

Metin Sınıflandırmada Öznitelik Seçim Yöntemlerinin Değerlendirilmesi

AYTUĞ ONAN

CELAL BAYAR ÜNİVERSİTESİ, BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

SERDAR KORUKOĞLU

EGE ÜNİVERSİTESİ, BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

İçerik

- ▶ Metin Sınıflandırma
- ▶ Öznitelik Seçimi
 - ▶ Filtre-tabanlı Öznitelik Seçimi
 - ▶ Sarmalama-tabanlı Öznitelik Seçimi
- ▶ Sezgisel Arama Yöntemleri
- ▶ Deneysel Süreç ve Sonuçlar

Metin Sınıflandırma

- ▶ Metin, **önemli bir bilgi kaynağı**,
- ▶ Metin sınıflandırma, **doğal dil işleme**, **veri madenciliği** ve **makine öğrenmesi** yöntemlerini kullanarak, belgelerin önceden tanımlanmış sınıflara atanması,
 - ▶ Haber Filtreleme ve Organizasyonu
 - ▶ Belge Organizasyonu ve Erişimi
 - ▶ Görüş sınıflandırma
 - ▶ İstenmeyen e-posta filtreleme

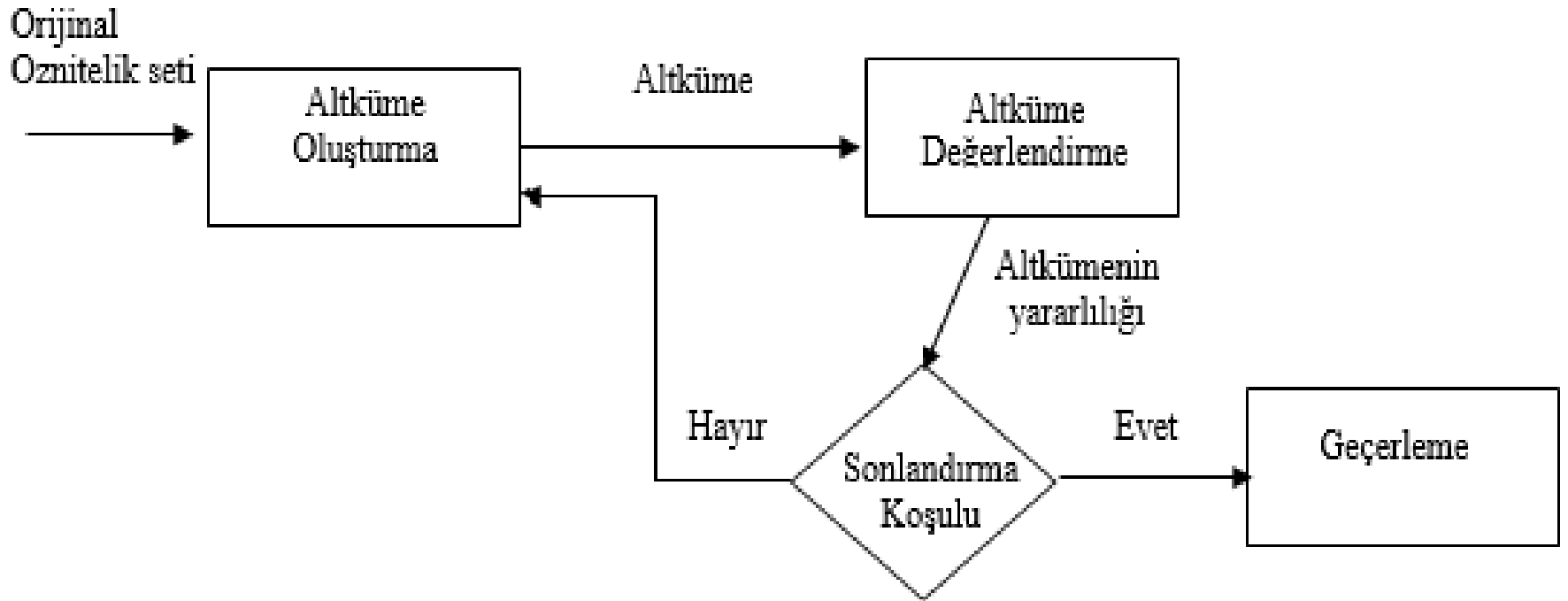
Metin Sınıflandırma

- ▶ **Önişleme**
 - ▶ Dizgeciklere ayırma, Köklere ayırma
- ▶ **Verinin Uygun Bir Veri Temsil Yöntemi ile Temsili**
 - ▶ Vektör uzay modeli → **Yüksek Boyutluluk**
- ▶ **Öznitelik Seçimi**
- ▶ **Sınıflandırma**
- ▶ **Performans Değerlendirme**

