

VERİ MADENCİLİĞİ

Veri Önileme

Yrd. Doç. Dr. Şule Gündüz Öğüdücü
www.cs.itu.edu.tr/~gunduz/courses/verimaden/

1

Konular

- Veri
- Veri Önileme
- Benzerlik ve farklılık

www.cs.itu.edu.tr/~gunduz/courses/verimaden/

2

Veri Nedir?

- nesneler ve nesnelerin niteliklerinden oluşan küme
 - kayıt (record), varlık (entity), örnek (sample, instance) nesne için kullanılabilir.
- Nitelik (attribute) bir nesnenin (object) bir özelliğidir
 - bir insanın yaşı, ortamın sıcaklığı..
 - boyut (dimension), özellik (feature, characteristic) olarak da kullanılır
- Nitelikler ve bu niteliklere ait değerler bir nesneyi oluşturur.

Tid	nitelikler				Dolan dırıcı
	Geri Odeme	Medeni Durum	Gelir		
1	Evet	Bekar	125K	-1	
2	Hayır	Evli	100K	-1	
3	Hayır	Bekar	70K	-1	
4	Evet	Evli	120K	-1	
5	Hayır	Boşanmış	95K	1	
6	Hayır	Evli	60K	-1	
7	Evet	Boşanmış	220K	-1	
8	Hayır	Bekar	85K	1	
9	Hayır	Evli	75K	-1	
10	Hayır	Bekar	90K	1	

www.cs.itu.edu.tr/~gunduz/courses/verimaden/

3

Değer Kümeleri

- Nitelik için saptanmış sayılar veya semboller
- Nitelik & Değer Kümeleri
 - aynı nitelik farklı değer kümelerinden değer alabilir
 - ağırlık: kg, lb
 - farklı nitelikler aynı değer kümesinden değer alabilirler
 - ID, yaş: her ikisi de sayısal

www.cs.itu.edu.tr/~gunduz/courses/verimaden/

4

Nitelik Türleri

- Belli aralıkta yeralan değişkenler (interval)
 - sıcaklık, tarih
- İkili değişkenler (binary)
 - cinsiyet
- Ayrık ve sıralı değişkenler (nominal, ordinal, ratio scaled)
 - göz rengi, posta kodu

www.cs.itu.edu.tr/~gunduz/courses/verimaden/

5

Konular

- Veri
- Veri Önileme
- Benzerlik ve farklılık

www.cs.itu.edu.tr/~gunduz/courses/verimaden/

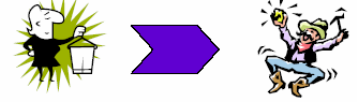
6

Problem

- Uygulamalarda toplanan veri yetersiz, tutarsız ya da gürültülü olabilir.
 - Hata sebepleri:
 - hatalı veri toplama gereçleri
 - veri giriş problemleri
 - veri girişi sırasında kullanıcıların hatalı yorumları
 - veri iletim hataları
 - teknolojik sınırlamalar
 - veri isimlendirmede veya yapısında uyumsuzluk
 - Sonuçları:
 - tekrarlanan kayıtlar
 - gelişkili veriler
 - yetersiz veriler

Problem

- Veri güvenilirmez
 - Veri madenciliği sonuçlarına güvenilebilir mi?
- Veri kaliteli ise veri madenciliği uygulamaları ile yararlı bilgi bulma şansı daha fazla.



Veri Önışleme

- Veri temizleme
- Veri birleřtirme
- Veri dönüşümü
- Veri azaltma

Konular

- Veri
- Veri Önışleme
 - Veri temizleme
 - eksik veri
 - gürültülü veri
 - Veri birleřtirme
 - Veri dönüşümü
 - Veri azaltma
- Benzerlik ve farklılık

Veri Temizleme

- Gerçek uygulamalarda veri eksik, gürültülü veya tutarsız olabilir.
- Veri temizleme işlemleri
 - Eksik nitelik değerlerini tamamlama
 - Aykırılıkların bulunması ve gürültülü verinin düzeltilmesi
 - Tutarsızlıkların giderilmesi



Eksik Veri

- Veri için bazı niteliklerin değerleri her zaman bilinemeyebilir.
- Eksik veri
 - diğer veri kayıtlarıyla tutarsızlığı nedeniyle silinmesi
 - Bazı nitelik değerleri hatalı olması dolayısıyla silinmesi
 - yanlış anlama sonucu kaydedilmeme
 - veri girişi sırasında bazı nitelikleri önemsiz görme

Eksik Veriler nasıl Tamamlanır?

