

# K-En Yakın Komşu Algoritması Parametrelerinin Sınıflandırma Performansı Üzerine Etkisinin İncelenmesi

**Erdal TAŞCI\***  
**Aytuğ ONAN\*\***

\*Ege Üniversitesi Bilgisayar  
Mühendisliği Bölümü

\*\*Celal Bayar Üniversitesi Bilgisayar  
Mühendisliği Bölümü



- Giriş
- KNN Algoritmasının Avantajları ve Dezavantajları
- KNN Algoritması ve Genel İşleyişi
- Uzaklık Ölçütleri
- Komşu Sayısı
- Ağırlıklandırma
- Materyal, Yöntem ve Sonuçlar
- Sonuçlar ve Tartışma

İçerik

- K-NN (K-Nearest Neighbor) (K-En Yakın Komşu) algoritması, en yakın komşunun, k değerine göre belirlendiği bir sınıflandırma yöntemidir [1].

Makine  
Öğrenmesi

Örüntü Tanıma

Veri  
Madenciliği

Yapay Zeka

Tıp

Biyoinformatik

Giriş

- Eğitiminin olmaması
- Gerçekleştirmenin kolay olması
- Analitik olarak izlenebilir olması
- Yerel bilgilere uyarlanabilir olması
- Paralel gerçekleştirmeye uygun olması
- Gürültülü eğitim verilerine karşı dirençli olması [2]

K-NN Algoritmasının Avantajları

- Yüksek miktarda bellek alanına gereksinim duyması
- Veriseti ve öznitelik boyutu arttıkça işlem yükünün artması
- Performansın  $k$  komşu sayısı, uzaklık ölçütü, ağırlıklandırma ölçütleri, öznitelik sayısı gibi parametrelere bağlı olması [3]

K-NN Algoritmasının Dezavantajları

- En temel örnek tabanlı öğrenme algoritmaları arasındadır.
- Yeni karşılaşılan bir örnek, eğitim setinde yer alan örnekler ile arasındaki benzerliğe göre sınıflandırılır [4].
- Eğitim setinde yer alan örnekler  $n$  boyutlu sayısal nitelikler ile belirtilir.
- Yeni örneğin sınıf etiketi, K-NN'deki sınıf etiketlerinin çoğunluk oylamasına göre atanır [5].

---

### Eđitim Algoritması

- Eđitim setinde yer alan her bir örneđi  $(x, f(x))$  eđitim örnekleri listesine ekle.

### Sınıflandırma Algoritması

- Sınıflandırılmak üzere verilen  $x_q$  örneđini ařađıdaki kurala göre sınıfla:
    - Eđitim örnekleri arasında yer alan  $x_1, \dots, x_k, x_q$  örneđine en yakın  $k$  tane örneđi temsil etmek üzere,  $x_q$  örneđinin sınıf etiketinin belirlenmesi:
$$f(x_q) \leftarrow \arg \max_{v \in V} \sum_{i=1}^k \delta(v, f(x_i))$$
    - Burada,  $a$  ve  $b$  eřit olduđu takdirde  $\delta(a, b) = 1$  olarak, aksi takdirde  $\delta(a, b) = 0$  olarak alınacaktır.
- 

K-NN Algoritmasının Genel İřleyiři [4]

Uzaklık Ölçütleri  
Komşu Sayısı  
Ağırlıklandırma


K-NN Parametreleri

 Minkowski Uzaklığı

 Öklit Uzaklığı

 Manhattan Uzaklığı

 Chebyshev Uzaklığı

 Dilca Uzaklığı

Uzaklık Ölçütleri

- Minkowski uzaklığı, Öklid uzayında tanımlı bir dizidir.
- Öklid uzaklığı, Manhattan uzaklığı gibi uzaklık ölçütlerinin genelleştirilmiş halidir.
- Herhangi iki nokta  $P$  ve  $Q$  arasındaki Minkowski uzaklığı  $P=(x_1, x_2, \dots, x_n)$  ve  $Q=(y_1, y_2, \dots, y_n)$  olmak üzere, Eşitlik 1'e göre hesaplanır:
- Minkowski ölçütünün  $p=2$  olduğu özel durumu, Öklid uzaklığını,  $p=1$  olduğu özel durumu Manhattan uzaklığını ve  $n \rightarrow \infty$  olduğu özel durum, Chebyshev uzaklığını vermektedir [6].

