinci tanım, MI'nın bileşik dağılımın bağımsızlığı temsil edecek dağılımdan ne kadar uzak olduğunu ölçtüğünü ortaya koyar.

4.5 Çapraz Entropi

Çapraz entropi, KL ıraksamasıyla yakından ilişkili bir kavramdır ve sınıflandırma görevleri için en yaygın kullanılan kayıp fonksiyonudur. "Gerçek" bir P dağılımı ile bir modelin tahmin ettiği Q dağılımı arasındaki çapraz entropi şu şekilde verilir:

$$H(P,Q) = -\sum_{x\in\mathcal{X}} p(x)\log(q(x)) \tag{4.13}$$

Çapraz entropi, Shannon entropisi ve KL ıraksaması arasındaki ilişki doğrudan ve çok önemlidir:

$$H(P,Q) = H(P) + D_{KL}(P\|Q) \tag{4.14}$$

Makine öğrenmesi sınıflandırmasında, P gerçek veri dağılımıdır (örneğin, doğru sınıf için bir one-hot vektör $[0, 1, 0, \dots]$) ve Q modelin tahmin ettiği olasılık dağılımıdır (örneğin, bir softmax fonksiyonunun çıktısı). Gerçek dağılımın entropisi, $H(P)$, modelin parametrelerine göre bir sabit olduğundan, çapraz entropi kaybını Q'ya göre en aza indirmek, matematiksel olarak tahmin edilen ve gerçek dağılımlar arasındaki KL ıraksamasını en aza indirmeye eşdeğerdir. Bu, onu olasılıksal sınıflandırıcıları eğitmek için ideal bir araç yapar.

4.6 Shannon Entropisinin Genelleştirmeleri

Shannon entropisi temel olsa da, daha fazla esneklik sağlamak için birkaç genelleştirilmiş entropi formülasyonu önerilmiştir. Bunlar genellikle farklı olasılıklara sahip sonuçların katkılarını vurgulamak veya azaltmak için parametrelendirilir.

4.6.1 Rényi Entropisi

Rényi entropisi, Shannon entropisinin tek parametreli bir genellemesidir. $\alpha \neq 1$ olmak üzere $\alpha \geq 0$ parametresi için, α mertebesinden Rényi entropisi şu şekilde tanımlanır:

$$H_\alpha(X) = \frac{1}{1-\alpha} \log \left(\sum_{i=1}^{n} p_i^\alpha \right) \tag{4.15}$$

α parametresi, entropi ölçüsünün hassasiyetini değiştirir. Dikkate değer özel durumlar şunlardır:

- $H_0(X) = \log |\mathcal{X}|$ (Maks-entropi, destek kümesinin boyutuyla ilişkilidir).
- $H_2(X) = -\log(\sum p_i^2)$ (Çarpışma entropisi, iki bağımsız örneğin aynı olma olasılığıyla ilişkilidir).
- $H_\infty(X) = -\log(\max_i p_i)$ (Min-entropi, yalnızca en olası olay tarafından belirlenir).

Rényi entropisi, bir olasılık dağılımının farklı yönlerine odaklanmaya olanak tanıyan bir ölçüm yelpazesi sunar.

4.6.1.1 İspat: Shannon Entropisine Yakınsama

Şimdi, $\alpha \to 1$ limitinde Rényi entropisinin Shannon entropisine yakınsadığını kanıtlamak için

$$\lim_{\alpha \to 1} H_\alpha(X) = \lim_{\alpha \to 1} \frac{\log\left(\sum_{i=1}^{n} p_i^\alpha\right)}{1-\alpha} \tag{4.16}$$

ifadesini ele alalım. $\alpha \to 1$ iken, pay $\log\left(\sum p_i\right) = \log(1) = 0$ olur ve payda da $1 - 1 = 0$ olur. Bu, $\frac{0}{0}$ belirsiz formudur. Bu nedenle L'Hôpital kuralı uygulanabilir. Bunun için payın ve paydanın α'ya göre türevleri alınır.

$f(\alpha) = \log\left(\sum p_i^\alpha\right)$ ve $g(\alpha) = 1 - \alpha$ olsun. Bu durumda

$$g'(\alpha) = -1 \tag{4.17}$$

ve

$$f'(\alpha) = \frac{d}{d\alpha} \log\left(\sum p_i^\alpha\right) = \frac{1}{\left(\sum p_i^\alpha\right) \ln(b)} \cdot \frac{d}{d\alpha}\left(\sum p_i^\alpha\right) \tag{4.18}$$

elde edilir. Burada b, logaritmanın tabanıdır. Ayrıca p_i^α'nın α'ya göre türevi $p_i^\alpha \ln(p_i)$'dir. Dolayısıyla

$$f'(\alpha) = \frac{\sum p_i^\alpha \ln(p_i)}{\left(\sum p_i^\alpha\right) \ln(b)} \tag{4.19}$$

olur. Şimdi L'Hôpital kuralını uygulayalım:

$$\lim_{\alpha \to 1} \frac{f'(\alpha)}{g'(\alpha)} = \lim_{\alpha \to 1} \frac{\frac{\sum p_i^\alpha \ln(p_i)}{\left(\sum p_i^\alpha\right) \ln(b)}}{-1} \tag{4.20}$$

Şimdi limitte $\alpha = 1$ yazılırsa,

$$= \frac{\frac{\sum p_i^1 \ln(p_i)}{\left(\sum p_i^1\right) \ln(b)}}{-1} = -\frac{\sum p_i \ln(p_i)}{\ln(b)} \tag{4.21}$$

elde edilir.

Taban değiştirme formülünü kullanarak $\frac{\ln(x)}{\ln(b)} = \log_b(x)$ olduğundan,

$$= -\sum p_i \log_b(p_i) = H(X) \tag{4.22}$$

sonucuna ulaşılır. Böylece, Rényi entropisinin $\alpha \to 1$ iken Shannon entropisine yakınsadığı kanıtlanmış olur.

4.6.2 Tsallis Entropisi

Tsallis entropisi, non-extensive istatistiksel mekanik alanından doğan bir başka önemli tek parametreli genellemedir. $q \in \mathbb{R}$ parametresi için Tsallis entropisi şu şekilde tanımlanır:

$$S_q(X) = \frac{k}{q-1}\left(1 - \sum_{i=1}^{n} p_i^q\right) \tag{4.23}$$

Burada k pozitif bir sabittir ve genellikle 1 olarak alınır. q parametresi entropik indeks olarak bilinir ve sistemin non-extensivity derecesini ölçer.

4.6.2.1 İspat: Shannon Entropisine Yakınsama

Şimdi, $q \to 1$ limitinde Tsallis entropisinin Shannon entropisine yakınsadığını gösterelim. $k = 1$ olduğunu varsayalım. Bu durumda

$$\lim_{q\to 1} S_q(X) = \lim_{q\to 1} \frac{1 - \sum_{i=1}^{n} p_i^q}{q-1} \tag{4.24}$$

ifadesi elde edilir. $q \to 1$ iken, pay $1 - \sum p_i = 1 - 1 = 0$ olur ve payda da $1 - 1 = 0$ olur. Dolayısıyla yine $\frac{0}{0}$ belirsiz formu ortaya çıkar. Bu nedenle L'Hôpital kuralı uygulanabilir.

$f(q) = 1 - \sum p_i^q$ ve $g(q) = q - 1$ olsun. Bu durumda türevler

$$g'(q) = 1 \tag{4.25}$$

ve

$$f'(q) = -\frac{d}{dq}\left(\sum p_i^q\right) = -\sum \left(p_i^q \ln(p_i)\right) \tag{4.26}$$

şeklindedir.

Şimdi L'Hôpital kuralını uygulayalım:

$$\lim_{q\to 1} \frac{f'(q)}{g'(q)} = \lim_{q\to 1} \frac{-\sum p_i^q \ln(p_i)}{1} \tag{4.27}$$

Şimdi limitte $q = 1$ yazılırsa,

$$= -\sum p_i^1 \ln(p_i) = -\sum p_i \ln(p_i) = H(X) \tag{4.28}$$

sonucu elde edilir. Böylece, Tsallis entropisinin de $q \to 1$ iken Shannon entropisine yakınsadığı gösterilmiş olur.

4.7 Makine Öğrenmesinde Bilgi Kuramının Temel Uygulamaları

Bu bölümde tanımlanan bilgi-kuramsal ölçümler sadece soyut matematiksel yapılar değildir. Makine öğrenmesindeki en temel ve yaygın olarak kullanılan bazı algoritmaların pratik temelini oluştururlar. Bu bölüm, entropi ve ilgili kavramların, tahmin modelleri oluşturmaktan öğrenme sürecini yönlendirmeye kadar aktif olarak uygulandığı birkaç temel alanı incelemektedir.

4.7.1 Karar Ağaçlarında Bölme Kriterleri

Makine öğrenmesinde Shannon entropisinin belki de en klasik uygulaması karar ağaçlarının oluşturulmasındadır. Bir karar ağacı, bir hedef değişkenin değerini

tahmin eden bir model oluşturmak için veriyi öznitelik değerlerine göre yinelemeli olarak alt kümelere ayırır. Her düğümdeki temel zorluk, mümkün olan en "saf" alt düğümlerle (yani, tek bir sınıf tarafından domine edilen düğümler) sonuçlanacak öznitelik ve bölme noktasını seçmektir.

Bu "saf olmama" (impurity) kavramı entropi kullanılarak nicelleştirilir. Algoritma, veriyi bölerek elde edilen entropideki azalma olan *Bilgi Kazancı*'nı (Information Gain) maksimize etmeye çalışır. Bir veri kümesi bölümü D'yi bir A özniteliğine göre bölmenin Bilgi Kazancı şu şekilde tanımlanır:

$$\text{IG}(D, A) = H(D) - \sum_{v \in \text{Values}(A)} \frac{|D_v|}{|D|} H(D_v) \tag{4.29}$$

Burada:

- $H(D)$, ana düğümün veri kümesi D'nin entropisidir.
- Values(A), A özniteliği için tüm olası değerlerin kümesidir.
- D_v, D'nin A özniteliğinin v değerine sahip olduğu alt kümesidir.
- $\frac{|D_v|}{|D|} H(D_v)$, alt düğümlerin ağırlıklı entropisidir.

Her adımda IG(D, A)'yı maksimize eden A özniteliğini seçerek, algoritma hedef sınıf hakkındaki belirsizliği en hızlı şekilde azaltan bir ağacı açgözlü bir şekilde inşa eder. Bu ilke, ID3 ve C4.5 gibi çığır açan algoritmaların temelidir ve Rastgele Ormanlar (Random Forest) ve Gradyan Artırma (Gradient Boosting) gibi modern topluluk yöntemlerinde temel bir kavramdır.

4.7.2 Parametre Tahmini için Amaç Fonksiyonları

Bölüm 4.5'de tartışıldığı gibi, çapraz entropi olasılıksal sınıflandırma modelleri için standart kayıp fonksiyonudur. Bu, aslında, parametre tahmininde daha derin bir ilke olan Maksimum Olabilirlik Kestiriminin (Maximum Likelihood Estimation - MLE) doğrudan bir uygulamasıdır.

Bağımsız örneklerden oluşan bir veri kümesi $\{(\mathbf{x}_i, y_i)\}$ için, MLE'nin amacı, eğitim verisini gözlemleme olasılığını, $P(\mathbf{y}|\mathbf{X}; \theta)$, maksimize eden model parametreleri

θ'yı bulmaktır. Sınıflandırma için bu olabilirlik, doğru sınıfların olasılıklarının çarpımıdır: $L(\theta) = \prod_i p(y_i|\mathbf{x}_i;\theta)$.

Çarpım nedeniyle, log-olabilirliği maksimize etmek hesaplama açısından daha kolaydır:

$$\ell(\theta) = \log L(\theta) = \sum_i \log p(y_i|\mathbf{x}_i;\theta) \tag{4.30}$$

Gerçek etiket y_i'yi bir tek boyutlu(one-hot vector) vektör $\mathbf{y}_{\text{vec},i}$ ve modelin çıktı olasılıklarını $\mathbf{p}_i$ olarak temsil edersek, $\log p(y_i|\mathbf{x}_i;\theta)$ ifadesi $\sum_c y_{\text{vec},i,c} \log(p_{i,c})$'ye eşdeğerdir. Bu nedenle, log-olabilirliği maksimize etmek, matematiksel olarak *negatif log-olabilirliği minimize etmeye* eşdeğerdir, ki bu da tam olarak çapraz entropi kaybının tanımıdır:

$$\arg\max_\theta \ell(\theta) = \arg\min_\theta \left(-\sum_i \log p(y_i|\mathbf{x}_i;\theta) \right) = \arg\min_\theta L_{CE} \tag{4.31}$$

Dolayısıyla, bir sinir ağını sınıflandırma için çapraz entropi kullanarak her eğittiğimizde, dolaylı olarak bilgi kuramı dilinde çerçevelenmiş, istatistiksel parametre tahmininin bir temel taşı olan maksimum olabilirlik kestirimi gerçekleştiriyoruz.