- Eksik nitelik değerleri olan veri kayıtlarını kullanma
- Eksik nitelik değerlerini elle doldur
- Eksik nitelik değerleri için global bir değişken kullan (Null, bilinmiyor,...)
- Eksik nitelik değerlerini o niteliğin ortalama değeri ile doldur
- Aynı sınıfa ait kayıtların nitelik değerlerinin ortalaması ile doldur
- Olasılığı en fazla olan nitelik değeriyle doldur

www.cs.itu.edu.tr/~gunduz/courses/verimaden/

13

Gürültülü Veri

- Ölçülen bir değerdeki hata
- Yanlış nitelik değerleri
 - hatalı veri toplama gereçleri
 - veri girişi problemleri
 - veri iletimi problemleri
 - teknolojik kısıtlar
 - nitelik isimlerinde tutarsızlık

www.cs.itu.edu.tr/~gunduz/courses/verimaden/

14

Gürültülü Veri nasıl Düzeltilir?

- Gürültüyü yok etme
 - Bölmeleme
 - veri sıralanır, eşit genişlik veya eşit derinlik ile bölünür
 - Demetleme
 - aykırılıkları belirler
 - Eğri uydurma
 - veriyi bir fonksiyona uydurarak gürültüyü düzeltir

www.cs.itu.edu.tr/~gunduz/courses/verimaden/

15

Bölmeleme

- Veri sıralanır: 4, 8, 15, 21, 21, 24, 25, 28, 34
 - Eşit genişlik: Bölme sayısı belirlenir. Eşit aralıklarla bölünür
 - Eşit derinlik: Her bölmede eşit sayıda örnek kalacak şekilde bölünür.
 - her bölme ortalamayla ya da bölmenin en alt ve üst sınırlarıyla temsil edilir

Bölme genişliği:3
1. Bölme: 4, 8, 15
2. Bölme: 21, 21, 24
3. Bölme: 25, 28, 34

Ortalamayla düzeltme:
1. Bölme: 9, 9, 9
2. Bölme: 22, 22, 22
3. Bölme: 29, 29, 29

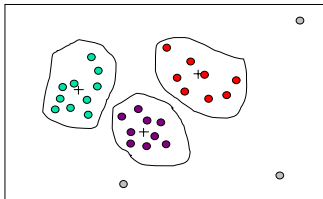
Alt-üst sınırla düzeltme:
1. Bölme: 4, 4, 15
2. Bölme: 21, 21, 24
3. Bölme: 25, 25, 34

www.cs.itu.edu.tr/~gunduz/courses/verimaden/

16

Demetleme

- Benzer veriler aynı demette olacak şekilde gruplanır
- Bu demetlerin dışında kalan veriler aykırılık olarak belirlenir ve silinir

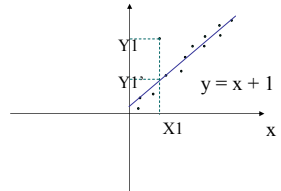


www.cs.itu.edu.tr/~gunduz/courses/verimaden/

17

Eğri Uydurma

- Veri bir fonksiyona uydurulur. Doğrusal eğri uydurmada, bir değişkenin değeri diğer bir değişken kullanılarak bulunabilir.



www.cs.itu.edu.tr/~gunduz/courses/verimaden/

18

Konular

- Veri
- Veri Önışleme
 - Veri temizleme
 - Veri birleřtirme
 - Veri dönüşümü
 - Veri azaltma
- Benzerlik ve farklılık

www.cs.itu.edu.tr/~gunduz/courses/verimaden/

19

Veri Birleřtirme

- Farklı kaynaklardan verilerin tutarlı olarak birleřtirilmesi
- řema birleřtirilmesi
 - Aynı varlıkların saptanması:
A.cust_id=B.cust_num
 - meta veri kullanılır
- Nitelik deęerlerinin tutarsızlıęının saptanması
 - Aynı nitelik için farklı kaynaklarda farklı deęerler olması
 - Farklı metrikler kullanılması



www.cs.itu.edu.tr/~gunduz/courses/verimaden/

20

Gereksiz Veri

- Farklı veri kaynaklarından veriler birleřtirilince gereksiz (fazla) veri oluşabilir
 - aynı nitelik farklı kaynaklarda farklı isimle
 - bir nitelięin deęeri başka bir nitelik kullanılarak hesaplanabilir
 - korelasyon hesaplaması
 - =0: nitelikler bağımsız, >0: pozitif korelasyon, <0: negatif korelasyon