Minkowski Uzaklığı

$$\left( \sum_{i=1}^n |x_i - y_i|^p \right)^{1/p} \quad (1)$$

- Öklid uzaklığı, sınıflandırma ve kümeleme algoritmalarında en sık kullanılan uzaklık ölçütüdür.
- Öklid uzaklığı, iki nokta arasındaki doğrusal uzaklık olup herhangi iki nokta, P ve Q arasındaki Öklid uzaklığı  $P=(x_1, x_2, \dots, x_n)$  ve  $Q=(y_1, y_2, \dots, y_n)$  olmak üzere, Eşitlik 2'ye göre hesaplanır [6]:

$$\left( \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \right) \quad (2)$$

Öklid Uzaklığı

- Manhattan uzaklığı, n boyutlu iki nokta arasındaki farkların mutlak değerlerinin toplamıdır. Herhangi iki nokta, P ve Q arasındaki Manhattan uzaklığı  $P=(x_1, x_2, \dots, x_n)$  ve  $Q=(y_1, y_2, \dots, y_n)$  olmak üzere, Eşitlik 3'e göre hesaplanır [6]:

$$\left( \sum_{i=1}^n |x_i - y_i| \right) \quad (3)$$

Manhattan Uzaklığı

- Chebyshev uzaklığı (maksimum değer uzaklığı), Minkowski uzaklığının,  $n \rightarrow \infty$  olduğu özel durum olup, iki nokta arasındaki farkların mutlak değerlerinin maksimumu olarak tanımlanmaktadır. Herhangi iki nokta, P ve Q arasındaki Chebyshev uzaklığı  $P=(x_1, x_2, \dots, x_n)$  ve  $Q=(y_1, y_2, \dots, y_n)$  olmak üzere, Eşitlik 4'e göre hesaplanır [7]:

$$\lim_{p \rightarrow \infty} \left( \sum_{i=1}^n |x_i - y_i|^p \right)^{1/p} = \max_{i=1}^n |x_i - y_i| \quad (4)$$

Chebyshev Uzaklığı

- Dilca (Distance Learning in Categorical Attribute) uzaklığı, kategorik öznitelik değerleri arasındaki uzaklığı ölçümlemek için kullanılan iki aşamalı bir ölçüttür [8].
- Bu ölçütte öncelikle, simetrik belirsizlik katsayısı yöntemi kullanılarak öznitelik seçimi işlemi gerçekleştirilerek eş-oluşum tablosu oluşturulmaktadır.
- Ardından, eş-oluşum tablosu üzerinde koşullu olasılık ve Öklid uzaklığına dayalı hesaplama gerçekleştirilerek uzaklık ölçümlenmektedir.

- Bilgi kazancı, yüksek değer içeren özniteliklere karşı taraflıdır.
- Simetrik belirsizlik katsayısı (symmetrical uncertainty) (SU), bilgi kazancının (information gain) (IG) bu problemi ortadan kaldırmak için, bilgi kazancının X ve Y özniteliklerinin entropi değerleri toplamına bölünmesi ile Eşitlik 5'e göre belirlenir:

$$SU = 2 \times \left[ \frac{IG}{H(Y) + H(X)} \right] \quad (5)$$

- Öznitelik seçiminin ardından, uzaklık hesaplaması Eşitlik 6'da belirtilen formüle göre gerçekleştirilir [8]:

$$d(x_i, x_j) = \sqrt{\sum_{Y \in \text{baglam}(X)} \sum_{y_k \in Y} (P(x_i|y_k) - P(x_j|y_k))^2} \quad (6)$$