Öznitelik Seçimi

- ▶ Veri setinden uygun bir **öznitelik alt kümesi** elde edilmesi,
 - ▶ **Ölçeklenebilirlik** ve **Sınıflandırma Başarımının** iyileştirilmesi,
 - ▶ $M < N$ olmak üzere, N tane öznitelik içeren bir veri seti için **belirli bir ölçüt fonksiyonu**, M 'nin tüm alt kümelerinde en iyi olacak şekilde **M öz niteliğin belirlenmesi süreci**

Öznitelik Seçim Süreci



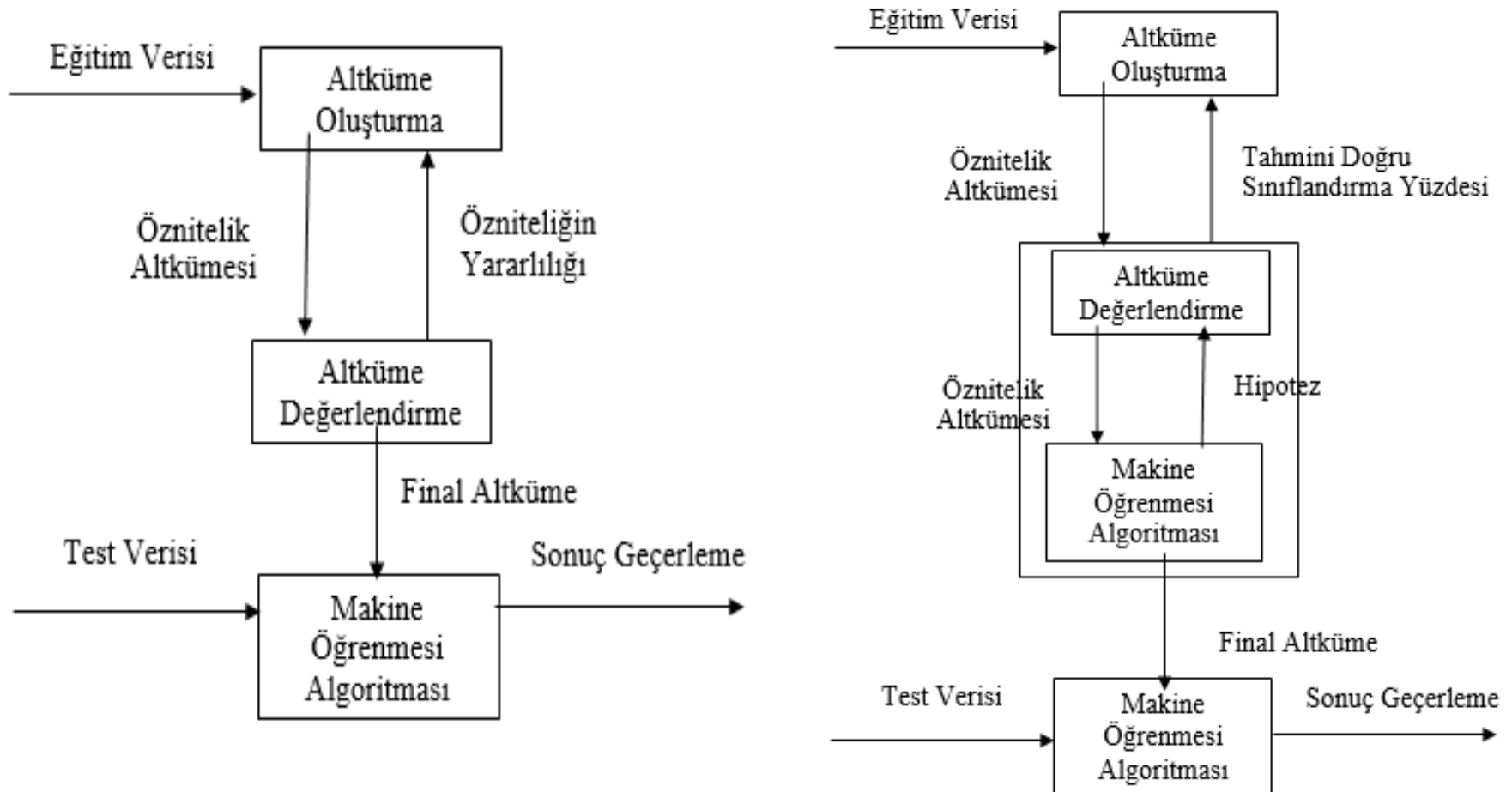
Öznitelik Seçim Süreci

- ▶ **Öznitelik Altkümesi Oluşturma:** Orijinal öznitelik setinden değerlendirmede kullanılmak üzere altkümeler oluşturan bir arama süreci.
- ▶ Arama sürecinin organizasyonu önemli:
 - ▶ N öznitelik için **2^N olası altkümenin** incelenmesini gerekli.
 - ▶ Genellikle **sezgisel arama yöntemleri** kullanılmakta:
 - ▶ *En iyi önce arama algoritması, genetik algoritmalar, açgözlü arama algoritması, parçacık sürüsü optimizasyonu, doğrusal ileri seçim algoritması, sıra arama algoritması, yeniden sıralama algoritması gibi.*

Öznitelik Seçim Süreci

- ▶ **Öznitelik Altkümesi Değerlendirme:** Öznitelik altkümesinin yararlılığının belirlenmesi.
 - ▶ **Filtre-tabanlı:** Verinin genel özelliklerine dayalı sezgiseller - Hedef ile en çok doğrusal ilişkili özniteliklerin seçilmesi
 - ▶ **Sarmalama-tabanlı:** Belirli bir öğrenme algoritmasına dayalı- Performans eniyilemesine odaklı / Daha yavaş
 - ▶ **Melez:** Özniteliklerin belirli bir ölçüte göre sıralanması (Filtre)+Sarmalama-tabanlı

Filtre/Sarmalama-Tabanlı Öznitelik Seçimi



Filtre-Tabanlı Öznitelik Seçim Süreci

- ▶ **Bireysel Öznitelik Ölçütleri (*Öznitelik Sıralama*)**
 - ▶ Özniteliklerin belirli bir ölçüte göre sıralanması ve belirli bir eşik değeri aşan özniteliklerin seçilmesi.
 - ▶ Hesaplama etkinliği bakımından başarılı.
 - ▶ Bilgi Kazancı, Ki-Kare Ölçütü, Simetrik Belirsizlik Katsayısı gibi.
- ▶ **Grup Öznitelik Ölçütleri**