$$r_{A,B} = \frac{\sum (A - \bar{A})(B - \bar{B})}{(n-1)\sigma_A\sigma_B}$$

$$\bar{A} = \frac{\sum A}{n}$$

$$\sigma_A = \sqrt{\frac{\sum (A - \bar{A})^2}{n-1}}$$

www.cs.itu.edu.tr/~gunduz/courses/verimaden/

21

Konular

- Veri
- Veri Önışleme
 - Veri temizleme
 - Veri birleřtirme
 - Veri dönüşümü
 - normalizasyon
 - nitelik oluřturma
 - Veri azaltma
- Benzerlik ve farklılık

www.cs.itu.edu.tr/~gunduz/courses/verimaden/

22

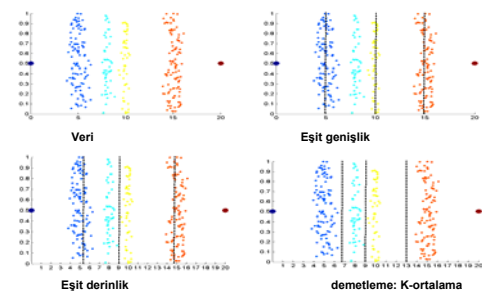
Veri Dönüşümü

- Veri, veri madencilięi uygulamaları için uygun olmayabilir
 - Seçilen algoritmaya uygun olmayabilir
 - Veri belirleyici deęil
- Çözüm
 - Veri düzeltme
 - Bölmeleme
 - Demetleme
 - Eğri Uydurma
 - Bir araya getirme
 - Genelleme
 - Normalizasyon
 - Nitelik oluřturma

www.cs.itu.edu.tr/~gunduz/courses/verimaden/

23

Bölmeleme ve Demetleme Yöntemleri ile Veri Düzeltme



www.cs.itu.edu.tr/~gunduz/courses/verimaden/

24

Normalizasyon

- min-max normalizasyon

$$v' = \frac{v - \min_A}{\max_A - \min_A} (\text{new_max}_A - \text{new_min}_A) + \text{new_min}_A$$

- z-score normalizasyon

$$v' = \frac{v - \text{mean}_A}{\text{stand_dev}_A}$$

- ondalık normalizasyon

$$v' = \frac{v}{10^j} \quad j: \text{Max}(|v'|) < 1 \text{ olacak şekilde en küçük tam sayı}$$

Nitelik Oluşturma

- Yeni nitelikler yarat
 - orjinal niteliklerden daha önemli bilgi içersin
 - veri madenciliği algoritmalarının başarımı daha iyi olsun

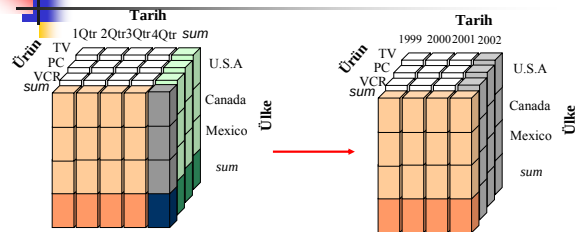
Konular

- Veri
- Veri Önileme
 - Veri temizleme
 - Veri birleştirme
 - Veri dönüşümü
 - Veri azaltma
 - nitelik birleştirme
 - nitelik azaltma
 - veri sıkıştırma
 - veri ayrıştırma ve kavram oluşturma
 - veri küçültme
- Benzerlik ve farklılık

Veri Azaltma

- Veri miktarı çok fazla olduğu zaman veri madenciliği algoritmalarının çalışması ve sonuç üretmesi çok uzun sürebilir
 - veriyi azaltma başarımı artırır
 - sonucun (nerdeyse) hiç değişmemesi gerekir
- Veri azaltma
 - nitelik birleştirme
 - nitelik azaltma
 - veri sıkıştırma
 - veri ayrıştırma ve kavram oluşturma
 - veri küçültme
 - eğri uydurma
 - demetleme
 - histogram
 - örnekleme

Nitelik Birleştirme



- Sorgulama için gerekli olan boyutlar kullanılıyor.

