SAKARYA ÜNİVERSİTESİ

Veri Madenciliği Uygulamaları

Hafta 7

Yrd. Doç.Dr. Nilüfer YURTAY

Sınıflandırma- Mesafeye Dayalı Algoritmalar

7.1 En Yakın k-komşu Algoritması

İnsanlar yeni problemleri çözmeye çalışırken genellikle daha önce çözdükleri benzer problemlerin çözümlerine bakarlar.

Bu teknikte yeni bir durum daha önce sınıflandırılmış benzer, en yakın komşuluktaki k tane olaya bakılarak sınıflandırılır.

K en yakın komşuluğundaki olayların ait olduğu sınıflar sayılır ve yeni durum sayısı fazla olan sınıfa dahil edilir.

Bu yöntemde ilk olarak nitelikler arasındaki mesafeyi ölçmek için bir ölçme yöntemi oluşturulur. Olaylar arasındaki uzaklıklar hesaplandıktan sonra, yeni olayların sınıflandırılması için hâlihazırda sınıflandırılmış olan durumlar temel olarak alınır.

Uzaklık karşılaştırmasına kaç adet olayın dahil edileceği (k 'nın belirlenmesi) ve komşuluk hesaplamalarının nasıl yapılacağına karar verilir.

Komşuluk hesaplamaları yapılırken, daha yakın komşulara daha büyük ağırlık değerleri atanabilir .

Bu yöntemin tercih edilme sebebi, sayısı bilinen veri kümeleri için hızlı ve verimli olmasıdır . Kayıtlar, bir veri uzayındaki noktalar olarak düşünülürse, birbirine yakın olan kayıtlar, birbirinin civarında (yakın komşusu) olur.

K en yakın komşuluğunda temel düşünce "komşunun yaptığı gibi yap" tır. Eğer belirli bir kişinin davranışı tahmin edilmek isteniyorsa, veri uzayında o kişiye yakın, kişinin davranışlarına bakılır.

Bu kişilere ait davranışlarının ortalaması hesaplanır ve bu ortalama belirlenen kişi için tahmin olur.

K en yakın komşuluğunda, k harfi araştırılan komşuların sayısıdır. 5-yakın komşuluğunda, 5 kişiye ve 1-yakın komşuluğunda 1 kişiye bakılır.

K en yakın komşuluğu bir öğrenme tekniği değildir. Daha çok bir araştırma yöntemidir. K en yakın komşuluğu, veri kümesini daha iyi anlamaya yardımcı olur.

K en yakın komşuluk yönteminde sınıflandırılmak istenen olay sayısı arttıkça hesaplamalar için gereken sürede hızlı bir şekilde artar, k en yakın komşuluk modelinin işlem hızını artırmak için genellikle bütün veri hafızada tutulur.

K en yakın komşuluğu tekniği ile n tane kayıttan oluşan bir veri kümesinde, her bir kayıt için tahmin yapılmak istendiğinde, her kayıt, diğer kayıtlarla karşılaştırılmak zorundadır.

Bu da büyük veri kümelerinde karesel karmaşıklığa yol açar. Eğer, bir milyon kayıtlı veri tabanında basit bir K en yakın komşuluğu incelemesi yapılacaksa, bir milyar karşılaştırma yapılması gerekir. Bu, araştırmada sorunlara neden olur.

Genelde veri madenciliği algoritmaları n kayıt sayısı kadar karmaşıklığa sahip olmalıdır. Bu nedenle K en yakın komşuluğu tekniği alt örneklemlerle ya da sınırlı sayıda veri kümesinde kullanılmalıdır.

Algoritma, bilinmeyen bir örneklemin hangi sınıfa dahil olduğunu belirlemek için örüntü uzayını araştırarak bilinmeyen örnekleme en yakın olan **k örneklemi** bulur.

Yakınlık Öklid uzaklığı ile tanımlanır. Daha sonra, bilinmeyen örneklem, k en yakın komşu içinden en çok benzediği sınıfa atanır.

k-en yakın komşu algoritması, aynı zamanda, bilinmeyen örneklem için bir gerçek değerin tahmininde de kullanılabilir.

Eğitim örnekleri yerleştirildikleri özellik uzayında birer nokta ile temsil edilirler.

Sınıfı bulunacak olan örnek bu uzayda kendine en yakın ve sayıca belirli bir örneklemin sınıf değerini alır.

Söz konusu yöntem örnek kümedeki gözlemlerin her birinin , sonradan belirlenen bir gözlem değerine olan uzaklıklarının hesaplanması ve en küçük uzaklığa sahip k sayıda gözlemin seçilmesi esasına dayanmaktadır.