 - ▶ Aday öznitelik altkümelerinin değerlendirilmesi.
 - ▶ Öznitelikler arası ilişkilerin dikkate alınması.
 - ▶ Farklı sezgisel arama algoritmaları ile birlikte kullanılabilir.
 - ▶ Korelasyon-tabanlı ya da Tutarlılık-tabanlı öznitelik seçimi gibi.

Korelasyon-tabanlı Öznitelik Seçimi (CBS)

- ▶ Öznitelik alt kümelerinin yararlılığı sezgisel bir fonksiyona dayalı olarak inceler.
- ▶ Her bir özneliğin sınıf etiketinin kestirimindeki belirleyiciliği ve öznelikler arası korelasyon dikkate alınır.

$$M_s = \frac{\overline{kr_{cf}}}{\sqrt{k + k(k-1)r_{ff}}}$$

k : Öznitelik sayısı,

M_s : S alt kümesinin sezgisel yararlılığı,

$\overline{r_{cf}}$: Ortalama öznelik-sınıf korelasyonu

r_{ff} : Ortalama öznelikler arası korelasyon

- ▶ Fonksiyon ile arama uzayındaki tüm olası kombinasyonlar için öznelik alt kümelerine ilişkin bir sıralama elde edilmektedir.
- ▶ Arama uzayındaki **tüm olası öznelik alt kümelerinin** incelenmesi oldukça **maliyetli** --- Genellikle bir arama algoritması ile birlikte kullanılmakta!

Tutarlılık Tabanlı Öznitelik Seçimi (ConsBS)

- ▶ **Öznitelik alt kümelerini** tutarlılık ölçütüne göre değerlendirir:

$$ConsBS_s = 1 - \frac{\sum_{i=0}^J |D_i| - |M_i|}{N}$$

- ▶ s , öznitelik altkümesi,
- ▶ J , s öznitelik altkümesi için öznitelik değerlerinin farklı kombinasyon sayısı,
- ▶ $|D_i|$ i öznitelik değeri kombinasyonunun görülme sayısı,
- ▶ $|M_i|$ i öznitelik değeri kombinasyonunun çoğunluk sınıfında görülme sayısı,
- ▶ N , veri setindeki toplam örnek sayısı,

Tutarlılık Tabanlı Öznitelik Seçimi (ConsBS)

- ▶ Tüm öznitelikleri içeren bir alt küme ile başlanır.
- ▶ Öznitelik alt küme uzayından rastgele olarak bir alt küme oluşturulur.
- ▶ Rastgele seçilen öznitelik alt kümesi ve mevcut öznitelik alt kümesinin tutarlılık dereceleri karşılaştırılır. – Daha iyi ise yeni küme seçilir.
- ▶ Süreç, belirli bir yinleme sayısınca sürdürülür.