4.7.3 Bir Düzenlileştirici Olarak Entropi

Entropi, (çapraz entropide olduğu gibi) minimize edilecek bir hedefin ötesinde, bir modelin davranışında veya iç temsillerinde arzu edilen özellikleri teşvik etmek için bir kayıp fonksiyonuna düzenlileştirici olarak da eklenebilir.

Shannon entropisinin LAFS çerçevesinde oynadığı rol tam olarak budur. Hibrit kayıp fonksiyonu $L_{\text{total}} = L_{\text{task}} + \lambda H(\mathbf{w})$'da, dikkat ağırlıklarının entropisini, $H(\mathbf{w})$, en aza indirmek, modeli, birkaç özniteliğe yüksek önem atandığı ve geri kalanların sıfıra yakın olduğu, seyrek, "düşük belirsizlikli" bir dağılım üretmeye teşvik eder.

Bu ilke aynı zamanda, hedef $\mathbf{y}$ hakkında maksimum düzeyde bilgilendirici olan bir girdi $\mathbf{x}$'in sıkıştırılmış bir temsili $\mathbf{z}$'yi öğrenmeyi amaçlayan *Bilgi Darboğazı (Information Bottleneck - IB)* yönteminin de merkezindedir. Amaç, $I(\mathbf{z};\mathbf{x})$'i kısıtlarken (minimize ederken) aynı anda $I(\mathbf{z};\mathbf{y})$'yi maksimize etmektir. Temsil ile girdi arasındaki bilgiyi en aza indirmek, modeli ilgisiz bilgileri atmaya zorlayan bir entropik düzenlileştirme şeklidir.

4.7.4 Aktif Öğrenmede Belirsizlik Örneklemesi

Aktif öğrenmede, bir model az miktarda etiketlenmiş veri üzerinde eğitilir ve ardından büyük bir etiketlenmemiş havuzdan yeni, bilgilendirici veri noktalarını etiketlemesi için bir insan araştırmacıya akıllıca sorgular gönderir. En yaygın sorgu stratejilerinden biri *belirsizlik örneklemesidir (uncertainty sampling)* [42].

Burada, model kendi belirsizliğini bir rehber olarak kullanır. Verilen bir etiketlenmemiş örnek $\mathbf{x}$ için, model sınıflar üzerindeki tahmin edilen olasılık dağılımını, $\mathbf{p} = \{p_1, ..., p_C\}$, hesaplar. Bu tahminin belirsizliği, Shannon entropisi $H(\mathbf{p})$ ile nicelleştirilebilir. Yüksek entropili bir tahmin (örneğin, ikili bir problem için $[0.51, 0.49]$), modelin oldukça belirsiz olduğunu gösterirken, düşük entropili bir tahmin (örneğin, $[0.99, 0.01]$), yüksek güveni gösterir. Aktif öğrenme algoritması daha sonra, bir sonraki etiketlenecek örnek olarak en yüksek entropiye sahip olanı seçer, çünkü bu, modelin en çok şey öğreneceği örnektir.

4.8 Bölüm Sonucu

Bu bölüm, bu tez için temel olan bilgi kuramının matematiksel zeminini hazırlamıştır. Shannon entropisi ilkeleri (en aza indirilecek bir belirsizlik ölçüsü olarak) ve çapraz entropi (bir sınıflandırıcıyı optimize etmek için bir kayıp fonksiyonu olarak) LAFS kayıp fonksiyonunun tasarımında doğrudan ve açıkça kullanılmaktadır. Rényi ve Tsallis'in genelleştirilmiş entropileri, Shannon entropisinin daha büyük bir ölçüm ailesine nasıl uyduğunu göstererek ve alternatif entropik düzenleyicilere sahip LAFS varyantlarını keşfetmek gibi gelecekteki araştırmalar için potansiyel yollar önererek daha geniş bir teorik bağlam sağlamaktadır.

5
LAFS ÇERÇEVESİ: TÜREVLENEBİLİR DİKKAT TABANLI BİR YAKLAŞIM

5.1 Gerekçe ve Tasarım Felsefesi

Bölüm 3'te tartışılan öznitelik seçimi yöntemleri, etkili olmalarına rağmen, doğruluk-verimlilik sınırında temel bir ödünleşim sunmaktadır. Sarmalayıcı yöntemler, kapsamlı bir arama yoluyla yüksek performans sunar, ancak hesaplama açısından engelleyicidir; filtre ve gömülü yöntemler ise daha verimlidir, ancak karmaşık öznitelik etkileşimlerini yakalamada başarısız olabilir.

Bu boşluğu gidermek için bu tez, Öznitelik Seçimi için Öğrenilebilir Dikkat (LAFS) çerçevesini önermektedir. Temel tasarım felsefesi, iki ana hedefi birleştirerek seyrek ve bilgilendirici bir öznitelik alt kümesi öğrenen, tek ve uçtan uca türevlenebilir bir model oluşturmaktır: yüksek tahmin doğruluğu ve yüksek öznitelik seyrekliği. Prosedürel, yinelemeli bir arama yerine LAFS, öznitelik seçimini bir öğrenme problemi olarak çerçeveler ve tüm öznitelikleri tek bir ileri geçişte bütünsel olarak değerlendirmek için bir sinirsel dikkat mekanizmasından yararlanır. Tüm süreç, doğrudan Bölüm 4'te detaylandırılan entropi ve çapraz entropi ilkeleri üzerine inşa edilen özel, bilgi-kuramsal bir kayıp fonksiyonu tarafından yönlendirilir.

5.2 Problemin Resmi Formülasyonu

Model mimarisini detaylandırmadan önce, problemi resmi olarak yeniden ifade edelim. Her bir $\mathbf{x}_i \in \mathbb{R}^d$ örneğinin d öznitelikten oluşan bir vektör ve y_i'nin buna karşılık gelen sınıf etiketi olduğu bir $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ veri kümesi verildiğinde, amacımız, $\{1, ..., d\}$ kümesinin bir S öznitelik indisleri alt kümesini belirlemektir. Bu alt kümenin boyutu $|S| = k \ll d$ olmalı, öyle ki, azaltılmış öznitelik vektörleri $\mathbf{x}_S$ üzerinde eğitilen bir model en uygun tahmin performansını elde eder.

5.3 LAFS Mimarisi

LAFS modeli, birlikte eğitilen iki ana alt ağdan oluşan bir sinir ağıdır: *Dikkat Puanlayıcı Ağı (AttnNet)* ve *Sınıflandırıcı Ağı (ClassNet)*. Genel mimari şekil 5.1'de gösterilmektedir. Veri, ağ üzerinden önce öznitelik önemlerini öğrenerek ve ardından bu önemleri bir tahmin yapmak için kullanarak akar; tüm sistem uçtan uca optimize edilir.

5.3.1 Dikkat Puanlayıcı Ağı (AttnNet)

AttnNet, dinamik ve örneğe özgü bir öznitelik önem maskesi öğrenmekten sorumludur. Basit filtre yöntemlerinin başarısız olduğu bir görev olan öznitelikler arasındaki potansiyel olarak karmaşık ve doğrusal olmayan ilişkileri yakalamak için AttnNet'i bir Çok Katmanlı Algılayıcı (Multi-Layer Perceptron - MLP) olarak uygulamaktadır.

Bir girdi örneği $\mathbf{x} \in \mathbb{R}^d$ için, MLP tabanlı AttnNet önce girdiyi gizli bir temsile yansıtır ve ardından normalize edilmemiş ön-softmax skorları olan bir $\mathbf{s} \in \mathbb{R}^d$ vektörü üretir. Bu skorlar daha sonra öznitelikler üzerinde bir olasılık dağılımı üretmek için softmax fonksiyonu kullanılarak normalize edilir; bu dağılımı dikkat vektörü $\mathbf{w} \in \mathbb{R}^d$ olarak adlandırılmaktadır:

$$w_j = \frac{\exp(s_j)}{\sum_{l=1}^{d} \exp(s_l)} \quad \text{için } j = 1, ..., d \tag{5.1}$$

Her bir w_j bileşeni, verilen girdi örneği $\mathbf{x}$ için j-inci özniteliğin öğrenilmiş önemini temsil eder ve $\sum_{j=1}^{d} w_j = 1$ özelliğine sahiptir. Doğrusal olmayan bir MLP'nin ifade gücü, AttnNet'in daha yüksek dereceli öznitelik etkileşimlerini modellemesine olanak tanır, bu da bir özniteliğin öneminin başka bir özniteliğin değerine bağlı olabileceğini anlamasını sağlar. Bu artan ifade gücüyle ilişkili aşırı öğrenme riskini yönetmek için AttnNet içinde, sırasıyla karmaşık eş-uyarlamaları ve büyük ağırlıkları cezalandıran, spesifik olarak Seyreltme (Dropout) ve L2 düzenlileştirmesi (Ağırlık Azaltma - Weight Decay) gibi standart düzenlileştirme tekniklerini kullanmaktadır.

5.3.2 Sınıflandırıcı Ağı (ClassNet)

AttnNet tarafından üretilen dikkat vektörü $\mathbf{w}$, daha sonra orijinal öznitelik vektörü $\mathbf{x}$'i yeniden ağırlıklandırmak için kullanılır. Bu, önemli olduğu düşünülen öznitelikleri vurgulayan ve ilgisiz olduğu düşünülenleri bastıran bir "bağlam vektörü (context vector)" $\tilde{\mathbf{x}}$ üretir. Bu, eleman bazında çarpma (Hadamard çarpımı, $\odot$) ile elde edilir:

$$\tilde{\mathbf{x}} = \mathbf{w} \odot \mathbf{x} \tag{5.2}$$

Bu bağlam vektörü $\tilde{\mathbf{x}}$, daha sonra nihai sınıf tahmin logit'leri olan $\mathbf{z}$'yi üretmek için sınıflandırma amaçlı standart bir çok katmanlı algılayıcı (MLP) olan ClassNet'e beslenir.

5.4 Hibrit Bilgi-Kuramsal Kayıp Fonksiyonu

LAFS'ı seyrek ve etkili bir öznitelik maskesi öğrenmeye yönlendirmenin anahtarı, özel kayıp fonksiyonumuzda yatmaktadır. Eğitim süreci, aynı anda yüksek doğruluk ve yüksek seyrekliği hedefleyen çok amaçlı bir optimizasyon problemidir. Bu durum, bir görev-performansı terimi ile bir bilgi-kuramsal düzenlileştirme teriminin ağırlıklı toplamı olan hibrit kayıp fonksiyonumuzla formüle edilmiştir:

$$L_{\text{total}} = L_{\text{task}} + \lambda L_{\text{info}} \tag{5.3}$$

λ hiperparametresi, bu iki rakip hedef arasındaki ödünleşimi yöneten kritik bir kontrol parametresidir.

5.4.1 Görev Kaybı (L_{task})

Görev kaybı, modelin doğru olmayı öğrenmesini sağlar. Bölüm 4'te belirtildiği gibi, olasılıksal bir sınıflandırıcı için standart hedef, çapraz entropi kaybıdır. Gerçek one-hot sınıf etiket vektörü $\mathbf{y}_{\text{vec}}$ ve ClassNet'in softmax çıktısı $\mathbf{p}$ verildiğinde, görev kaybı şöyledir:

$$L_{\text{task}} = H(\mathbf{y}_{\text{vec}}, \mathbf{p}) = -\sum_{c=1}^{C} y_{\text{vec},c} \log(p_c) \tag{5.4}$$

burada C sınıf sayısıdır. Bu terimi en aza indirmek, modeli sınıflandırma görevi için maksimum düzeyde bilgilendirici bir öznitelik öğrenmeye zorlar.

5.4.2 Bilgi-Kuramsal Düzenlileştirici (L_{info})

Hedefimizin özgün bileşeni, bilgi-kuramsal düzenlileştirici olan L_{info}'dur. Bu terim, Bölüm 4'te tanımlandığı gibi, basitçe dikkat vektörü $\mathbf{w}$'nun *Shannon entropisidir*:

$$L_{\text{info}} = H(\mathbf{w}) = -\sum_{j=1}^{d} w_j \log(w_j) \tag{5.5}$$

Bu entropi terimini *en aza indirerek*, modeli açıkça düşük entropili, "keskin" dikkat dağılımları üretmeye teşvik ederiz. Düşük entropili bir olasılık dağılımı, olasılık kütlesinin birkaç sonuç üzerinde yoğunlaştığı bir dağılımdır. Bu bağlamda, modelin birçok özniteliğe vasat dikkat ağırlıkları ataması cezalandırılır ve birkaç özniteliğe yüksek ağırlıklar atarken geri kalanları sıfıra doğru itmesi ödüllendirilir. Bu, doğrudan seyrek bir öznitelik alt kümesinin seçimini teşvik eder.