Uzaklıkların hesaplanmasında, aşağıdaki öklit uzaklık formülü ile hesaplanır.

k değeri iyi belirlendiği takdirde olumlu sonuçlar verir.

"k–en yakın komşu algoritması şu adımlardan oluşur:

1.k parametresi belirlenir. Komşuluklarının sayısı belirlenir.

2.Koşuluklara ait uzaklıklar hesaplanır.

3. Hesaplanan uzaklılara göre satırlar sıralanarak bunlar içersinden en küçük k tanesi belirlenir.

4 .Belirlenen satırların hangi sınıfa ait olduğu belirlenerek , tekrarlanan sınıf değeri seçilir.

5. Seçilen sınıf , tahmin edilmesi beklenen gözlem değerinin sınıfı olarak kabul edilir.

7.2 Örnek Çalışma

X	Y	Z
1	3	Negatif
2	5	Pozitif
2	3	Pozitif
3	9	Negatif
4	7	Negatif
5	2	Pozitif
6	8	Pozitif
8	6	Negatif
10	6	Negatif
9	1	Negatif

Yukarıdaki gözlem tablosunu kullanarak , yeni eklenen X=7 ve Y=3 yani (7,3) gözleminin hangi Z sınıfına dahil olduğunu k-en yakın komşu algoritması ile bulmaya çalışalım.

formülü ile

(7,3) noktasının tüm gözlem değerleri ile arasındaki uzaklıkları hesaplayalım.

$$d(1) = \sqrt{(1-7)^2 + (3-3)^2} = 6.00$$

$$d(2) = \sqrt{(2-7)^2 + (5-3)^2} = 5{,}39$$

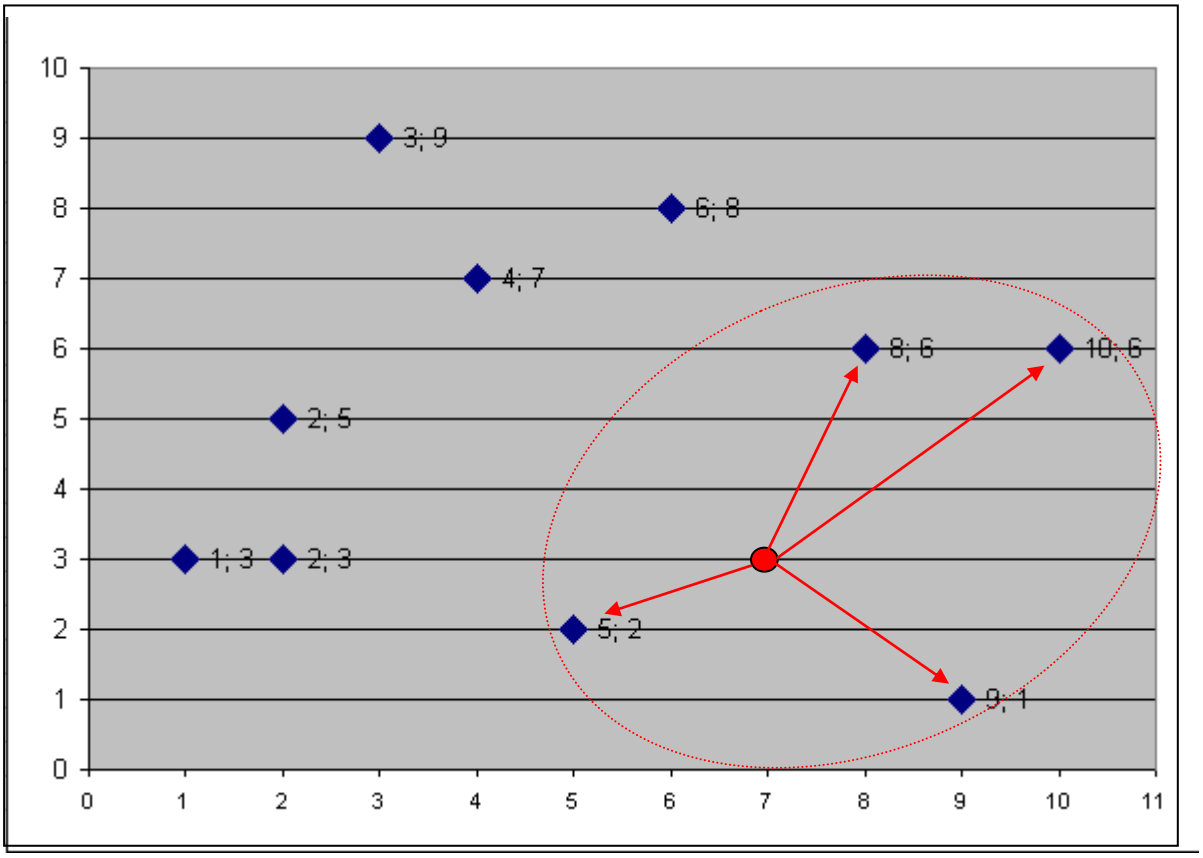
$$d(3) = \sqrt{(2-7)^2 + (3-3)^2} = 5{,}00$$