Sezgisel Arama Yöntemleri

- ▶ Öznitelik altkümelerinin değerlendirilmesi-->**Maliyetli**
- ▶ **Sezgisel Arama**
 - ▶ *En iyi önce arama (BFS)*
 - ▶ *Genetik Algoritmalar (GA)*
 - ▶ *Açgözlü arama (GS)*
 - ▶ *Doğrusal İleri Seçim algoritması (LFS)*
 - ▶ *Parçacık Sürü Optimizasyonu (PSO)*
 - ▶ *Sıra arama algoritması (RS)*
 - ▶ *Yeniden Sıralama Algoritması (RRS)*

Sezgisel Arama Yöntemleri

- ▶ **En iyi önce arama algoritması (BFS)**
 - ▶ Mevcut düğümün çocuk düğümleri arasında **en iyi değere** sahip olanın genişletilmek üzere seçilir.
- ▶ **Genetik algoritmalar (GA)**
 - ▶ Evrimsel hesaplamaya dayalı arama yöntemleridir.
 - ▶ Çaprazlama, mutasyon gibi operatörler kullanılarak bir sonraki neslin oluşturulduğu yinelemeli bir süreç.

Sezgisel Arama Yöntemleri

- ▶ **Açgözlü arama algoritması (GS)**
 - ▶ Arama uzayındaki tüm öznitelikler ile ya da hiçbir öznitelik içermeyen küme ile başlanır.
 - ▶ İleri doğru açgözlü aramada, boş altküme ile başlanır.
 - ▶ Her bir adımda, yeni bir öznitelik mevcut öznitelik altkütmesine eklenir.

Sezgisel Arama Yöntemleri

- ▶ **Doğrusal ileri seçim algoritması (LFS)**
 - ▶ En iyi önce arama algoritmasına dayalı,
 - ▶ Hesaplama maliyetini azaltmak için arama sürecinin her bir adımında değerlendirilecek öznitelik sayısı sınırlı tutulur.
- ▶ **Sıra arama algoritması (RS)**
 - ▶ Öznitelikler belirli bir bireysel filtre-tabanlı öznitelik seçim yöntemine göre sıralanır.
 - ▶ Arama süreci, en yüksek ölçüt değerine sahip özniteliklerin sırasıyla eklenmesi ile sürdürülür.

Sezgisel Arama Yöntemleri

- ▶ **Yeniden Sıralama Algoritması (RRS)**
 - ▶ Özniteliklerin sıralamasını dinamik olarak değiştirerek,
 - ▶ Belirli bir özneliğin eklenmesi ile gereksiz hale gelen özniteliklerin ortadan kaldırılmasını
 - ▶ Gerekli hale gelen özniteliklerin eklenmesini amaçlayan bir sıralama algoritmasıdır.

Sezgisel Arama Yöntemleri

- ▶ **Parçacık Sürüsü Optimizasyonu (PSO)**
 - ▶ Eniyileme problemini, P adet parçacıktan oluşan bir toplum aracılığıyla çözmeye çalışan bir yöntemdir.
 - ▶ Burada, parçacıklar çözüm uzayı etrafında pozisyon ve hızlarına dayalı olarak hareket eder.
 - ▶ Belirli bir parçacığın hareketi hem ilgili parçacığın mevcut en iyi pozisyonuna hem de sürüdeki en iyi mevcut çözüme dayalı olarak yönlendirilir.

Veri Setleri

- **Veri temsili:** Terim sıklığı ve 1-gram temsili.

Tablo 1: Veri setlerine ilişkin temel özellikler

Veri Seti	Pozitif Görüş Kutbu Sayısı	Negatif Görüş Kutbu Sayısı	Öznitelik Sayısı (1-gram)
Camera	250	248	1352
Camp	402	402	2045
Doctor	739	739	1578
Drug	401	401	1438
Laptop	88	88	2010
Lawyer	110	110	2474
Music	291	291	1398
Radio	502	502	1923
TV	235	235	2834

DeneySEL Süreç

- ▶ **WEKA 3.7.11** Geliştirici Versiyonu.
- ▶ **Parametreler:** Temel parametreler sabit tutulmuştur.
- ▶ 10-kat çapraz geçirme
- ▶ Değerlendirme Ölçütü:
 - ▶ Doğru Sınıflandırma