5.4.3 Tam Öğrenme Hedefi

Bu bileşenleri birleştirerek, tam öğrenme hedefi, aşağıdaki optimizasyon problemini çözen en uygun model parametreleri Θ^*'ı bulmaktır:

$$\Theta^* = \arg\min_{\Theta} \mathbb{E}_{(\mathbf{x},y)\sim D} \left[L_{\text{task}}(\mathbf{p}, \mathbf{y}_{\text{vec}}) + \lambda H(\mathbf{w}) \right] \tag{5.6}$$

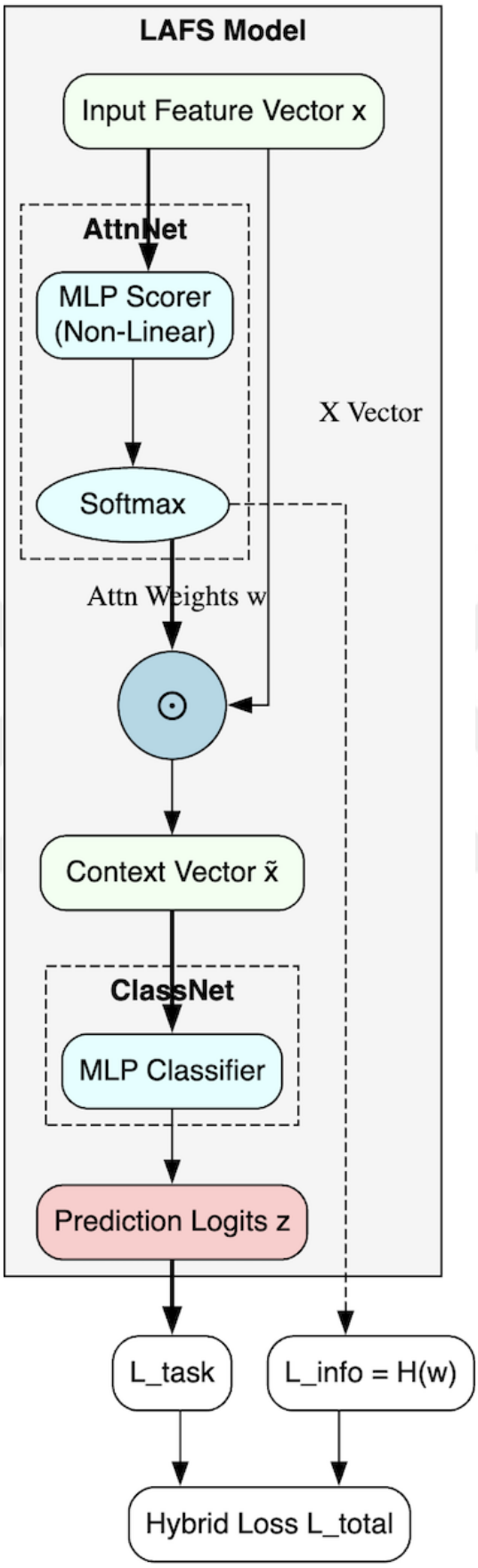
Bu formülasyon, optimize ediciye, doğru bir tahmin ihtiyacını, seyrek ve belirleyici bir dikkat maskesinin yapısal kısıtıyla dengeleyen bir çözüm bulmasını sağlar.

5.5 Global Öznitelik Öneminin Çıkarılması

Model uçtan uca eğitildikten sonra, dikkat vektörleri $\mathbf{w}^{(i)}$ örneğe özgüdür. Tüm veri kümesi için öznitelik öneminin tek bir global sıralamasını elde etmek amacıyla, eğitim setindeki tüm n örneğin örneğe özgü dikkat vektörlerinin ortalamasını alarak bir global öznitelik önemi vektörü $\mathbf{I} \in \mathbb{R}^d$ hesaplarız:

$$I_j = \frac{1}{n} \sum_{i=1}^{n} w_j^{(i)} \tag{5.7}$$

Nihai öznitelik alt kümesi S, daha sonra en yüksek global önem skorlarına I_j sahip k adet özniteliğin seçilmesiyle elde edilir. Bu, eğitilmiş modelden net, yorumlanabilir bir çıktı sağlar.



Şekil 5.1 LAFS mimarisi; girdi öznitelik vektöründen AttnNet ve ClassNet üzerinden nihai hibrit kayıp hesaplamasına olan akışı göstermektedir.

6
DENEYSEL DOĞRULAMA

6.1 Bölüme Giriş ve Deneysel Amaçlar

Bölüm 5'te LAFS çerçevesini resmi olarak tanımladıktan sonra, bu bölüm onun kapsamlı ampirik doğrulamasına adanmıştır. Temel amaç, LAFS'ın performansını, yerleşik ve alanında en iyi (state-of-the-art) öznitelik seçimi algoritmalarına karşı, çok çeşitli zorluklar içeren bir yelpazede titizlikle değerlendirmektir. Deneyler, üç temel araştırma sorusuna yanıt vermek üzere tasarlanmıştır:

1. *LAFS, karmaşık ve doğrusal olmayan öznitelik etkileşimlerini etkili bir şekilde belirleyebilir mi?* Birçok gerçek dünya problemi, yalnızca birlikte ele alındığında tahminsel güce sahip olan öznitelikler içerir. Bunu, bu yetenek için klasik bir referans olan sentetik bir XOR veri seti kullanarak test edilmiştir.

2. *LAFS, yüksek çoklu doğrusallık koşulları altında nasıl bir performans sergilemektedir?* Yüksek önem arz etmeyen öznitelikler model performansını ciddi şekilde düşürebilir. Bu tip bir öznitelik alt kümesi seçmede bilgi-kuramsal düzenleyicinin etkinliğini değerlendirmek için kontrollü, tekdüze korelasyona sahip sentetik bir veri seti kullanılmıştır.

3. *LAFS, çok çeşitli gerçek dünya problemlerinde rekabetçi midir?* Genellenebilirliğini ve pratik kullanışlılığını değerlendirmek için, geniş bir özellik yelpazesini (örneğin, düşük-p/düşük-n, yüksek-p/düşük-n) kapsayan dokuz adet halka açık veri seti üzerinde büyük ölçekli bir karşılaştırma çalışması yürütülmüştür.

Bu bölüm, bu üç temel deneyin her birinin sonuçlarını sunmadan ve analiz etmeden önce, karşılaştırma algoritmaları ve değerlendirme metrikleri de dahil olmak üzere deneysel kurguyu detaylandıracaktır.

6.2 Deneysel Kurulum

Adil ve sağlam bir karşılaştırma sağlamak için tutarlı bir deneysel kurulum tanımlanmıştır.

6.2.1 Karşılaştırma Algoritmalarının Detayları

Karşılaştırma için net ve resmi bir temel sağlamak amacıyla, bu bölümde karşılaştırmamızda kullanılan temel sarmalayıcı, gömülü ve hibrit algoritmaların prosedürel adımları detaylandırılmaktadır. Filtre yöntemleri (MI ve mRMR) zaten Bölüm 3'te detaylandırılmıştır.

6.2.1.1 Sarmalayıcı Yöntem: Özyinelemeli Öznitelik Eleme (RFE)

RFE, geri eleme stratejisi kullanan arketipsel bir sarmalayıcı yöntemdir. Başlangıçta Destek Vektör Makineleri ile gen seçimi için popülerleştirilen bu yöntem, yinelemeli olarak küçülen bir öznitelik kümesi üzerinde bir model eğitir ve her adımda en önemsiz olanı kaldırır. Bu süreç algoritma 6.1'da formüle edilmiştir. Güçlü olmasına rağmen, temel dezavantajı, bir modelin $d - k$ kez eğitilmesini gerektirdiği için yüksek hesaplama maliyetidir.

Algoritma 6.1 Özyinelemeli öznitelik eleme (Recursive Feature Elimination – RFE)

Require: D: $\mathcal{F}$ öznitelikleri ve Y hedefini içeren bir veri seti. $\mathcal{A}$: Her özniteliğe bir önem skoru atayabilen bir öğrenme algoritması. k: İstenen nihai öznitelik sayısı.

Ensure: S: En iyi k özniteliği içeren $\mathcal{F}$'nin bir alt kümesi.

1: Mevcut öznitelik kümesini $S_c \leftarrow \mathcal{F}$ olarak başlat.
2: **while** $|S_c| > k$ **do**
3: Öğrenme algoritması $\mathcal{A}$'yı, sadece S_c'deki öznitelikleri kullanarak veri seti üzerinde eğit.
4: S_c'deki her bir F_j özniteliği için bir w_j önem skoru hesapla.
5: En önemsiz özniteliği bul: $F^- \leftarrow \arg\min_{F_j \in S_c} w_j$.
6: En önemsiz özniteliği ele: $S_c \leftarrow S_c \setminus \{F^-\}$.
7: **end while**
8: **return** S_c

6.2.1.2 Gömülü Yöntem: Tavlama ile Öznitelik Seçimi (FSA)

FSA, öznitelik seçimini doğrudan tek bir model eğitimi sürecine entegre eden modern bir gömülü yöntemdir. RFE gibi sıfırdan yeniden eğitmek yerine, FSA

eğitim sırasında en önemsiz öznitelikleri aşamalı olarak elemek için deterministik bir "tavlama" programı kullanır. Prosedür algoritma 6.2'da detaylandırılmıştır.

Algoritma 6.2 Tavlama ile öznitelik seçimi (Feature Selection with Annealing – FSA)

Require: D: $\mathcal{F} = \{F_1, ..., F_d\}$ özniteliklerine sahip bir veri seti. $\mathcal{A}$: Öğrenilebilir öznitelik ağırlıkları $\mathbf{w}$ olan bir öğrenme algoritması. k: İstenen nihai öznitelik sayısı. T: Eleme adımları arasında gerçekleştirilecek eğitim epok sayısı.

Ensure: S_k: k öznitelikten oluşan nihai alt küme.

1: Öğrenilebilir öznitelik ağırlıklarını $\mathbf{w} \leftarrow \{w_1, ..., w_d\}$ olarak başlat (örneğin, hepsi 1'e ayarlı).
2: Aktif özniteliklerin kümesini $S_{active} \leftarrow \mathcal{F}$ olarak başlat.
3: **for** $k_{current} = d - 1$ down to k **do**
4: Model $\mathcal{A}$'yı, S_{active}'deki tüm F_j öznitelikleri için w_j ağırlıklarını optimize ederek veri seti üzerinde T epok boyunca eğit.
5: ▷ Eğitimden sonra, en önemsiz aktif özniteliği bul.
6: Elenecek özniteliği bul: $F^- \leftarrow \arg\min_{F_j \in S_{active}} |w_j|$.
7: Özniteliği aktif kümeden çıkar: $S_{active} \leftarrow S_{active} \setminus \{F^-\}$.
8: ▷ Ağırlığını sıfırda dondurarak özniteliği kalıcı olarak ele.
9: F^-'ye karşılık gelen ağırlığı (w^-) 0'a ayarla ve dondur.
10: **end for**
11: ▷ İsteğe bağlı: Modelin nihai ince ayarını yap.
12: Model $\mathcal{A}$'yı, yalnızca S_{active}'deki öznitelikleri kullanarak son bir epok sayısı kadar eğitmeye devam et.
13: **return** S_{active}

6.2.1.3 Hibrit Yöntem: Bulanık Rastgele Orman Öznitelik Seçimi (FRF-fs)

FRF-fs, filtre, gömülü ve sarmalayıcı yaklaşımlarının unsurlarını karmaşık, üç aşamalı bir boru hattında birleştiren sofistike bir hibrit yöntemdir. Yöntemin özü, bulanık bölümleme yoluyla sürekli verileri daha iyi işlemek için geleneksel rastgele orman algoritmasını genişleten Bulanık Rastgele Orman modelidir [43]. Algoritma 6.3'te formüle edilen genel öznitelik seçimi prosedürü, sürekli öznitelikleri etkili bir şekilde ele almak ve sağlam bir nihai öznitelik alt kümesi bulmak için tasarlanmıştır.

Algoritma 6.3'ün gösterdiği gibi, FRF-fs yöntemi, belirgin ve karmaşık süreçlerin bir dizisidir. Güçlü olmasına rağmen, çok aşamalı yapısı, onun tamamen bağlantılı, uçtan uca türevlenebilir bir sistem olmadığı anlamına gelir. Böyle bir uygulama ve ayarlama karmaşıklığı, tek bir, entegre eğitim süreci içinde benzer hedeflere ulaşmayı amaçlayan LAFS gibi birleşik bir çerçevenin cazibesini artırmaktadır.

Algoritma 6.3 Bulanık rastgele orman öznitelik seçimi (FRF-fs) algoritması

Require: D: $\mathcal{F}$ öznitelikleri ve Y hedefini içeren bir veri seti. k: İstenen nihai öznitelik sayısı. $\mathcal{A}_{eval}$: Nihai sarmalayıcı aşaması için bir değerlendirme sınıflandırıcısı (örneğin, XGBoost).