- Burada,  $x_i$  ve  $x_k$  incelenmekte olan özneliğin aldığı değer çiftleri ve  $x_i, x_j \in X$  dir. Her bir  $Y$  bağlam özneliği için  $x_i$  ve  $x_k$  değerlerine dayalı koşullu olasılık hesaplanıp ardından Öklid uzaklığı alınmaktadır.
- Her bir öznelik için bağlam öznelikleri, simetrik belirsizlik katsayısına dayalı olarak belirli bir sezgisel değerlendirme ölçütü aracılığıyla yapılmaktadır [8].

- K-NN algoritmasında, komşu sayısı ( $k$ ) parametresinin değerine dayalı olarak sınıflandırma yapılmaktadır.
- Sınıflandırma sürecinde,  $k=1$  için, sadece en yakın komşunun bulunduğu sınıfa atanırken,  $k$  sayısı örnek sayısına ( $N$ ) yaklaştıkça veri setinde yer alan tüm veriler dikkate alınmakta ve oylamaya göre seçim yapılmaktadır.

- Komşular için ağırlık değerleri atanması ile sınıflandırılmakta olan örneğe daha yakın olan komşu örneklerin, çoğunluk oylamasına daha fazla katkı koyması amaçlanır.
- En çok kullanılan ağırlık değeri atama yöntemleri, her bir komşunun ağırlığının,  $d$ , komşular arası uzaklık olmak üzere,  $1/d$  ya da  $1/d^2$  şeklinde alınmasıdır [9].

- Veri Setleri

- UCI Machine Learning Repository'den alınmıştır.

- [10]

Tablo 1: Veri setlerine ilişkin temel özellikler

Veri Seti	Örnek Sayısı	Öznitelik Sayısı	Sınıf Sayısı
Breast Cancer Wisconsin	699	10	2
Cardiotocography	2126	23	3
Ionosphere	351	34	2
Leaf	340	16	30
Parkinsons	197	23	2
Thoracic Surgery	470	17	2

Materyal, Yöntem ve Sonuçlar

- Deneysel alıřmalar, WEKA 3.7.11 yazılımı kullanılarak gerekleřtirilmiřtir.
- Komřu sayısı parametresi iin 1'den 10'a kadar olan deęerler kullanılmıřtır.
- Uzaklık lütü kapsamında, Minkowski, klid, Manhattan, Chebyshev ve Dilca uzaklıklarından oluřan 5 farklı durum incelenmiřtir.
- Veri setlerinin eęitim ve test setleri olarak ayrılmasında ise, 10-kat apraz geerleme yntemi kullanılmıř, sınıflandırıcının genelleřtirme performansı hesaplanmıřtır.

Tablo 2: Hata Matrisi [11]

	T / G	Gerçek Sınıf	
		p	n
Tahminlenen Sınıf	Y	Gerçek Pozitif (GP)	Yanlış Pozitif (YP)
	N	Yanlış Negatif (YN)	Gerçek Negatif (GN)
Sütun Toplamı		P	N