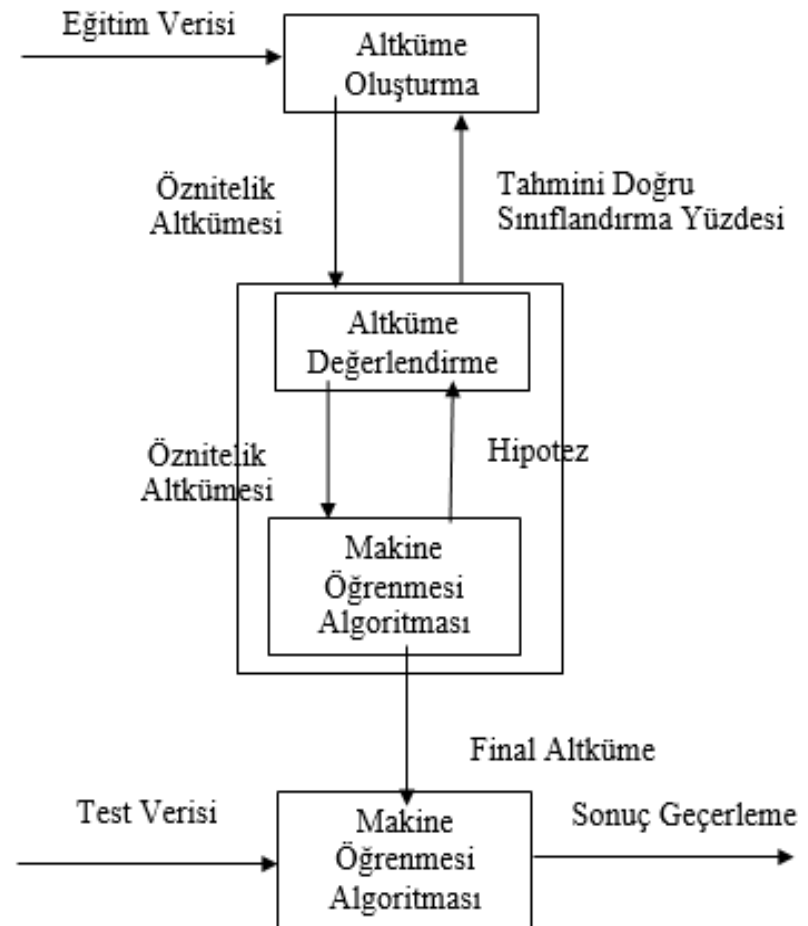
$$ACC = \frac{TN + TP}{TP + FP + FN + TN}$$

Öznitelik Seçimi Deneyleleri

- ▶ **Deney#1: Grup Öznitelik Ölçütlerinin Performans Analizi**
 - ▶ **Ölçütler:** Korelasyon-tabanlı (CBS), Tutarlılık-tabanlı (ConsBS)
 - ▶ **Sezgiseller:** BFS, Genetic, Greedy, LFS, PSO, Rank, ReRank
 - ▶ **Sınıflandırıcılar:**
 - ▶ Naive Bayes (NB),
 - ▶ Destek vektör makineleri (SVM),
 - ▶ C4-5,
 - ▶ K-en yakın komşu (KNN),
 - ▶ Radyal tabanlı fonksiyon ağırları (RBF)

Öznitelik Seçimi Deneyleeri

- ▶ **Deney#2: Sarmalama Tabanlı Öznitelik Seçim Yöntemleri**
 - ▶ **Temel Sınıflandırıcı:**
Naive Bayes (NB), KNN
 - ▶ **Arama Algoritmaları:**
BFS, Genetic, LFS, PSO, Rank
 - ▶ **Özniteliklerin Sınanması:**
Naive Bayes (NB), KNN