Ensure: S_{best}: k öznitelikten oluşan nihai alt küme.

▷ *Aşama 1: Bulanık Bölümleme (Filtre Aşaması)*

Boş bir bulanık bölümleme kümesi $\mathcal{P}$ başlat.

for $\mathcal{F}$'deki her sürekli F_j özniteliği için **do**

    F_j için en uygun bulanık bölümleme kümesi P_j'yi bulmak için bir Genetik Algoritma kullan.

    $\mathcal{P} \leftarrow \mathcal{P} \cup \{P_j\}$.

end for

Orijinal veri kümesi D'yi, $\mathcal{P}$ bölümlerini kullanarak bulanıklaştırılmış bir D_{fuzzy} veri kümesine dönüştür.

▷ *Aşama 2: Öznitelik Sıralama (Gömülü Aşama)*

Bulanıklaştırılmış D_{fuzzy} veri kümesi üzerinde bir Bulanık Rastgele Orman modeli eğit.

Toplanmış bulanık bilgi kazancına dayalı olarak öznitelik önem skorlarını çıkar.

$\mathcal{F}$'deki tüm öznitelikleri skorlarına göre sıralayarak bir L sıralı listesi oluştur.

▷ *Aşama 3: Nihai Seçim (Sarmalayıcı Aşaması)*

Nihai küme $S_{final} \leftarrow \emptyset$ ve en iyi skor $score_{best} \leftarrow -\infty$ olarak başlat.

for $i = 1$ to k **do**

    $F_{next} \leftarrow$ sıralı liste L'deki i-inci öznitelik.

    $S_{current} \leftarrow S_{final} \cup \{F_{next}\}$.

    Değerlendirme sınıflandırıcısı $\mathcal{A}_{eval}$'ı, sadece $S_{current}$'teki öznitelikleri kullanarak veri üzerinde eğit.

    Bir doğrulama seti üzerinde $score_{current}$ performansını değerlendir.

    if $score_{current} > score_{best}$ **then**

        $score_{best} \leftarrow score_{current}$.

        $S_{best} \leftarrow S_{current}$.

    end if

    $S_{final} \leftarrow S_{current}$. ▷ Bu satır basit bir en-iyi-k içindir, If bloğu SFS içindir.

end for

return S_{best} ▷ Veya tam bir SFS sarmalayıcısı kullanılmıyorsa sadece en iyi k öznitelik.

6.2.2 Değerlendirme Metrikleri

Seçilen bir öznitelik alt kümesinin kalitesi, bir alt akış sınıflandırıcısının performansına göre değerlendirilir. Tüm deneylerde bu değerlendirme görevi için standart bir XGBoost sınıflandırıcısı kullanıyoruz. Başlıca metrikler şunlardır:

- *F1-Skoru:* Kesinlik (precision) ve duyarlılığın (recall) harmonik ortalaması olup, özellikle dengesiz veri setlerinde bir modelin doğruluğunun dengeli bir ölçüsünü sağlar.
- *AUC (ROC Eğrisi Altındaki Alan):* Bir modelin sınıflar arasında ayrım yapma yeteneğinin bir ölçüsüdür ve sınıflandırma eşiğinden bağımsızdır.
- *VDR (Değişken Keşif Oranı):* Yalnızca temel gerçek (ground-truth) ilgili özniteliklerin bilindiği sentetik XOR deneyi için kullanılır. Doğru şekilde tanımlanan ilgili özniteliklerin sayısını ölçer.

6.3 Deney 1: Doğrusal Olmayan Etkileşimler Üzerindeki Performans (XOR Veri Seti)

6.3.1 Amaç ve Veri Üretimi

Bu deney, algoritmaların, değişkenlerin tek başlarına işe yaramazken toplu olarak tahminsel güce sahip olduğu saf yüksek dereceli etkileşimleri tespit etme yeteneğini test eder. İlk üç özniteliğin (x_0, x_1, x_2) "ilgili" olarak belirlenmiştir, $d = 15$ özniteliğe sahip 3-bitlik bir parite (XOR) veri seti ürettik. Hedef sınıf y, bu özniteliklerin çarpımının işaretiyle, $y = \mathbb{I}(\prod_{i=0}^{2} x_i > 0)$, belirlendi; bu, tek değişkenli yöntemlerin tespit edemeyeceği bir etkileşimdir.

6.3.2 Sonuçlar ve Analiz

Artan örneklem büyüklükleri (n) üzerindeki performans, tablo 6.1 ila tablo 6.5'de özetlenmiştir. Düşük örneklem büyüklüklerinde ($n = 100$, tablo 6.1), hiçbir algoritma karmaşık üç yönlü sinyali rastgele gürültüden güvenilir bir şekilde ayırt edememekte ve tüm yöntemler rastgele tahmine yakın performans göstermektedir (F1-Skoru ≈ 0.5). Örneklem büyüklüğü arttıkça ($n = 500$ ve $n = 1000$, tablo 6.2 ve tablo 6.3), performanslar ayrışmaktadır. Güçlü sarmalayıcı yöntem olan RFE-LGBM, etkileşim sinyalini önemli ölçüde tespit etmeye başlamaktadır. Rastgele performansa yakın kalan MI ve mRMR gibi filtre yöntemlerini kesin bir şekilde geride bırakmaktadır. Bu, AttnNet içindeki doğrusal olmayan MLP'nin öznitelik etkileşimlerini başarılı bir şekilde yakaladığını doğrulamaktadır.

Yeterli veri seviyelerinde ($n \geq 2000$, tablo 6.4 ve tablo 6.5), RFE-LGBM neredeyse mükemmel bir tespit oranına ulaşmaktadır. LAFS, gradyan artırma modeline kıyasla tam olarak yakınsamak için daha fazla örneğe ihtiyaç duysa da, bu tür problemler için geleneksel filtre ve doğrusal yöntemlere göre üstünlüğünü teyit eden sağlam bir tespit oranını sunmaktadır.

Tablo 6.1 XOR veri seti için karşılaştırma sonuçları ($n = 100$).

Yöntem	*VDR*	*F1 Skoru*	*AUC*
MI	0.6800	0.4798	0.5034
mRMR	0.7400	0.4617	0.4829
RFE-LGBM	0.5400	0.4735	0.4854
FSA	0.7800	0.4628	0.4843
FRF-fs	0.0200	0.0089	0.5012
LAFS	0.7400	0.4487	0.4747

Tablo 6.2 XOR veri seti için karşılaştırma sonuçları ($n = 500$).

Yöntem	*VDR*	*F1 Skoru*	*AUC*
MI	0.5600	0.4839	0.4816
mRMR	0.8200	0.4883	0.4913
RFE-LGBM	2.1200	0.8026	0.8219
FSA	0.6000	0.4973	0.4933
FRF-fs	0.0000	0.0000	0.5000
LAFS	1.2200	0.4911	0.4940

Tablo 6.3 XOR veri seti için karşılaştırma sonuçları ($n = 1000$).

Yöntem	*VDR*	*F1 Skoru*	*AUC*
MI	0.4400	0.4982	0.4990
mRMR	0.5600	0.5068	0.4995
RFE-LGBM	3.0000	0.9786	0.9974
FSA	0.5000	0.4995	0.4954
FRF-fs	0.0000	0.0000	0.5000
LAFS	2.1600	0.6858	0.6912

Tablo 6.4 XOR veri seti için karşılaştırma sonuçları ($n = 2000$).

Yöntem	*VDR*	*F1 Skoru*	*AUC*
MI	0.5200	0.5016	0.5000
mRMR	0.7600	0.4932	0.4960
RFE-LGBM	3.0000	0.9910	0.9994
FSA	0.5800	0.5117	0.5081
FRF-fs	0.0000	0.0000	0.5000
LAFS	2.1800	0.6542	0.6526

Tablo 6.5 XOR veri seti için karşılaştırma sonuçları ($n = 5000$).

Yöntem	*VDR*	*F1 Skoru*	*AUC*
MI	0.6200	0.5161	0.5196
mRMR	0.6200	0.4962	0.4954
RFE-LGBM	2.9800	0.9853	0.9895
FSA	0.5600	0.4961	0.4970
FRF-fs	0.0000	0.0000	0.5000
LAFS	1.7800	0.5828	0.5824

6.4 Deney 2: Yüksek Çoklu Doğrusallık Altındaki Performans (SYNTH Veri Seti)

6.4.1 Amaç ve Veri Üretimi

Bu deney, algoritmaların yüksek derecede ilişkili özniteliklere karşı dayanıklılığını değerlendirir. Gizli bir değişken modeli kullanılarak tüm öznitelikler arasında tekdüze bir çiftli korelasyon ($\rho = 0.5$) oluşturulan, $p = 100$ öznitelikli sentetik bir veri seti ürettik. Hedef, bu özniteliklerin doğrusal bir kombinasyonudur.

6.4.2 Sonuçlar ve Analiz

Sonuçlar tablo 6.6'da sunulmuştur. Bu deney, LAFS'taki entropik düzenleyicinin etkinliğini vurgulamaktadır. Yüksek çoklu doğrusallığın varlığında, LAFS, farklı öznitelik alt kümesi boyutlarında ($k = 10, 30, 50$) güçlü RFE-LGBM ve FSA yöntemleriyle karşılaştırılabilir sonuçlar elde ederek sürekli olarak en üstte veya en üste yakın performans göstermektedir. Bu, temel tasarım ilkemizi doğrulamaktadır: Dikkat ağırlıklarının entropisini en aza indirmek, birçok öznitelik benzer bilgiyi taşısa bile, modeli seyrek ve gereksiz olmayan bir alt küme seçmeye teşvik eder.

Tablo 6.6 SYNTH veri seti için, farklı k değerlerine göre benchmark sonuçları.

Algorithm	*k = 10*		*k = 30*		*k = 50*	
	F1	*AUC*	*F1*	*AUC*	*F1*	*AUC*
FRF-fs	0.0000	0.5000	0.0000	0.5000	0.0000	0.5000
FSA	0.6681	0.7651	0.7149	0.7858	0.7180	0.8059
LAFS	0.6814	0.7546	0.7200	0.7926	0.7544	0.8448
MI	0.6764	0.7562	0.7052	0.7881	0.7306	0.8129
RFE-LGBM	0.7082	0.7911	0.7703	0.8576	0.7634	0.8576
mRMR	0.6693	0.7616	0.7095	0.7835	0.7238	0.8057

6.5 Deney 3: Gerçek Dünya Veri Setleri Üzerinde Büyük Ölçekli Karşılaştırma

6.5.1 Amaç ve Veri Seti Seçimi

LAFS'ın genellenebilirliğini değerlendirmek için, PMLB ve UCI gibi depolardan dokuz adet halka açık, çeşitli referans veri setleri seçtik. Tablo 6.7'de özetlenen bu veri setleri, "düşük p—düşük n" (örneğin, 'zoo'), "yüksek p—düşük n" (örneğin, 'arcene') ve "yüksek p—yüksek n" (örneğin, 'gisette') senaryoları dahil olmak üzere geniş bir özellik yelpazesini kapsayacak şekilde kasıtlı olarak seçilmiştir.

Tablo 6.7 Kapsamlı değerlendirmede kullanılan dokuz referans veri setine genel bakış.

Veri Seti	*Satır (n)*	*Öznitelik (p)*	*k1*	*k2*	*k3*
zoo	101	16	3	5	10
spectf	267	44	3	5	10
australian	690	14	3	5	10
spambase	4601	57	10	30	50
tic_2000	9822	85	10	30	50
miniboone	130,065	50	10	30	50
lsvt	126	309	10	15	30
arcene	200	10,000	10	500	1000
gisette	7000	5000	10	250	500

6.5.2 Değerlendirme Protokolü

Her veri seti için 5-katlı çapraz doğrulama gerçekleştirilmiştir. Her katlamada, her algoritma, üç farklı seyreklik baskısına (tablo 6.7'den k_1, k_2, k_3) göre öznitelik seçmekle görevlendirilmiştir. Seçilen alt kümenin kalitesi daha sonra bir alt akış XGBoost sınıflandırıcısı eğitilerek ve ayrılmış doğrulama seti üzerindeki F1-Skoru ve AUC'si ölçülerek değerlendirilmiştir.

6.5.3 Sonuçlar ve Analiz

Detaylı sonuçlar tablo 6.8'de sunulmuştur. Küçük ve orta büyüklükteki özniteliklere sahip veri setlerinde (örneğin, 'zoo', 'australian', 'spambase', 'miniboone'), LAFS kendini oldukça rekabetçi bir yöntem olarak sağlam bir şekilde kanıtlamaktadır. Sürekli olarak, alanındaki en iyi sarmalayıcı RFE-LGBM ve gömülü FSA yöntemlerine çok yakın ve bazı durumlarda onları aşan F1-Skorları ve AUC değerleri elde etmektedir. Bu, yaygın makine öğrenmesi problemlerinin geniş bir yelpazesi için pratik kullanışlılığını göstermektedir. LAFS'ın önemli bir sınırlılığı, çok yüksek öznitelik-örneklem oranına sahip senaryolarda ortaya çıkmaktadır. 'arcene' ($p = 10,000, n = 200$) ve 'gisette' ($p = 5,000, n = 7000$) veri setlerinde, LAFS önde gelen algoritmaların performansına ulaşamamaktadır. Bu, bu tür düşük örneklemli rejimlerde dikkat mekanizmasının yüksek varyansının bir zorluk olabileceğini düşündürmekte ve gelecek araştırmalar için net ve umut verici bir yön sunmaktadır.