$$ACC = \frac{GP + GN}{P + N} \quad (7)$$

Performans Ölçütleri

Uzaklık	Veriseti	Sınıflandırma Doğruluk Oranı (%)									
		1-NN	2-NN	3-NN	4-NN	5-NN	6-NN	7-NN	8-NN	9-NN	10-NN
Dilca	Ionosphere	89.77	89.91	90.45	<b>90.74</b>	90.00	90.57	89.38	90.46	89.21	89.95
Manhattan	Ionosphere	<b>90.74</b>	90.45	88.80	90.09	88.49	89.69	88.35	88.95	87.12	88.09
Minkowski	Ionosphere	87.10	<b>89.77</b>	86.02	87.21	85.10	85.76	84.30	85.22	84.30	84.87
Euclidean	Ionosphere	87.10	<b>89.77</b>	86.02	87.21	85.10	85.76	84.30	85.22	84.30	84.87
Chebyshev	Ionosphere	87.10	<b>87.69</b>	82.25	80.43	78.69	79.41	78.83	79.40	79.18	79.78
Dilca	Breast Cancer Wisconsin	93.46	91.43	94.12	93.52	<b>94.66</b>	94.28	94.45	93.98	94.22	94.02
Manhattan	Breast Cancer Wisconsin	96.44	95.29	<b>96.65</b>	96.12	96.10	95.90	96.24	96.38	96.44	96.41
Minkowski	Breast Cancer Wisconsin	95.35	94.56	96.45	96.14	<b>96.94</b>	96.60	96.65	96.64	96.77	96.60
Euclidean	Breast Cancer Wisconsin	95.35	94.56	96.45	96.14	<b>96.94</b>	96.60	96.65	96.64	96.77	96.60
Chebyshev	Breast Cancer Wisconsin	94.75	94.74	95.85	95.91	96.31	<b>96.50</b>	96.44	<b>96.50</b>	96.41	96.28
Dilca	Cardiotocography	98.96	98.97	<b>99.10</b>	98.90	98.97	98.83	98.84	98.80	98.80	98.74
Manhattan	Cardiotocography	99.09	<b>99.13</b>	99.11	99.07	99.08	99.05	99.07	99.04	99.06	98.94
Minkowski	Cardiotocography	99.02	99.02	<b>99.15</b>	99.00	99.05	99.03	99.00	98.93	98.95	98.80
Euclidean	Cardiotocography	99.02	99.02	<b>99.15</b>	99.00	99.05	99.03	99.00	98.93	98.95	98.80
Chebyshev	Cardiotocography	98.89	98.81	<b>98.90</b>	98.80	98.62	98.50	98.47	98.40	98.40	98.34
Dilca	Leaf	<b>58.38</b>	57.26	56.97	54.76	56.29	53.88	54.06	53.56	52.18	51.47
Manhattan	Leaf	<b>66.26</b>	61.79	60.68	62.97	61.91	61.76	62.38	61.68	61.82	61.18
Minkowski	Leaf	<b>62.62</b>	57.47	58.50	58.03	56.68	56.35	56.85	57.56	57.91	56.32
Euclidean	Leaf	<b>62.62</b>	57.47	58.50	58.03	56.68	56.35	56.85	57.56	57.91	56.32
Chebyshev	Leaf	<b>58.74</b>	53.74	54.26	49.47	50.65	50.65	50.65	49.29	49.74	48.74
Dilca	Parkinsons	<b>89.05</b>	86.81	88.34	88.09	<b>89.05</b>	88.29	87.94	87.23	86.77	87.07
Manhattan	Parkinsons	93.81	91.80	<b>94.78</b>	93.76	93.92	92.69	91.35	91.66	90.17	90.02
Minkowski	Parkinsons	<b>95.91</b>	93.49	93.48	93.65	92.73	91.67	92.17	91.09	91.14	90.51
Euclidean	Parkinsons	<b>95.91</b>	93.49	93.48	93.65	92.73	91.67	92.17	91.09	91.14	90.51
Chebyshev	Parkinsons	<b>92.88</b>	89.58	90.73	90.11	90.89	91.24	90.29	90.20	89.85	89.44
Dilca	Thoracic Surgery	78.66	76.00	82.60	81.11	83.19	82.79	84.45	84.43	<b>84.98</b>	84.81
Manhattan	Thoracic Surgery	77.02	72.04	83.00	79.85	84.64	83.53	84.64	84.21	<b>84.70</b>	83.96
Minkowski	Thoracic Surgery	77.02	71.72	82.81	80.38	84.79	83.77	<b>84.91</b>	84.43	84.87	84.28
Euclidean	Thoracic Surgery	77.02	71.72	82.81	80.38	84.79	83.77	<b>84.91</b>	84.43	84.87	84.28
Chebyshev	Thoracic Surgery	78.51	77.34	84.21	84.38	84.87	84.87	85.09	85.09	<b>85.11</b>	85.09