Filtre-Tabanlı

Tablo 2: Filtre-Tabanlı Öznitelik Seçim Yöntemlerine İlişkin Deneysel sonuçlar

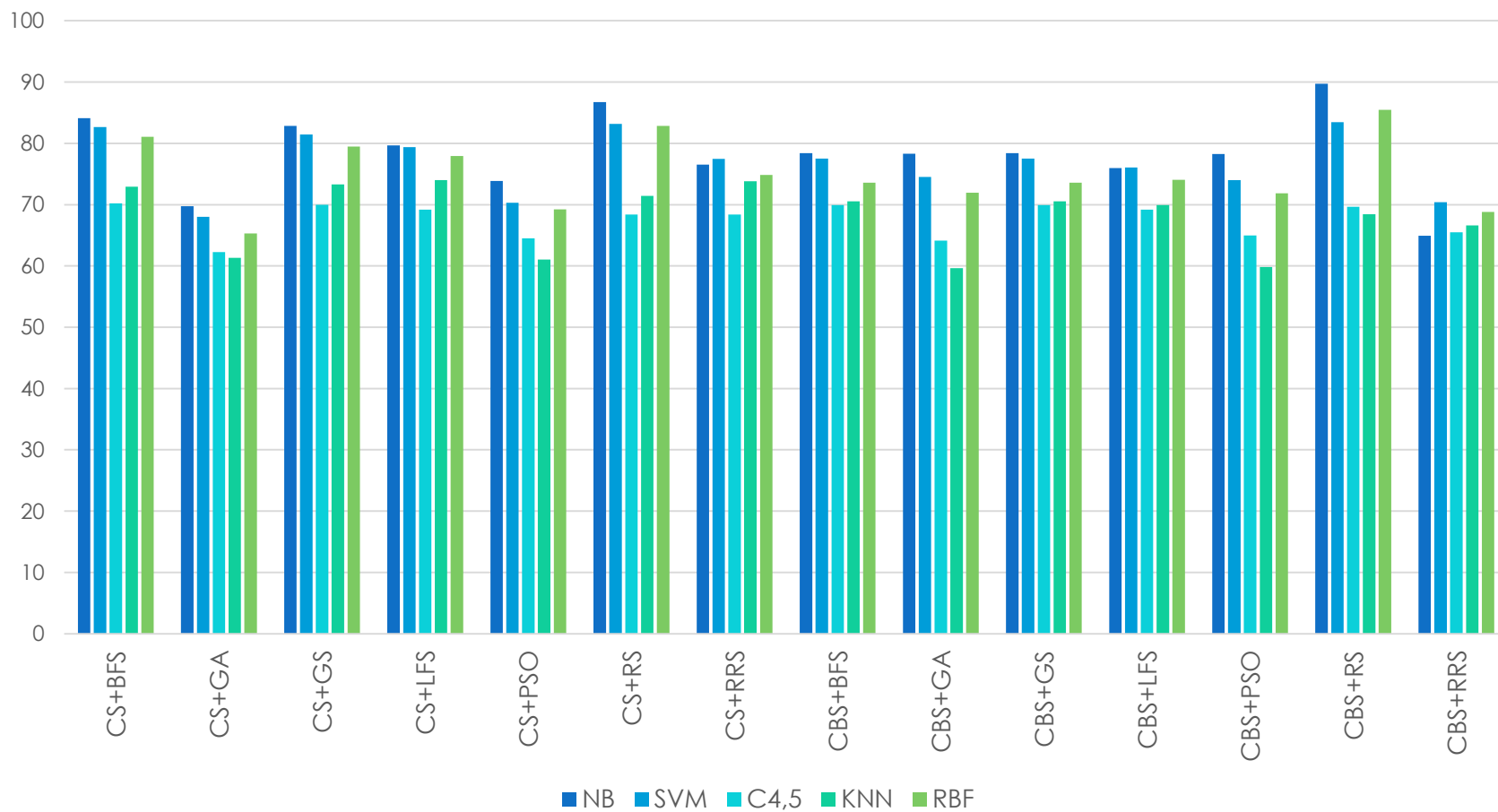
Yöntem	NB	SVM	C4.5	KNN	RBF	Ortalama
CS+BFS	84.09	82.67	<u>70.23</u>	72.93	81.07	78.20
CS+GA	69.76	68	62.27	61.34	65.31	65.34
CS+GS	82.83	81.44	69.98	73.29	79.48	77.40
CS+LFS	79.65	79.39	69.19	<u>74</u>	77.93	76.03
CS+PSO	73.84	70.33	64.52	61.05	69.23	67.79
CS+RS	86.75	83.16	68.41	71.45	82.86	78.53
CS+RRS	76.52	77.45	68.4	73.8	74.86	74.21
CBS+BFS	78.41	77.49	69.95	70.55	73.59	74.00
CBS+GA	78.33	74.51	64.15	59.66	71.96	69.72
CBS+GS	78.41	77.49	69.95	70.55	73.59	74.00
CBS+LFS	75.95	76.08	69.17	69.95	74.05	73.04
CBS+PSO	78.24	74.01	64.95	59.83	71.84	69.77
CBS+RS	<u>89.72</u>	<u>83.47</u>	69.63	68.45	<u>85.46</u>	<u>79.35</u>
CBS+RRS	64.94	70.39	65.5	66.63	68.83	67.26
Ortalama	<u>78.39</u>	76.85	67.59	68.11	75.00	