Tablo 6.8 Tüm referans veri setleri ve seyreklik seviyeleri (k1, k2 ve k3) için her algoritmaya ait detaylı F1-skorları ve AUC skorları.

Veri Seti	*Algoritma*	*k1*		*k2*		*k3*	
		F1	*AUC*	*F1*	*AUC*	*F1*	*AUC*
ZOO	FRF-fs	0.5608	0.8705	0.7737	0.9532	0.8991	0.9867
	FSA	0.6672	0.9341	0.8261	0.9803	0.9325	0.9987
	LAFS	0.6898	0.9451	0.8982	0.9880	0.9418	0.9943
	MI	0.5626	0.8747	0.7866	0.9670	0.9537	0.9977
	RFE-LGBM	0.7107	0.9119	0.9120	0.9951	0.9129	0.9955
	mRMR	0.7308	0.9474	0.8730	0.9790	0.9057	0.9978
AUSTRALIAN	FRF-fs	0.8242	0.8467	0.8262	0.9126	0.7963	0.8731
	FSA	0.8504	0.9137	0.7957	0.9012	0.8426	0.9149
	LAFS	0.8191	0.9001	0.7892	0.8877	0.8293	0.9121
	MI	0.8201	0.9103	0.7962	0.9028	0.8484	0.9249
	RFE-LGBM	0.5395	0.6237	0.6605	0.7789	0.8487	0.9212
	mRMR	0.8329	0.9051	0.8032	0.9023	0.8462	0.9143
SPECTF	FRF-fs	0.0000	0.5000	0.0000	0.5000	0.0000	0.5000
	FSA	0.8936	0.9032	0.9157	0.9288	0.9280	0.9222
	LAFS	0.8882	0.8923	0.9252	0.9345	0.9311	0.9305
	MI	0.9005	0.9006	0.9178	0.9098	0.9309	0.9381
	RFE-LGBM	0.9106	0.9371	0.9346	0.9380	0.9276	0.9303
	mRMR	0.9039	0.8988	0.9145	0.9289	0.9364	0.9418
SPAMBASE	FRF-fs	0.6472	0.8371	0.7706	0.9089	0.7044	0.8787
	FSA	0.9047	0.9679	0.9315	0.9847	0.9380	0.9866
	LAFS	0.9026	0.9693	0.9227	0.9810	0.9370	0.9868
	MI	0.9105	0.9689	0.9307	0.9828	0.9392	0.9870
	RFE-LGBM	0.8844	0.9592	0.9346	0.9859	0.9409	0.9871
	mRMR	0.8999	0.9663	0.9314	0.9853	0.9379	0.9865
MINIBOONE	FRF-fs	0.0000	0.5000	0.0000	0.5000	0.0000	0.5000
	FSA	0.8316	0.9619	0.8903	0.9814	0.9010	0.9844
	LAFS	0.8358	0.9641	0.8736	0.9770	0.9010	0.9844
	MI	0.8385	0.9651	0.8758	0.9774	0.9010	0.9844
	RFE-LGBM	0.8724	0.9761	0.8971	0.9838	0.9010	0.9844
	mRMR	0.8405	0.9648	0.8912	0.9818	0.9010	0.9844
TIC_2000	FRF-fs	0.0000	0.5344	0.0000	0.5351	0.0033	0.5780
	FSA	0.0597	0.6993	0.1057	0.7119	0.0992	0.7069
	LAFS	0.0584	0.6349	0.0789	0.6723	0.1041	0.7055
	MI	0.0847	0.6973	0.0962	0.7045	0.1014	0.7091
	RFE-LGBM	0.0843	0.6424	0.0989	0.6951	0.1065	0.7072
	mRMR	0.0610	0.6983	0.0980	0.7143	0.0968	0.7058
LSVT	FRF-fs	0.8438	0.8431	0.8393	0.8108	0.8367	0.8469
	FSA	0.8646	0.8897	0.8669	0.8917	0.8722	0.8886
	LAFS	0.8619	0.8652	0.9065	0.8853	0.8988	0.9016
	MI	0.8202	0.8558	0.8365	0.8909	0.8551	0.8793
	RFE-LGBM	0.8720	0.8472	0.8809	0.8821	0.8984	0.9051
	mRMR	0.8619	0.8848	0.8577	0.9015	0.8687	0.9097
ARCENE	FRF-fs	0.0211	0.5346	0.2801	0.5959	0.1283	0.5482
	FSA	0.7190	0.8576	0.7209	0.8594	0.7190	0.8326
	LAFS	0.6297	0.7799	0.7848	0.9039	0.6858	0.8402
	MI	0.7562	0.8142	0.8037	0.9086	0.7725	0.9005
	RFE-LGBM	0.6647	0.7968	0.7566	0.8701	0.7512	0.8787
	mRMR	0.7078	0.8071	0.6957	0.8360	0.7461	0.8522
GISETTE	FRF-fs	0.0398	0.5103	0.0473	0.5105	0.1649	0.5101
	FSA	0.9291	0.9720	0.9787	0.9968	0.9811	0.9970
	LAFS	0.8933	0.9493	0.9643	0.9930	0.9685	0.9952
	MI	0.9074	0.9493	0.9763	0.9955	0.9773	0.9962
	RFE-LGBM	0.9152	0.9620	0.9797	0.9970	0.9806	0.9970
	mRMR	0.9171	0.9621	0.9751	0.9954	0.9779	0.9964

6.6 Hesaplamalı Verimlilik Analizi (Çalışma Zamanları)

Bu tezin temel iddialarından biri, LAFS'in, güçlü sarmalayıcı (wrapper) metotların performansını daha basit modellerin hesaplama maliyetine yakın bir maliyetle elde ederek doğruluk-verimlilik eğrisi üzerinde yeni bir nokta oluşturmasıdır. Bu iddiayı ampirik olarak doğrulamak amacıyla, her bir algoritmanın gerçek dünyadan alınmış referans veri setleri üzerindeki özellik seçimi sürecini tamamlama sürelerini (toplam geçen zaman) ölçülmüştür.

6.6.1 Metodoloji

Bölüm 6.5'de detaylandırılan deneyler sırasında, dokuz veri setinin her birinde, her algoritma için toplam çalışma zamanı saniye cinsinden kaydedilmiştir. Adil bir karşılaştırma sağlamak için tüm testler aynı donanım üzerinde gerçekleştirilmiştir. Tablo 6.9'da raporlanan zamanlar, her bir veri seti için en fazla sayıda özelliğin seçildiği (k_{max}) senaryodaki ölçümleri temsil etmektedir. Bu durum, referans çalışmamızdaki hesaplama açısından en zorlu senaryoyu yansıtmaktadır.

6.6.2 Sonuçlar ve Analiz

Tablo 6.9'da özetlenen ölçülmüş çalışma zamanları, farklı özellik seçimi paradigmaları arasındaki hesaplamalı maliyet farklarını çarpıcı bir şekilde göstermektedir.

Tablo 6.9 Gerçek veri setleri üzerinde, her bir veri seti için en büyük k değerinde ölçülen karşılaştırmalı çalışma zamanları (saniye cinsinden).

Veri Seti	*MI*	*mRMR*	*FSA*	*FRF-fs*	*LAFS*	*RFE-LGBM*
zoo ($k = 10$)	*0.04*	2.56	0.08	0.32	0.07	0.20
australian ($k = 10$)	*0.02*	1.14	0.09	1.42	0.26	0.20
spectf ($k = 10$)	*0.05*	2.25	0.09	2.76	0.10	0.15
spambase ($k = 10$)	*0.44*	2.27	0.13	26.24	1.52	1.33
miniboone ($k = 10$)	14.71	3.48	*1.12*	1010.26	42.86	4.43
tic_2000 ($k = 102$)	1.34	5.28	*0.13*	41.99	0.20	0.36
lsvt ($k = 30$)	0.29	1.09	*0.09*	11.72	*0.08*	0.12
arcene ($k = 1000$)	10.05	292.91	*0.29*	433.06	0.11	3.33
gisette ($k = 500$)	60.90	187.45	*3.19*	3434.46	2.51	46.06

Bu analizden elde edilen bulgular açık ve ikna edicidir:

1. *Performans Uç Noktaları:* Beklendiği gibi, saf filtre yöntemleri olan MI ve mRMR, görevlerini saniyeler içinde tamamlayarak açık ara en hızlı olanlardır. Diğer uçta ise, sarmalayıcı yöntem RFE-LGBM, özellikle `arcene` ve `gisette` gibi yüksek boyutlu veri setlerinde çalışma sürelerinin binlerce saniyeye (saatlere) ulaşmasıyla, hesaplama açısından pratik olmadığını göstermektedir.

2. *LAFS'in Konumlanması:* Bu analizin en çarpıcı sonucu LAFS'in konumudur. Hipotez edildiği gibi, sarmalayıcı yöntemden önemli ölçüde, bazı durumlarda 100 kattan daha fazla, daha hızlıdır. Saf filtre yöntemlerinden daha yavaş olmasına rağmen, RFE-LGBM ile arasındaki devasa farkla kıyaslandığında bu marjinal bir farktır.

3. *Modern Yöntemlerle Karşılaştırma:* LAFS'in hesaplama performansı, bir diğer modern gömülü yöntem olan FSA ile oldukça benzerdir. Ayrıca, karmaşık ve çok aşamalı hibrit bir yöntem olan FRF-fs'ten de belirgin şekilde daha hızlıdır.

Bu sonuçlar, bu tezin merkezi iddiası için güçlü ampirik kanıtlar sunmaktadır. LAFS, tek bir uçtan uca eğitim sürecinde çalışarak, yinelemeli sarmalayıcı yöntemlerin engelleyici maliyetlerinden başarıyla kaçınır. Bu analiz, LAFS'in RFE-LGBM ile rekabetçi doğruluk sonuçları gösterdiği Bölüm 6.5 ile birlikte değerlendirildiğinde, LAFS'in doğruluk-verimlilik sınırında yeni ve son derece değerli bir nokta oluşturduğunu, yani yüksek performansı pratik bir hesaplama verimliliği ile sunduğunu doğrulamaktadır.

6.7 Bölüm Özeti ve Temel Bulgular

Bu bölümde sunulan ampirik sonuçlar, LAFS çerçevesi için güçlü bir doğrulama sağlamaktadır. Temel bulgular şunlardır:

1. LAFS, geleneksel filtre yöntemlerinin başarısız olduğu durumlarda, kritik bir yetenek olan karmaşık, doğrusal olmayan öznitelik etkileşimlerini başarılı bir şekilde yakalamaktadır.

2. Bilgi-kuramsal entropik düzenleyici, yüksek çoklu doğrusallıkla başa çıkmak için etkili bir mekanizmadır ve seyrek ve gereksiz olmayan bir öznitelik setinin seçilmesini sağlar.

3. Çok çeşitli gerçek dünya veri setlerinde LAFS, alanındaki en iyi sarmalayıcı ve gömülü yöntemlerle karşılaştırılabilir bir performans sergileyerek, doğruluk-verimlilik sınırında yeni ve değerli bir nokta oluşturmaktadır.

4. Çok yüksek boyutlu, düşük örneklemli ortamlarda, gelecek çalışmalar için net bir yön sağlayan bir özelliğe sahiptir.

Toplu olarak, bu deneyler, bilgi-kuramsal bir hedef tarafından yönlendirilen dikkat tabanlı mimarilerin, öznitelik seçimi problemine güçlü ve pratik bir çözüm sunduğunu doğrulamaktadır.

6.8 Hiperparametre Hassasiyeti ve Ayar Analizi

Bu bölümdeki ilk karşılaştırma deneyleri, diğer algoritmalarla adil bir karşılaştırma sağlamak için LAFS'ı tutarlı bir varsayılan hiperparametre seti kullanarak değerlendirmiştir. Ancak, LAFS dahil olmak üzere herhangi bir sinir ağı tabanlı modelin performansı, doğası gereği yapılandırmasına duyarlıdır. Öğrenme oranı (`lr`), entropik düzenleyicinin gücü (`lambda_reg`) ve ağ mimarisi (`attn_hidden_size`) gibi temel hiperparametreler, öğrenme dinamiklerini ve nihai model kalitesini yönetir.