X	Y	Uzaklık
1	3	6
2	5	5,39
2	3	5
3	9	7,21
4	7	5
5	2	2,24
6	8	5,10
8	6	3,16
10	6	4,24
9	1	2,83

En küçük uzaklıkların belirlenmesi için satırlar sıralanarak en küçük **k=4** tanesi belirleniyor. Belirlenen dört nokta **(7,3)** noktasına en yakın değerlerdir.

X	Y	Uzaklık	Sıralama
1	3	6	9
2	5	5,39	8
2	3	5	6
3	9	7,21	10
4	7	5	5
5	2	2,24	1
6	8	5,10	7
8	6	3,16	3
10	6	4,24	4
9	1	2,83	2

Grafik ile inceleyecek olursak



biçiminde bir komşuluk izlenir.

En küçük satırlara ilişkin sınıfların belirlenmesi işleminde gözlem değerlerinin içinde **hangi değerin baskın olduğuna göre** karar verilir.

X	Y	Z
1	3	Negatif
2	5	Pozitif
2	3	Pozitif
3	9	Negatif
4	7	Negatif
5	2	Pozitif
6	8	Pozitif
8	6	Negatif
10	6	Negatif
9	1	Negatif

Gözlem değerlerin içinde **bir pozitif** ve **üç negatif** değer olduğundan **(7,3)** noktasının sınıfı **negatif** olarak belirlenir.

7.3 Ağırlıklı Oylama

Ağırlıklı oylama yöntemi gözlem değerleri için aşağıdaki bağıntıya göre ağırlıklı uzaklıkların hesaplanması yöntemine dayanır.

$$d(i,j)' = \frac{1}{d(i,j)^2}$$

Sınıf değerlerinin herbiri için uzaklıkların toplamı hesaplanarak ağırlıklı oylama değeri bulunur.

En büyük ağırlıklı oylama değerine sahip olan sınıf değeri yeni gözlem değerinin ait olduğu sınıf olarak belirlenir.

7.4 Örnek Çalışma

X	Y	BAKİYE
0,07	0,25	ARTI
0,02	0,02	ARTI
0,25	0,08	ARTI
1	0,2	EKSİ
0,26	0,3	ARTI
0,14	0,26	ARTI
0,28	0,36	ARTI
0,04	0,11	EKSİ
0,03	0,55	ARTI
0,02	0,87	EKSİ

(0.10, 0.50) gözleminin hangi sınıfa dahil olduğunu k-en yakın komşu algoritmasından ve yukarıdaki tablodan yararlanarak bulunuz.

İlk adım: K'nın belirlenmesi
k=3 olarak seçersek **(0.10, 0.50)** gözlemine en yakın **3** komşuyu arayacağız.

İkinci adım: Uzaklıkların hesaplanması
Öklid uzaklık formülü kullanılarak uzaklıklar hesaplandığında oluşan tabloyu belirleyelim.