Sarmalama Tabanlı

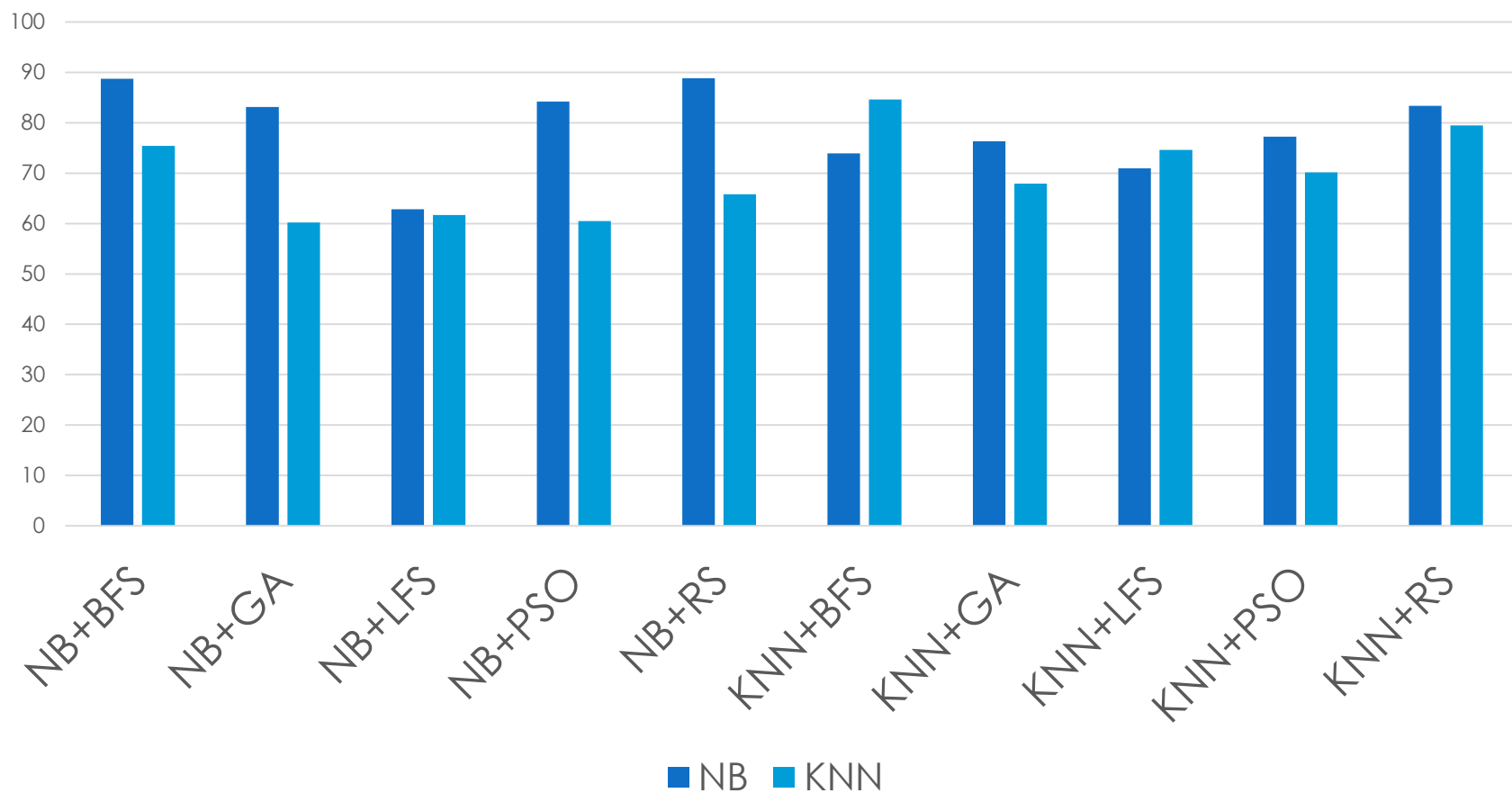
Tablo 3: Sarmalama-Tabanlı Öznitelik Seçim Yöntemlerine İlişkin Deneysel sonuçlar

Yöntem	NB	KNN	Ortalama
NB+BFS	88.74	75.41	82.08
NB+GA	83.14	60.19	71.67
NB+LFS	62.83	61.69	62.26
NB+PSO	84.2	60.51	72.36
NB+RS	<u>88.84</u>	65.79	77.32
KNN+BFS	73.91	<u>84.6</u>	79.26
KNN+GA	76.31	67.93	72.12
KNN+LFS	70.97	74.59	72.78
KNN+PSO	77.24	70.12	73.68
KNN+RS	83.33	79.45	81.39
Ortalama	78.95	70.03	