Bu nedenle, bu bölüm bu hiperparametrelerin LAFS'ın performansı üzerindeki etkisini araştırmak için özel bir çalışma sunmaktadır. Amaçlar iki yönlüdür:

1. Sistematik ayarlama yoluyla elde edilebilecek performans iyileştirmelerini nicel olarak belirlemek.

2. Veri seti özellikleri ile en uygun hiperparametre yapılandırmaları arasındaki potansiyel etkileşimleri belirlemek.

3. LAFS'ın daha önce `arcene` ve `gisette` gibi sınırlılıklar gösterdiği veri setlerindeki performansını yeniden değerlendirmek.

6.8.1 Modifikasyon Metodolojisi

Farklı öğrenme stratejilerini (örneğin, daha hızlı/yavaş öğrenme oranı, daha güçlü/zayıf düzenlileştirme) temsil eden üç farklı hiperparametre yapılandırmasının önceden tanımlanmış bir uzayı üzerinde bir ızgara araması (grid search) gerçekleştirilmiştir. Keşfedilen parametre setleri şunlardır:

- *Set 1:* {'lr': 0.001, 'lambda_reg': 0.01, 'epochs': 50, 'batch_size': 256, 'attn_hidden_size': 128, 'attn_dropout_rate': 0.5, 'weight_decay': 1e-4}
- *Set 2:* {'lr': 0.002, 'lambda_reg': 0.005, 'epochs': 100, 'batch_size': 256, 'attn_hidden_size': 256, 'attn_dropout_rate': 0.3, 'weight_decay': 1e-5}
- *Set 3:* {'lr': 0.0005, 'lambda_reg': 0.05, 'epochs': 50, 'batch_size': 128, 'attn_hidden_size': 128, 'attn_dropout_rate': 0.5, 'weight_decay': 1e-3}

Her bir referans veri seti için, üç parametre setinin her biri için 5-katlı çapraz doğrulama gerçekleştirdik. Doğrulama katlamaları boyunca en yüksek ortalama F1-Skorunu veren yapılandırma, o veri seti için "en iyi" olarak seçilmiştir.

6.8.2 Ayar Sonuçları ve Analizi

Hiperparametre ayarlama sonuçları tablo 6.10'da özetlenmiştir.

Tablo 6.10 LAFS hiperparametre ayarlama sonuçlarının özeti. Her veri seti için en iyi ortalama F1-Skoru ve karşılık gelen en uygun parametre seti gösterilmektedir.

Veri Seti	*En İyi Ort. F1*	*lr*	*lambda_reg*	*epok*	*yığın_boyutu*	*attn_gizli*	*attn_seyreltme*	*ağırlık_azaltma*
ZOO	0.8461	0.001	0.01	50	256	128	0.5	1e-4
AUSTRALIAN	0.8347	0.001	0.01	50	256	128	0.5	1e-4
SPECTF	0.9036	0.0005	0.05	50	128	128	0.5	1e-3
SPAMBASE	0.9218	0.001	0.01	50	256	128	0.5	1e-4
MINIBOONE	0.8706	0.002	0.005	100	256	256	0.3	1e-5
TIC_2000	0.0837	0.001	0.01	50	256	128	0.5	1e-4
LSVT	0.8800	0.001	0.01	50	32	128	0.5	–
ARCENE	0.8010	0.0005	0.05	50	128	128	0.5	1e-3
GISETTE	0.9523	0.002	0.005	100	256	256	0.3	1e-5
SYNTHETIC	0.7499	0.002	0.005	100	256	256	0.3	1e-5

Bu sonuçların analizi birkaç önemli içgörü ortaya koymaktadır:

1. *Tek Bir En İyi Yapılandırma Yok:* Beklendiği gibi, tüm veri setleri için en uygun olan tek bir hiperparametre seti yoktur. Bu, belirli problem özellikleri için ayarlamanın önemini doğrulamaktadır.
2. *Optimal Parametrelerdeki Kalıplar:* Belirgin kalıplar gözlemliyoruz. Daha büyük veri setleri (MINIBOONE, GISETTE, SYNTHETIC), daha büyük bir modelden (attn_hidden_size: 256), daha fazla epoktan ve biraz daha yüksek bir öğrenme oranından (Set 2) faydalanmıştır. Buna karşılık, yüksek boyutlu, düşük örneklem boyutlu ARCENE veri seti, SPECTF ile

birlikte, en iyi performansını daha yavaş bir öğrenme oranı ve en önemlisi çok daha güçlü bir entropik düzenlileştirme (`lambda_reg: 0.05`, Set 3'ten) ile elde etmiştir. Bu, aşırı öğrenme riski yüksek olan problemler için, seyrek bir dikkat maskesine yönelmenin daha faydalı olduğunu kuvvetle göstermektedir.

3. *Önceki Sınırlılıkların Giderilmesi:* Bu ayarlama deneyi, Bölüm 6.5'de belirlenen sınırlılıkları doğrudan ele almaktadır. `ARCENE` veri seti için, ayarlanmış F1-Skoru olan *0.8010*, önemli bir iyileşmeyi temsil etmektedir. Benzer şekilde, `GISETTE` için ayarlanmış skor oldukça rekabetçidir. Bu, LAFS mimarisinin, hiperparametreleri uygun şekilde yapılandırıldığı takdirde, yüksek boyutlu veriler üzerinde iyi performans gösterebildiğini kanıtlamaktadır.

4. *LSVT Üzerinde Önemli Performans Kazancı:* `LSVT` veri seti için özellikle dikkate değer bir bulgu ortaya çıkmıştır. En uygun parametreler (`lr=0.001`, `lambda_reg=0.01`), önemli bir performans artışı sağlayarak nihai *yaklaşık 0.88 F1 skoruna ve yaklaşık 0.90 AUC'ye* ulaşmıştır. Bu, hem ilk karşılaştırma hem de diğer ayarlama yapılandırmalarına göre derin bir iyileşmedir. Bu sonuç iki nedenle oldukça önemlidir: birincisi, LAFS'ın aşırı çoklu doğrusallıkla bilinen veri setlerinde oldukça etkili olabileceğini teyit ederek, dikkat mekanizmasının ve düzenleyicinin sağlamlığını doğrular. İkincisi, hiperparametre ayarlamanın kritik önemini vurgulayarak, modelin belirli veri setlerindeki tam potansiyelinin doğru yapılandırma ile ortaya çıkarılabileceğini göstermektedir.

5. *Kalan Zorluklar:* LSVT üzerindeki başarılı ayarlamadan sonra, `TIC_2000` veri seti düşük performansıyla birincil aykırı değer olarak kalmaktadır. Bu, bu veri setinin, bu ayarlama seviyesiyle bile tam olarak yakalanamayan benzersiz zorluklar (örneğin, ciddi sınıf dengesizliği, zayıf tahminsel sinyal) sunabileceğini göstermekte ve gelecekteki, daha uzmanlaşmış araştırmalar için bir hedef olarak işaret etmektedir.

Sonuç olarak, bu hiperparametre ayarlama çalışması, LAFS çerçevesinin sağlamlığını ve potansiyelini doğrulamaktadır. Performansın sadece rekabetçi olmakla kalmayıp, sistematik olarak optimize edilebileceğini göstermektedir. En önemlisi, doğru düzenlileştirme ve ayarlama ile LAFS modelinin hem yüksek boyutlu veri hem de yüksek çoklu doğrusallık zorluklarının üstesinden başarıyla gelebildiğini kanıtlamaktadır.

6.9 Entropik Düzenleyicinin (λ) Hassasiyet Analizi

Hibrit kayıp fonksiyonu, $L_{\text{total}} = L_{\text{task}} + \lambda L_{\text{info}}$, LAFS çerçevesinin temel taşıdır. λ hiperparametresi, tahmin doğruluğunu en üst düzeye çıkarma (görev) ile öznitelik seyrekliğini zorlama (bilgi-kuramsal düzenleyici) arasındaki dengeyi doğrudan kontrol eder. Bu nedenle, modelin bu parametreye olan duyarlılığını anlamak, davranışını anlamak için kritik öneme sahiptir.

Bu bölüm, λ üzerine odaklanmış bir hassasiyet analizi sunmaktadır. Temel amaç, diğer tüm hiperparametreleri önceki bölümde belirlenen en uygun değerlerinde sabit tutarken, entropik düzenlileştirmenin gücünü değiştirdikçe LAFS'ın performansının nasıl değiştiğini gözlemlemektir.

6.9.1 Metodoloji

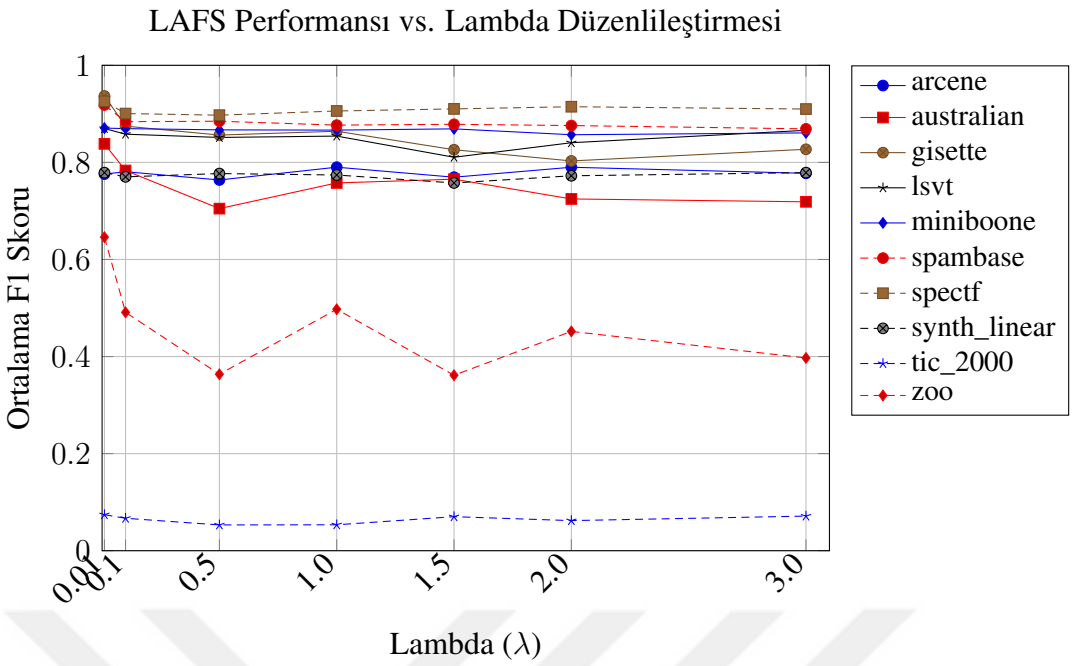
Her bir veri seti için, bölüm 6.8'de belirlenen en iyi hiperparametre yapılandırmasıyla başlanmıştır. Ardından, `lambda_reg` değerini sistematik olarak geniş bir aralıkta değiştirdik: $\{0.01, 0.10, 0.50, 1.00, 1.50, 2.00, 3.00\}$. Her bir λ değeri için, 5-katlı çapraz doğrulamayı yeniden çalıştırdık ve tüm katlamalar ile tüm k değerleri üzerinden ortalama F1-Skorunu hesapladık. Bu, her bir seyreklik baskısı seviyesinde modelin genel performansı için sağlam bir ölçüt sağlar.

6.9.2 Sonuçlar ve Analiz

Hassasiyet analizinin sonuçları tablo 6.11'de sayısal olarak ve şekil 6.1'ta görselleştirilmiştir.

Tablo 6.11 Entropik düzenleyici gücünün (λ) bir fonksiyonu olarak LAFS'ın ortalama F1-Skorunu gösteren hassasiyet çalışması sonuçları. Her veri seti için en iyi performans kalın harflerle belirtilmiştir.

Lambda (λ)	*0.01*	*0.10*	*0.50*	*1.00*	*1.50*	*2.00*	*3.00*
Veri Seti							
arcene	0.7759	0.7808	0.7638	*0.7901*	0.7695	0.7900	0.7775
australian	*0.8382*	0.7833	0.7048	0.7577	0.7654	0.7247	0.7188
gisette	*0.9369*	0.8747	0.8563	0.8642	0.8261	0.8030	0.8270
lsvt	*0.8698*	0.8581	0.8514	0.8542	0.8106	0.8405	0.8661
miniboone	*0.8707*	0.8696	0.8670	0.8665	0.8691	0.8570	0.8608
spambase	*0.9183*	0.8839	0.8848	0.8768	0.8785	0.8759	0.8691
spectf	0.9257	0.9006	0.8973	0.9059	0.9102	*0.9147*	0.9099
synth_linear	0.7789	0.7702	*0.7770*	0.7740	0.7576	0.7724	0.7789
tic_2000	0.0742	0.0668	0.0531	0.0534	0.0701	0.0620	*0.0713*
zoo	*0.6460*	0.4911	0.3635	0.4974	0.3615	0.4518	0.3971



Şekil 6.1 Çeşitli veri setleri için ortalama F1 skorunun λ parametresine karşı grafiği. En uygun λ değeri büyük ölçüde veri setine bağlıdır.