Uzaklık	Veriseti	Sınıflandırma Doğruluk Oranı (%)									
		1-NN	2-NN	3-NN	4-NN	5-NN	6-NN	7-NN	8-NN	9-NN	10-NN
Dilca	Ionosphere	89.77	89.77	90.48	89.91	90.03	90.20	89.38	89.69	89.21	89.49
Manhattan	Ionosphere	90.74	90.74	88.86	89.49	88.95	88.61	88.69	88.18	87.75	87.49
Minkowski	Ionosphere	87.10	87.41	86.02	86.02	85.33	85.19	84.30	84.56	84.30	84.27
Euclidean	Ionosphere	87.10	87.41	86.02	86.02	85.33	85.19	84.30	84.56	84.30	84.27
Chebyshev	Ionosphere	87.10	86.56	81.88	79.49	78.72	78.64	78.83	78.92	79.18	79.09
Dilca	Breast Cancer Wisconsin	93.46	93.46	94.15	94.46	94.41	94.71	94.71	94.51	94.59	94.51
Manhattan	Breast Cancer Wisconsin	96.44	96.47	96.78	96.81	96.57	96.75	96.57	96.73	96.64	96.50
Minkowski	Breast Cancer Wisconsin	95.35	95.47	96.45	96.25	96.95	97.04	96.74	96.97	96.78	96.77
Euclidean	Breast Cancer Wisconsin	95.35	95.47	96.45	96.25	96.95	97.04	96.74	96.97	96.78	96.77
Chebyshev	Breast Cancer Wisconsin	94.75	95.09	96.10	96.30	96.41	96.55	96.52	96.48	96.38	96.22
Dilca	Cardiotocography	98.96	98.98	99.03	99.06	98.93	98.99	98.85	98.89	98.84	98.85
Manhattan	Cardiotocography	99.09	99.09	99.13	99.15	99.05	99.01	99.01	98.99	99.02	98.99
Minkowski	Cardiotocography	99.02	99.02	99.20	99.23	99.05	99.05	99.00	99.00	98.96	99.01
Euclidean	Cardiotocography	99.02	99.02	99.20	99.23	99.05	99.05	99.00	99.00	98.96	99.01
Chebyshev	Cardiotocography	98.89	98.90	99.04	98.97	98.71	98.65	98.61	98.51	98.43	98.42
Dilca	Leaf	58.38	58.38	58.35	58.56	56.97	56.82	56.09	55.35	55.03	54.59
Manhattan	Leaf	66.26	66.26	64.88	65.41	64.06	64.68	64.03	64.59	64.76	64.38
Minkowski	Leaf	62.62	62.62	61.32	60.79	59.88	61.26	61.44	62.03	62.59	61.82
Euclidean	Leaf	62.62	62.62	61.32	60.79	59.88	61.26	61.44	62.03	62.59	61.82
Chebyshev	Leaf	58.74	58.74	58.56	57.00	55.47	56.59	56.24	55.62	55.50	54.94
Dilca	Parkinsons	89.05	89.10	88.34	88.85	89.16	89.15	88.29	88.14	87.54	87.83
Manhattan	Parkinsons	93.81	93.81	94.83	95.35	94.43	94.43	92.53	92.67	92.27	92.42
Minkowski	Parkinsons	95.91	95.91	94.51	94.26	93.34	92.52	93.50	92.06	93.34	92.26
Euclidean	Parkinsons	95.91	95.91	94.51	94.26	93.34	92.52	93.50	92.06	93.34	92.26
Chebyshev	Parkinsons	92.88	92.88	91.76	92.42	92.22	92.16	91.58	90.56	90.77	90.46
Dilca	Thoracic Surgery	78.66	78.85	81.28	81.74	82.15	82.49	82.87	83.23	83.40	83.57
Manhattan	Thoracic Surgery	77.02	77.02	81.11	80.91	82.57	82.74	83.60	83.87	83.94	84.49
Minkowski	Thoracic Surgery	77.02	77.02	81.32	81.30	82.89	82.72	83.77	83.66	84.51	84.40
Euclidean	Thoracic Surgery	77.02	77.02	81.32	81.30	82.89	82.72	83.77	83.66	84.51	84.40
Chebyshev	Thoracic Surgery	78.51	80.11	83.47	84.40	84.00	84.34	84.91	84.91	84.94	85.02