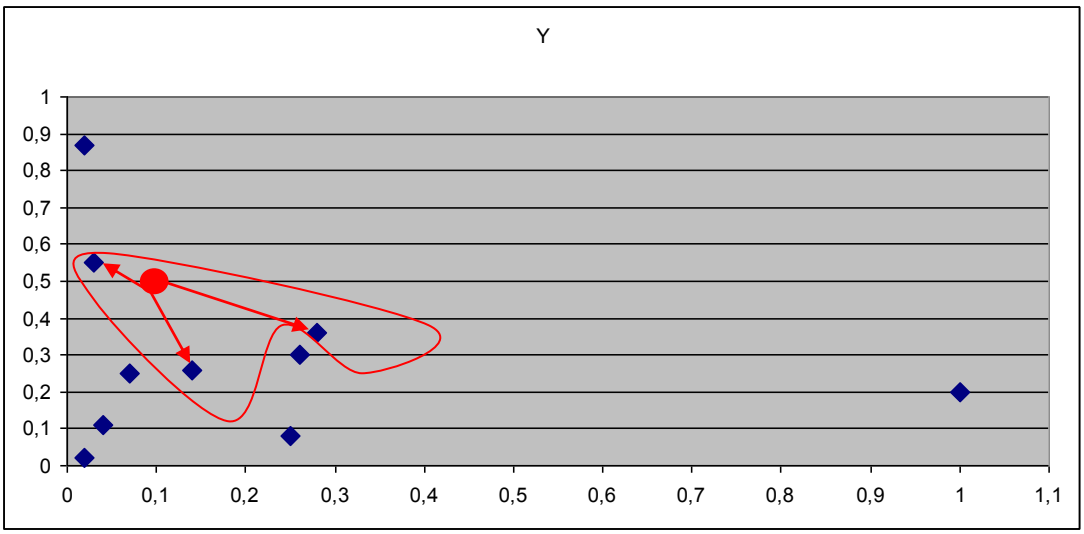
X	Y	UZAKLIK	SIRA
0,07	0,25	0,25	4
0,02	0,02	0,49	9
0,25	0,08	0,45	8
1	0,2	0,95	10
0,26	0,3	0,26	5
0,14	0,26	0,24	3
0,28	0,36	0,23	2
0,04	0,11	0,39	7
0,03	0,55	0,09	1
0,02	0,87	0,38	6

Üçüncü adım: En küçük uzaklıkların belirlenmesi

k=3 olarak seçilen gözlemin belirlenmesi

X	Y	UZAKLIK	SIRA	BAKİYE
0,07	0,25	0,25	4	ARTI
0,02	0,02	0,49	9	ARTI
0,25	0,08	0,45	8	ARTI
1	0,2	0,95	10	EKSİ
0,26	0,3	0,26	5	ARTI
0,14	0,26	0,24	3	ARTI
0,28	0,36	0,23	2	ARTI
0,04	0,11	0,39	7	EKSİ
0,03	0,55	0,09	1	ARTI
0,02	0,87	0,38	6	EKSİ

Grafik ile gösterirsek:



Ağırlıklı Oylama Yöntemin Uygulanması :

$$d(i,j)' = \frac{1}{d(i,j)^2}$$

bağıntısını kullanılarak hesaplamalar yapılır.

X	Y	UZAKLIK	AĞIRLIKLI OYLAMA	SIRA
0,07	0,25	0,25	15,77	4
0,02	0,02	0,49	4,22	9
0,25	0,08	0,45	5,03	8
1	0,2	0,95	1,11	10
0,26	0,3	0,26	15,24	5
0,14	0,26	0,24	16,89	3
0,28	0,36	0,23	19,23	2
0,04	0,11	0,39	6,42	7
0,03	0,55	0,09	135,14	1
0,02	0,87	0,38	6,98	6

$$d(9, gözlem1)' = \frac{1}{(0.09)^2} = 135.14$$

$$d(7, gözlem2)' = \frac{1}{(0.23)^2} = 19.23$$

$$d(6, gözlem3)' = \frac{1}{(0.24)^2} = 16.89$$

Elde edilen bu değerlerin tablo üzerine eklenmesi ile yeni tablo;

X	Y	UZAKLIK	AĞIRLIKLI OYLAMA	SIRA	BAKİYE
0,07	0,25	0,25	15,77	4	ARTI
0,02	0,02	0,49	4,22	9	ARTI
0,25	0,08	0,45	5,03	8	ARTI
1	0,2	0,95	1,11	10	EKSİ
0,26	0,3	0,26	15,24	5	ARTI
0,14	0,26	0,24	16,89	3	ARTI
0,28	0,36	0,23	19,23	2	ARTI
0,04	0,11	0,39	6,42	7	EKSİ
0,03	0,55	0,09	135,14	1	ARTI
0,02	0,87	0,38	6,98	6	EKSİ

biçiminde son halini alır. Bakiyeler içinde hepsi ARTI olduğu için aranan yeni gözlem değerinin sınıfının da ARTI'ya ait olduğu belirlenir.

Kaynaklar

- Veri Madenciğği Yöntemleri Dr. Yalçın Özkan 06’2008
- Veri Madenciliği DR ğökhan Silahtaroğlu 06’2008
- İstanbul Ticaret Üniversitesi Derğisi Ver Madenclğ Modeller Veuyğulama Alanları (Serhat ÖZEKES