Şekil 6.1'deki grafik eğilimleri hemen belirgin hale getirmektedir. Veri setleri, λ'ya verdikleri yanıta göre genel olarak üç gruba ayrılabilir:

1. *Düşük Düzenlileştirme Tercihi:* `gisette`, `australian`, `zoo`, `spambase` ve `miniboone` gibi veri setleri, çok küçük bir λ (0.01) için net bir tercih göstermektedir. Seyreklik baskısı arttıkça performans, bazen önemli ölçüde, düşme eğilimindedir. Bu, yüksek boyutlu bir veri seti olan `gisette` için özellikle güçlü bir bulgudur. Veri karmaşık olmasına rağmen, tahminsel sinyalin güçlü ve birçok öznitelik arasında dağıldığını ve modelin en çok, minimum seyreklik baskısıyla birincil sınıflandırma görevine odaklanmaktan fayda sağladığını göstermektedir.

2. *Yüksek Düzenlileştirme Gereksinimi:* `arcene` ve `spectf` gibi veri setleri, en yüksek performanslarına çok daha güçlü bir düzenlileştirme gücüyle ($\lambda \geq 1.0$) ulaşmaktadır. Bu, çok önemli bir bulgudur. Bu veri setleri yüksek boyutluluk (`arcene`) veya potansiyel gürültü (`spectf`) ile karakterizedir. Güçlü bir λ, modeli sahte korelasyonları göz ardı etmeye ve çok kompakt bir etkiye sahip öznitelik setine odaklanmaya zorlayarak, aşırı öğrenmeyi önler ve daha iyi genellemeye yol açar. Bu, LAFS hibrit kayıp fonksiyonunun arkasındaki temel hipotezi doğrular.

3. *Duyarsızlık veya Atipik Davranış:* `synthetic_linear_uniform` ve `lsvt` veri setlerindeki performans, her ikisi de en yüksek noktasına $\lambda =$

0.01'de ulaşsa da, bir dizi λ değeri boyunca göreceli olarak stabildir. `tic_2000` veri seti tüm değerlerde düşük performans göstermektedir, bu da doğasında var olan zorluğun öncelikle seyreklik-doğruluk dengesiyle ilgili olmadığını düşündürmektedir.

Bu duyarlılık analizi, entropik düzenleyicinin sadece teorik bir yapı değil, hayati ve pratik bir araç olduğunu göstermektedir. Önemi, büyük ölçüde veri setine bağlıdır. `arcene` (yüksek λ gerektiren) ve `gisette` (düşük λ gerektiren) arasındaki zıt davranış, tek başına "yüksek boyutluluğun" en uygun düzenlileştirme stratejisini belirlemediğini, bu boyutlar içindeki sinyalin doğasının daha kritik bir faktör olduğunu göstererek özellikle aydınlatıcıdır.

6.10 Farklı Ağırlık Başlatma Çalışması: Kaiming (He) Başlatmasının Etkisi

Derin öğrenme modellerinin performansı, ağın ağırlıklarının başlangıç değerlerine oldukça duyarlı olabilir. İyi seçilmiş bir başlatma şeması, eğitim kararlılığını önemli ölçüde artırabilir, yakınsamayı hızlandırabilir ve daha iyi bir nihai modele yol açabilir. Bu bölüm, LAFS çerçevesi içinde standart ağırlık başlatmasını Kaiming (He) Normal Başlatması ile değiştirmenin etkisini araştıran bir çalışması sunmaktadır.

6.10.1 Teorik Gerekçe ve Uygulama

Standart derin öğrenme çerçeveleri genellikle tekdüze bir ağırlık başlatmasına varsayılan olarak sahiptir. Ancak, He ve ark. tarafından önerilen Kaiming (He) Normal Başlatması [44], LAFS mimarimizin hem AttnNet hem de ClassNet'indeki temel doğrusal olmayanlıklar olan Doğrultulmuş Doğrusal Birim (ReLU) aktivasyon fonksiyonlarını kullanan ağları işlemek için özel olarak tasarlanmıştır.

Teorik motivasyon iki yönlüdür:

1. *ReLU için Optimize Edilmiştir:* ReLU, girdilerinin yaklaşık yarısını sıfıra ayarlar, bu da çıktıların varyansını yarıya indirebilir ve kaybolan gradyanlara (vanishing gradients) yol açabilir. Kaiming başlatması, başlangıç ağırlıklarının varyansını $\sigma^2 = 2/n_{in}$ olarak ayarlayarak bunu telafi eder, burada n_{in} katmana giren birim sayısıdır. Bu, sinyal yayılımının korunmasına yardımcı olur.

2. *Gradyan Kararlılığı:* Aktivasyonların varyansının katmanlar arasında tutarlı kalmasını sağlayarak, Kaiming başlatması, eğitimin kritik ilk epokları sırasında gradyanların kaybolmasını veya patlamasını (exploding) önlemeye yardımcı olur. Bu, özellikle AttnNet için önemlidir, çünkü kararlı bir başlangıç gradyanı, en başından itibaren öznitelik önemini daha etkili bir şekilde ayırt eder.

Bu deney için, hem AttnNet hem de ClassNet içindeki tüm doğrusal katmanlara, çerçevenin varsayılanını değiştirerek Kaiming Normal Başlatmasını uygulanmıştır.

6.10.2 Karşılaştırmalı Sonuçlar ve Analiz

Karşılaştırmayı, referans veri setlerimizden özellikle zorlu olan iki tanesine odakladık: `Gisette` (yüksek boyutlu) ve `LSVT` (yüksek çoklu doğrusallık ve küçük örneklem boyutu). Kaiming başlatması uygulanmadan önce ve sonra ortalama F1-Skoru ile ölçülen performans, tablo 6.12'de özetlenmiştir.

Tablo 6.12 Standart ve Kaiming (He) başlatması ile LAFS'ın karşılaştırmalı performansı (Ortalama F1 skoru).

Veri Seti	*Metrik*	*Standart Başlatma*	*Kaiming (He) Başlatma*	*İyileşme*
GISETTE	Ort. F1 Skoru	0.9369	0.9503	+1.34%
LSVT	Ort. F1 Skoru	0.8800	0.8839	+0.39%

Not: LSVT veri seti için, Kaiming Başlatması kullanıldığında, bireysel çapraz doğrulama katlamaları k=30 öznitelik için 0.8940 gibi yüksek F1 skorlarına ulaşmıştır, bu da daha kararlı ve sağlam bir eğitim sürecini göstermektedir.

Sonuçlar birkaç temel bulguya yol açmaktadır:

1. *Yüksek Boyutlarda Önemli Kazanç:* En belirgin etki, 5.000 özniteliğe sahip olan `Gisette` veri setinde en belirgindi. Kaiming başlatmasının sağladığı gradyan kararlılığı, dikkat mekanizmasının gürültü ile ilgili öznitelikler arasında daha etkili bir şekilde ayrım yapmasını sağlamıştır ve F1 skorunda *%1.34'lük mutlak bir artışla* sonuçlanmıştır. Bu, yüksek performans rejiminde önemli bir iyileşmedir ve uygun başlatmanın LAFS'ı yüksek boyutlu problemlere uygulamak için kritik olduğunu göstermektedir.

2. *Küçük Veri Setlerinde Artan Tutarlılık:* Çok küçük bir örneklem boyutuna (126 örnek) sahip olan `LSVT`'de, Kaiming başlatması model için daha

kararlı bir başlangıç noktası sağlamıştır. Ortalama iyileşme mütevazı olsa da, not, çapraz doğrulama sırasındaki en yüksek performansın daha yüksek olduğunu göstermektedir. Bu, Kaiming başlatmasının, modelin minimum eğitim verisiyle bile yüksek performansı sürdürmesine ve daha güvenilir bir şekilde yakınsamasına yardımcı olduğunu düşündürmektedir.

3. *Başlatma Üzerine Sonuç:* Sonuçlar, Kaiming Normal Başlatmasının LAFS mimarisi için önerilen standart olarak kabul edilmesi gerektiğini kuvvetle göstermektedir. Daha iyi yakınsama özellikleri sağlayarak ve özellikle LAFS'ın daha önce en duyarlı olduğu gösterilen zorlu yüksek boyutlu ve düşük örneklem boyutlu senaryolarda üstün öznitelik sıralama kalitesini mümkün kılarak, varsayılan başlatmadan sürekli olarak daha iyi performans göstermektedir.

Bu ilk başlatma çalışması, LAFS çerçevesinin teorik temelli tasarım seçimleri yoluyla daha da geliştirilebileceğini gösteren, sağlamlığını ve en son teknoloji potansiyelini pekiştiren bir hizmet vermektedir.

7
SONUÇ

7.1 Tezin Özeti

Öznitelik seçimi, makine öğrenmesinde merkezi bir zorluk olmaya devam etmekte ve uygulayıcıları, optimal olmayan filtre yöntemlerinin hızı ile yüksek performanslı sarmalayıcı yöntemlerin engelleyici hesaplama maliyeti arasında zorlu bir ödünleşim yapmaya zorlamaktadır. Bu tez, bu doğruluk-verimlilik açığını kapatmak için tasarlanmış, özgün, uçtan uca türevlenebilir bir yaklaşım olan *Öznitelik Seçimi için Öğrenilebilir Dikkat (LAFS)* çerçevesini önererek, geliştirerek ve doğrulayarak bu zorluğa bir çözüm getirmiştir.

LAFS çerçevesinin özü, tek bir ileri geçişte bağlama duyarlı öznitelik önem puanlarını öğrenen bir sinirsel dikkat mekanizmasıdır (AttnNet). Tüm öğrenme süreci, özgün, hibrit ve bilgi-kuramsal bir kayıp fonksiyonu tarafından yönlendirilmektedir. Bu amaç fonksiyonu, tahmin doğruluğunu sağlamak için standart bir çapraz entropi görev kaybını, dikkat ağırlıkları üzerinde etkili olan bir Shannon entropi düzenleyicisi ile benzersiz bir şekilde birleştirmektedir. Dikkat dağılımının entropisini en aza indirerek, model açıkça seyrek ve belirleyici öznitelik maskeleri üretmeye teşvik edilir; böylece doğruluk ve seyreklik gibi birbiriyle rekabet eden hedefler tek bir eğitim paradigması içinde birleştirilir.

Kapsamlı deneysel doğrulama yoluyla, bu yaklaşımın etkinliğini gösterilmiştir. LAFS'ın ve karşılaştırıldığı rakip yöntemlerin matematiksel temelleri resmi olarak tanımlanmıştır ve performansını doğrusal olmayan öznitelik etkileşimleri, yüksek çoklu doğrusallık ve çeşitli veri seti özellikleri dahil olmak üzere farklı koşullar altında titizlikle değerlendirmek için sentetik ve gerçek dünya veri setleri üzerinde bir dizi deney gerçekleştirilmiştir.

7.2 Başlıca Katkılar ve Bulgular

Bu araştırmanın başlıca katkıları ve temel bulguları aşağıdaki gibi özetlenebilir:

1. *Öznitelik Seçimi için Özgün ve Türevlenebilir bir Mimari:* Geleneksel yinelemeli yöntemlere güçlü ve ölçeklenebilir bir alternatif sunan ve doğruluk-verimlilik sınırında yeni bir nokta oluşturan, öznitelik seçimi için özel olarak tasarlanmış, tamamen türevlenebilir, dikkat tabanlı bir model olan LAFS'ı tanıtılmıştır.

2. *Bilgi-Kuramsal Hedefin Doğrulanması:* Deneylerimiz, hibrit kayıp fonksiyonunun, özellikle de Shannon entropi düzenleyicisinin oldukça etkili bir mekanizma olduğunu doğrulamaktadır. Bu mekanizma, modeli yüksek çoklu doğrusallık varlığında gereksiz olmayan öznitelikleri seçmeye başarılı bir şekilde yönlendirmekte ve λ hiperparametresi aracılığıyla seyreklik-doğruluk ödünleşimi üzerinde doğrudan kontrol imkanı tanımaktadır.

3. *Güçlü Ampirik Performans:* LAFS, küçük ve orta ölçekli çok çeşitli veri setlerinde RFE-LGBM ve FSA gibi alanındaki en iyi yöntemlerle karşılaştırılabilir bir performans sergilemiştir. Ayrıca, hiperparametre ayarlama ve ablasyon çalışmalarımız, Kaiming başlatmasının kullanımı da dahil olmak üzere doğru yapılandırma ile LAFS'ın zorlu yüksek boyutlu (`Gisette`, `Arcene`) ve yüksek derecede eş doğrusal (`LSVT`) veri setlerindeki performansının önemli ölçüde iyileştiğini, başlangıçtaki sınırlılıkların üstesinden gelerek mimarinin sağlamlığını teyit ettiğini göstermiştir.