- $k$  deęerinin uygun deęerde seęilmesi oldukęa önem tařımaktadır.
- $k$  deęeri bydkçe daha dzgn karar sınırları oluřmasına karřın hesaplama yk artacak,  $k$  deęeri kçldkçe ise K-NN grltl veriye daha hassas olacak fakat hızlı ęalıřacaktır.
- Uzaklık fonksiyonları da rneklerin daęılımına uygun olarak seęilmelidir.
- znitelik sayısının artıř gstermesiyle boyut artıř gstermekte ve boyutun kapladığı blgeye dřen nokta sayısı azalmaktadır.

Sonuç, Tartıřma ve neriler

- İlişkisiz özniteliklerin artması, ayırt edicilik kriteri yüksek olan özniteliklerin bilgilerini etkisiz hale getirecek ve K-NN algoritması parametrelerinin güvenilirliğini azaltacaktır.
- Bu kapsamda özniteliklerin seçimi, sınıflandırma öncesi ön işleme aşaması olarak kullanılmalıdır.
- Farklı ağırlıklandırma ölçütleri geliştirilerek veya parametre eniyilemesi yapılarak sınıflandırma performansı artırılabilir.

Sonuç, Tartışma ve Öneriler

- [1]. Cover, T.M. and Hart, P.E., "Nearest neighbor pattern classification". IEEE Transactions on Information Theory, IT-13(1):21–27 (1967).
- [2]. Bhatia, N. and Vandana, "Survey of nearest neighbor techniques", **International Journal of Computer Science and Information Security**, 8(2):302-305 (2010).
- [3]. Liu, H. and Zhang, S., "Noisy data elimination using mutual k-nearest neighbor for classification mining", **Journal of Systems and Software**, 85(5):1067-1074 (2012).
- [4]. Mitchell, T., "Machine Learning", **McGraw Hill**, New York, (1997).
- [5]. Han, J. and Kamber, M., "Data mining: concepts and techniques", **Morgan Kaufmann Publishers**, Burlington, (2006).
- [6]. Kresse, W. and Danko, D.M., "Springer Handbook of Geographic Information", **Springer-Verlag**, Berlin, (2012).
- [7]. Xu, G., Zong, Y. and Yang, Z., "Applied Data Mining", **CRC Press**, New York, (2013).
- [8]. Ienco, D., Pensa, G. and Meo, R., "From context to distance: learning dissimilarity for categorical data clustering", **ACM Transactions on Knowledge Discovery**, 6(1): 1-27 (2012).
- [9]. Doad, P.K. and Bartere, M.M., "A Review : Study of Various Clustering Techniques", **International Journal of Engineering Research & Technology**, 2(11):3141-3145 (2013).
- [10]. Lichman, M. UCI Machine Learning Repository [<http://archive.ics.uci.edu/ml>], **Irvine, CA: University of California, School of Information and Computer Science**, (2013).
- [11]. Taşçı, E., "Akciğer tomografileri kullanılarak yapay zeka ve görüntü işleme tekniklerine dayalı otomatik nodül bölge tespit yöntemi geliştirilmesi", **Yüksek Lisans Tezi, Ege Üniversitesi**, 104p (Yayınlanmamış) (2013).

## Kaynaklar