Filtre-tabanlı



Sarmalama-Tabanlı



Sonuçlar

- ▶ Öznitelik seçim yöntemlerinin performansı
 - ▶ Kullanılan sezgisel arama yöntemine,
 - ▶ Sınıflandırma algoritmasına,
 - ▶ Veri setinin genel özelliklerine göre değişmek ile birlikte;
 - ▶ Sarmalama tabanlı öznitelik seçim yöntemlerinde elde edilen ortalamalar genellikle daha yüksek,
 - ▶ En yüksek başarımlar (%89.72)--> **CBS+RS** ve **NB** ile
 - ▶ En yüksek ikinci başarımlar (%88.84)--> **NB+RS** ve **NB** ile elde edilmekte

Kaynaklar

- ▶ [1] Aggarwal, C.C., Zhai, C.X., "A survey of text classification algorithms", 77-128. **Mining text Data**, Aggarwal, C.C., Zhai, C.X. (Eds.), **Springer-Verlag**, New York, (2012).
- ▶ [2] Korde, V., Mahender, C. N., "Text classification and classifiers: a survey", **International Journal of Artificial Intelligence & Applications**, 3(2): 85-99 (2012).
- ▶ [3] Narendra, P.M., Fukunaga, K., "A branch and bound algorithm for feature selection", **IEEE Transactions on Computers**, 26(9): 917-922 (1977).
- ▶ [4] Chandrashekar, G., Sahin, F., "A survey on feature selection methods", **Computers and Electrical Engineering**, 40: 16-28 (2014).
- ▶ [5] Diao, R., "Feature selection with harmony search and its applications", **Ph.D. Thesis, Aberystwyth University**, 213p (Unpublished) (2014).
- ▶ [6] Hall, M.A., "Correlation-based feature selection for machine learning", **Ph.D. Thesis, University of Waikato**, 198p (Unpublished) (1999).
- ▶ [7] Hall, M.A., Smith, L.A., "Feature selection for machine learning: comparing a correlation-based filter approach to the wrapper", **Proceedings of the Twelfth International Florida Artificial Intelligence Research Society Conference**, USA, 235-239 (1999).
- ▶ [8] Liu, H., Setiono, R., "A probabilistic approach to feature selection: a filter solution", **Proceedings of the Thirteenth International Conference on Machine Learning**, 319-327 (1996).
- ▶ [9] Hall, M. A., Holmes, G., "Benchmarking attribute selection techniques for discrete class data mining", **IEEE Transactions on Knowledge and Data Engineering**, 15(6):1437-1447 (2003).

Kaynaklar

- ▶ [10] Gütlein, M., "Large scale attribute selection using wrappers", **Diploma Thesis, University of Freiburg**, 135p (Unpublished) (2006).
- ▶ [11] Gütlein, M., Frank, E., Hall, M., Karwath, A., "Large-scale attribute selection using wrappers", **Proceedings of IEEE Symposium on Computational Intelligence and Data Mining**, 332-339 (2009).
- ▶ [12] Witten, I.H., Frank, E., Hall, M.A., "Data Mining: Practical Machine Learning Tools and Techniques", **Morgan Kaufmann Publishers**, Burlington, (2011).
- ▶ [13] Bermejo, P., Ossa, L., Gamez, J. A., Puerta, J. M., "Fast wrapper feature subset selection in high-dimensional datasets by means of filter re-ranking", **Knowledge-Based Systems**, 25:35-44 (2012).
- ▶ [14] Han, J., Kamber, M., "Data mining: concepts and techniques", **Morgan Kaufmann Publishers**, Burlington, (2006).
- ▶ [15] Abe, S., "Support vector machines for pattern classification", **Springer**, London, (2010).
- ▶ [16] Bors, A.G., "Introduction to the radial basis function (rbf) networks", **Online Symposium for Electronics Engineering**, 1(1): 1-7 (2001).
- ▶ [17] Niuniu, X., Yuxun, L., "Review of decision trees", **Proc. of the Third IEEE Int. Conf. on Computer Science and Information Technology**, China, 105-109 (2010).
- ▶ [18] Onan, A., Korukoğlu, S., "Görüş Madenciliğinde Sınıflandırıcı Toplulukları", **Proceedings of the 23rd Signal Processing and Communications Applications Conference (SIU)**, Turkey, 212-215 (2015).
- ▶ [19] Whitehead, M., Yaeger, L., "Building a general purpose cross-domain sentiment mining model", **Proceedings of IEEE World Congress on Computer Science and Information Engineering**, 472-476 (2009).