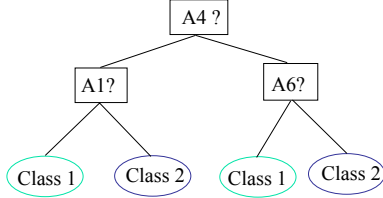
Nitelik Azaltma

- Nitelik seçme
 - Veri madenciliği uygulaması için gerekli olan nitelikleri seç
 - Nitelikler altkümüsi kullanılarak elde edilen sınıfların dağılımları gerçek dağılıma eşit ya da çok yakın olmalı
- Sezgisel yöntemler kullanılarak nitelikler azaltılabilir.
 - istatistiksel anlamlılık testi (statistical significance)
 - bilgi kazancı (information gain)
 - karar ağaçları

Örnek

Başlangıç nitelikler kümesi:

{A1, A2, A3, A4, A5, A6}



Seçilen nitelik kümesi: {A1, A4, A6}

www.cs.itu.edu.tr/~gunduz/courses/verimaden/

31

Veri Sıkıştırma

- Verinin boyutunu azaltır
 - daha az saklama ortamı
 - veriye ulaşmak daha çabuk
- Kayıplı ve kayıpsız veri sıkıştırma
 - bazı yöntemler bazı veri tiplerine uygun
 - her veri tipi için kullanılan yöntemler de var
- Eğer veri madenciliği yöntemi sıkıştırılmış veri üzerinde doğrudan çalışabiliyorsa elverişli

www.cs.itu.edu.tr/~gunduz/courses/verimaden/

32

Veri Ayırıştırma

- Bazı veri madenciliği algoritmaları sadece ayrık veriler ile çalışır
- Sürekli bir nitelik değerini bölerek her aralığı etiketler
- Verinin değeri, bulunduğu aralığın etiketi ile değişir
- Veri boyutu küçülür
- Kavram oluşturmak için kullanılır

www.cs.itu.edu.tr/~gunduz/courses/verimaden/

33

Kavram Oluşturma

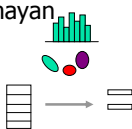
- Sayısal veriler
 - çok geniş aralıkta olabilir
 - değerleri çok sık değişebilir
- Sayısal veriler için kavram oluşturma
 - bölmeleme
 - histogram
 - demetleme
 - entropi
 - kesmeleme

www.cs.itu.edu.tr/~gunduz/courses/verimaden/

34

Veri Küçültme

- Veriyi farklı şekillerde gösterme
 - parametrik
 - eğri uydurma
 - parametrik olmayan
 - histogram
 - demetleme
 - örnekleme

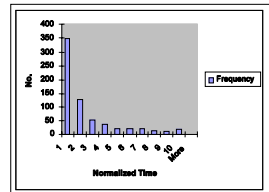


www.cs.itu.edu.tr/~gunduz/courses/verimaden/

35

Histogram ile Veri Küçültme

- Verinin dağılımı
- Veriyi bölerek her bölüm için veri değerini gösterir (toplam, ortalama)
 - eşit genişlik (equi-width): bölmelerin genişliği eşit
 - eşit yükseklik (equi-height): her bölmedeki veri sayısı eşit
 - v-optimal: en az varyansı olan histogram $\sum (count_i * value_i)$
 - MaxDiff: bölme genişliğini kullanıcı belirler

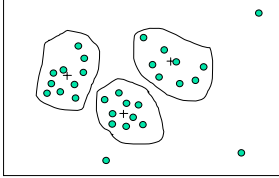


www.cs.itu.edu.tr/~gunduz/courses/verimaden/

36

Demetleme ile Veri Küçültme

- Veri demetlere ayrılır
- Veri demetleri temsil eden örnekler (demet merkezleri) ve aykırılıklar ile temsil edilir
- Etkisi verinin dağılımına bağlı
- Hiyerarşik demetleme yöntemleri kullanılabilir.



www.cs.itu.edu.tr/~gunduz/courses/verimaden/

37

Örnekleme ile Veri Küçültme

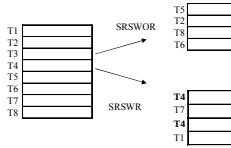
- Büyük veri kümesini daha küçük bir alt küme ile temsil etme
- Alt küme nasıl seçiliyor?
 - yerine koymadan örnekleme (SRSWOR)
 - yerine koyarak örnekleme (SRSWR)
 - demet örnekleme (yerine koymadan veya koyarak)
 - katman örnekleme (katman: nitelik değerine göre grup)