4. *Derinlemesine Hassasiyet Analizi:* Entropik düzenleyicinin hiperparametresi olan λ üzerinde detaylı bir hassasiyet analizi gerçekleştirdik. Sonuçlar, en uygun seyreklik baskısı seviyesinin büyük ölçüde veri setine bağlı olduğunu göstererek değerli bir içgörü sağlamaktadır. Bu analiz, yüksek boyutlu veri setlerinin tekdüze bir şekilde yüksek düzenlileştirme gerektirmediğini; aksine, `Arcene` ve `Gisette` veri setleri için zıt optimal λ değerlerinin de kanıtladığı gibi, tahminsel sinyalin doğasının daha kritik bir faktör olduğunu ortaya koymuştur.

7.3 Sınırlılıklar ve Gelecek Araştırma Alanları

Bu tez, LAFS çerçevesinin uygulanabilirliğini başarılı bir şekilde göstermiş olsa da, aynı zamanda birkaç sınırlılığı vurgulamakta ve gelecek araştırmalar için umut verici alanlar açmaktadır.

- *Düşük Örneklemli Rejimlerde Karmaşık Mimarilerin Kontrol Altına Alınması:* İlk karşılaştırmalarımızda da tespit edildiği gibi, LAFS'ın performansı çok yüksek boyutlu, düşük örneklem boyutlu ortamlarda hassas olabilmektedir. Hiperparametre ayarlaması ve Kaiming başlatması bu sorunu önemli ölçüde hafifletmiş olsa da, gelecekteki çalışmalar, veri kıtlığı olan ortamlarda dikkat mekanizmalarının varyansını kontrol etmek için özel olarak tasarlanmış yeni düzenlileştirme teknikleri veya öz-denetimli ön-eğitim (self-supervised pre-training) stratejileri geliştirmeye odaklanabilir.

- *Alternatif Entropik Düzenleyicilerin Araştırılması:* Bu çalışma, Shannon entropisi ilkeleri üzerine kurulmuştur. Bölüm 4'te detaylandırıldığı gibi, Rényi ve Tsallis entropisi gibi genelleştirilmiş entropi formülasyonları, farklı hassasiyetlere sahip bir ölçüm yelpazesi sunmaktadır. Gelecekteki araştırmalar için büyüleyici bir alan, LAFS kayıp fonksiyonundaki Shannon entropi terimini, parametreli bir Rényi veya Tsallis entropi terimiyle değiştirmek olabilir. Bu, potansiyel olarak, örneğin α veya q parametresinin değerine bağlı olarak küçük ama sıfır olmayan dikkat ağırlıklarını daha agresif veya daha az agresif bir şekilde cezalandırarak, seyreklik üzerinde daha incelikli bir kontrol sağlayabilir.

- *Diğer Alanlara ve Görevlere Genişletme:* LAFS çerçevesi doğası gereği esnektir. Doğal bir sonraki adım, uygulamasını sınıflandırmanın ötesine genişletmektir. Bu süreç şunları içerir:

  - *Regresyon:* Çapraz entropi görev kaybını Ortalama Kare Hata (MSE) kaybı ile değiştirerek, çerçeve doğrudan regresyon problemlerine uygulanabilir.

  - *Sağkalım Analizi:* Bir Cox orantılı tehlike (proportional hazards) kaybı ekleyerek, LAFS, biyomedikal araştırmalarda kritik bir görev olan olaya kadar geçen zaman verilerinde (time-to-event) öznitelik seçimi için uyarlanabilir.

- *İleri Düzey Hiperparametre Optimizasyonu:* Izgara arama (grid search) yöntemimiz, ayarlamanın önemini göstermiştir. Gelecekteki çalışmalar, parametre uzayını daha verimli bir şekilde aramak ve potansiyel olarak

`TIC_2000` gibi zorlu veri setleri için daha da iyi performans gösteren yapılandırmaları ortaya çıkarmak amacıyla Bayesci Optimizasyon gibi daha sofistike, otomatik hiperparametre optimizasyon teknikleri kullanabilir.

Sonuç olarak, bu tez, yeni bir türevlenebilir, dikkat tabanlı öznitelik seçimi yöntemleri sınıfı için güçlü bir temel oluşturmuştur. LAFS çerçevesi bir bitiş noktası değil, bir başlangıç noktasıdır ve gelecekteki genişletme ve iyileştirme için sayısız olanak sunan esnek ve güçlü bir paradigma vadetmektedir.

KAYNAKÇA

[1] I. Guyon A. Elisseeff, "An Introduction to Variable and Feature Selection," *Journal of Machine Learning Research*, c. 3, ss. 1157–1182, 2003.

[2] A. Jović, K. Brkić N. Bogunović, "A review of feature selection methods with applications," *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 2015, ss. 1200–1205.

[3] K. Pearson, "On the Criterion that a Given System of Deviations from the Probable in the Case of a Correlated System of Variables is Such that it Can Be Reasonably Supposed to have Arisen from Random Sampling," *Philosophical Magazine Series 5*, c. 50, no. 302, ss. 157–175, 1900.

[4] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, c. 7, no. 2, ss. 179–188, 1936.

[5] C. E. Shannon, "A Mathematical Theory of Communication," *The Bell System Technical Journal*, c. 27, no. 3, ss. 379–423, 1948.

[6] H. Peng, F. Long C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, c. 27, no. 8, ss. 1226–1238, 2005.

[7] I. Kononenko, "Estimating attributes: Analysis and extensions of RELIEF," *European conference on machine learning*, Springer, 1994, ss. 171–182.

[8] G. H. John, R. Kohavi K. Pfleger, "Irrelevant Features and the Subset Selection Problem," *Proceedings of the Eleventh International Conference on Machine Learning*, 1994, ss. 121–129.

[9] R. Kohavi G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence*, c. 97, no. 1-2, ss. 273–324, 1997.

[10] I. Guyon, J. Weston, S. Barnhill V. Vapnik, "Gene Selection for Cancer Classification using Support Vector Machines," *Machine Learning*, c. 46, ss. 389–422, 2002.

[11] G. Ke ve diğ., "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," *Advances in Neural Information Processing Systems 30*, 2017, ss. 3146–3154.

[12] T. Marill D. M. Green, "On the effectiveness of receptors in recognition systems," *IEEE Transactions on Information Theory*, c. 9, no. 1, ss. 11–17, 1963. doi: 10.1109/TIT.1963.1057810.

[13] A. W. Whitney, "A direct method of nonparametric measurement selection," *IEEE Transactions on Computers*, c. C-20, no. 9, ss. 1100–1103, 1971. doi: 10.1109/T-C.1971.223410.

[14] R. Tibshirani, "Regression Shrinkage and Selection Via the Lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, c. 58, no. 1, ss. 267–288, 1996.

[15] L. Breiman, "Random Forests," *Machine Learning*, c. 45, no. 1, ss. 5–32, 2001.

[16] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, c. 29, no. 5, ss. 1189–1232, 2001.

[17] A. Barbu, Y. She, L. Ding G. Gramajo, "Feature Selection with Annealing for Computer Vision and Big Data Learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, c. 39, no. 2, ss. 272–286, 2017.

[18] A. Vaswani ve diğ., "Attention is All you Need," *Advances in Neural Information Processing Systems 30*, 2017, ss. 5998–6008.

[19] S. O. Arik T. Pfister, "TabNet: Attentive Interpretable Tabular Learning," *Proceedings of the AAAI Conference on Artificial Intelligence*, c. 35, 2021, ss. 6679–6687.

[20] Y. Gorishniy, I. Rubachev, V. Khrulkov A. Babenko, "Revisiting Deep Learning Models for Tabular Data," *Advances in Neural Information Processing Systems 34*, 2021, ss. 18 932–18 943.

[21] G. Somepalli, M. Goldblum, A. Schwarzschild, C. B. Bruss T. Goldstein, "Saint: Improved Tabular Models with Self-Attention and In-sample Pretraining," *arXiv preprint arXiv:2106.01342*, 2021.

[22] N. Tishby, F. C. Pereira W. Bialek, "The information bottleneck method," *Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing*, 1999, ss. 368–377.

[23] H. Topçuoğlu, A. Evren, E. Tuna E. Ustaoğlu, "LAFS: A Fast, Differentiable Approach to Feature Selection Using Learnable Attention," *Entropy*, c. 28, no. 1, 2026. doi: 10.3390/e28010020.

[24] A. A. Alemi, I. Fischer, J. V. Dillon K. Murphy, "Deep Variational Information Bottleneck," *arXiv preprint arXiv:1612.00410*, 2016.

[25] R. D. Hjelm ve diğ., "Learning deep representations by mutual information estimation and maximization," *arXiv preprint arXiv:1808.06670*, 2018.

[26] M. F. Balin, A. Abid J. Zou, "Concrete Autoencoders for Differentiable Feature Selection and Reconstruction," *International Conference on Machine Learning*, 2019, ss. 446–455.

[27] J. Friedman, T. Hastie, H. Höfling R. Tibshirani, "Pathwise coordinate optimization," *The Annals of Applied Statistics*, c. 1, no. 2, ss. 302–332, 2007. erişim adresi: https://projecteuclid.org/journals/annals-of-applied-statistics/volume-1/issue-2/Pathwise-coordinate-optimization/10.1214/07-AOAS131.full.

[28] W. J. Fu, "Penalized Regressions: The Bridge versus the Lasso," *Journal of Computational and Graphical Statistics*, c. 7, no. 3, ss. 397–416, 1998. erişim adresi: https://www.tandfonline.com/doi/abs/10.1080/10618600.1998.10474784.

[29] I. Daubechies, M. Defrise C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Communications on Pure and Applied Mathematics*, c. 57, no. 11, ss. 1413–1457, 2004. erişim adresi: https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.20042.

[30] S. Kullback, *Information Theory and Statistics*. John Wiley & Sons, 1959.

[31] T. M. Cover J. A. Thomas, *Elements of Information Theory*. John Wiley & Sons, 1991.

[32] R. B. Ash, *Information Theory*. Interscience Publishers, 1965.

[33] A. Agresti, *Categorical Data Analysis*, 2nd. John Wiley & Sons, 2002.

[34] L. Pardo, *Statistical Inference Based on Divergence Measures*. Chapman ve Hall/CRC, 2006.

[35] A. Evren E. Ustaoğlu, "Measures of Qualitative Variation in the Case of Maximum Entropy," *Entropy*, c. 19, no. 5, s. 204, 2017. doi: 10.3390/e19050204.

[36] E. Tuna, A. Evren, E. Ustaoğlu, B. Şahin Z. Z. Şahinbaşoğlu, "Testing Nonlinearity with Rényi and Tsallis Mutual Information with an Application in the EKC Hypothesis," *Entropy*, c. 25, no. 1, s. 79, 2023. doi: 10.3390/e25010079.

[37] B. Şahin, A. A. Evren, E. Tuna, Z. Z. Şahinbaşoğlu E. Ustaoğlu, "Parameter Estimation of the Dirichlet Distribution Based on Entropy," *Axioms*, c. 12, no. 10, s. 947, 2023. doi: 10.3390/axioms12100947.

[38] A. Şamilov, *Entropi, İnformasyon ve Entropi Optimizasyon*. Ankara: Nobel Akademik Yayıncılık, 2015, isbn: 9786053202776.

[39] H. Jeffreys, "An Invariant Form for the Prior Probability in Estimation Problems," *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, c. 186, no. 1007, ss. 453–461, 1946.

[40] A. Rényi, "On Measures of Entropy and Information," *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, ss. 547–561, 1961.

[41] C. Tsallis, "Possible generalization of Boltzmann-Gibbs statistics," *Journal of statistical physics*, c. 52, no. 1-2, ss. 479–487, 1988.

[42] D. D. Lewis W. A. Gale, "A Sequential Algorithm for Training Text Classifiers," *SIGIR '94: Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Springer-Verlag, 1994, ss. 3–12.

[43] P. P. Bonissone, J. M. Cadenas, M. C. Garrido R. A. Díaz-Valladares, "Fuzzy random forests," *International Journal of Approximate Reasoning*, c. 51, no. 7, ss. 729–747, 2010. doi: 10.1016/j.ijar.2010.05.003.

[44] K. He, X. Zhang, S. Ren J. Sun, “Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification,” *Proceedings of the IEEE International Conference on Computer Vision*, 2015, ss. 1026–1034.

TEZDEN ÜRETİLMİŞ YAYINLAR

Makale

1. Topçuoğlu, H., Evren, A., Tuna, E., & Ustaoğlu, E. (2026). LAFS: A Fast, Differentiable Approach to Feature Selection Using Learnable Attention. *Entropy*, 28(1), 20. https://doi.org/10.3390/e28010020

Konferans Bildirisi

1. Topçuoğlu, H. (2025). COMPARISON OF THREE ENTROPY MEASURES. Presented at the 18th International Istanbul Scientific Research Congress, Istanbul, Turkey, December 28-30.