www.cs.itu.edu.tr/~gunduz/courses/verimaden/

38

Örnek

- Örnekleme

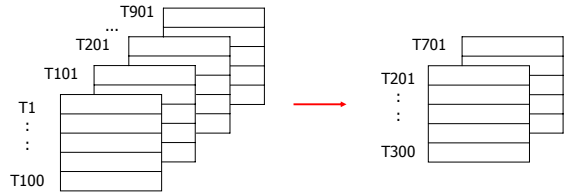


www.cs.itu.edu.tr/~gunduz/courses/verimaden/

39

Örnek

- Demetleme



www.cs.itu.edu.tr/~gunduz/courses/verimaden/

40

Örnek

- Katman Örnekleme

T1	genç	T1	genç
T2	genç	T4	genç
T3	genç	T6	orta yaşlı
T4	genç	T7	orta yaşlı
T5	orta yaşlı	T9	orta yaşlı
T6	orta yaşlı	T11	orta yaşlı
T7	orta yaşlı	T13	yaşlı
T8	orta yaşlı		
T9	orta yaşlı		
T10	orta yaşlı		
T11	orta yaşlı		
T12	yaşlı		
T14	yaşlı		

www.cs.itu.edu.tr/~gunduz/courses/verimaden/

41

Konular

- Veri
- Veri Önışleme
- Benzerlik ve farklılık

www.cs.itu.edu.tr/~gunduz/courses/verimaden/

42

Benzerlik ve Farklılık

- Benzerlik
 - iki nesnenin benzerliğini ölçen sayısal değer
 - nesneler birbirine daha benzer ise daha büyük
 - genelde 0-1 aralığında değer alır
- Farklılık
 - iki nesnenin birbirinden ne kadar farklı olduğunu gösteren sayısal değer
 - nesneler birbirine daha benzer ise daha küçük
 - en küçük farklılık genelde 0
 - üst sınır değişebilir

www.cs.itu.edu.tr/~gunduz/courses/verimaden/

43

Öklid Uzaklığı

- Veri kümesi

$$\begin{bmatrix} x_{11} & \dots & x_{1f} & \dots & x_{1p} \\ \vdots & & \vdots & & \vdots \\ x_{i1} & \dots & x_{if} & \dots & x_{ip} \\ \vdots & & \vdots & & \vdots \\ x_{n1} & \dots & x_{nf} & \dots & x_{np} \end{bmatrix}$$

- Uzaklık matrisi

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & \ddots & \\ d(n,1) & d(n,2) & \dots & \dots & 0 \end{bmatrix}$$

- Öklid uzaklığı (Euclidean Distance) nesneler arası farklılığı bulmak için kullanılır.

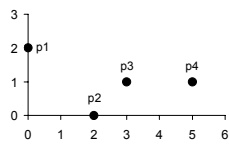
- p adet niteliği (boyutu) olan i ve j nesneleri arasındaki uzaklık

$$d(i, j) = \sqrt{|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \dots + |x_{ip} - x_{jp}|^2}$$

www.cs.itu.edu.tr/~gunduz/courses/verimaden/

44

Örnek: Öklid Uzaklığı



nesne	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

Uzaklık Matrisi

www.cs.itu.edu.tr/~gunduz/courses/verimaden/

45

Minkowski Uzaklığı

- Öklid uzaklığının genelleştirilmiş hali

$$d(i, j) = \sqrt[q]{(|x_{i1} - x_{j1}|^q + |x_{i2} - x_{j2}|^q + \dots + |x_{ip} - x_{jp}|^q)} \quad q: \text{pozitif tam sayı}$$

- $q=1 \rightarrow$ Manhattan uzaklığı

www.cs.itu.edu.tr/~gunduz/courses/verimaden/

46

Uzaklık Özellikleri

- $q=1 \Rightarrow$ Manhattan Uzaklığı
- $q=2 \Rightarrow$ Euclidean Uzaklığı
- Uzaklık ölçütünün sağlaması gereken özellikler:
 1. $d(i, j) \geq 0$
 2. $d(i, i) = 0$
 3. $d(i, j) = d(j, i)$
 4. $d(i, j) \leq d(i, h) + d(h, j)$
- Uzaklıklar ağırlıklı olarak da hesaplanabilir:

$$d(i, j) = \sqrt{w_1 |x_{i1} - x_{j1}|^2 + w_2 |x_{i2} - x_{j2}|^2 + \dots + w_p |x_{ip} - x_{jp}|^2}$$

www.cs.itu.edu.tr/~gunduz/courses/verimaden/

47

Benzerlik Özellikleri

- İki nesne arası benzerlik özellikleri
 1. $\text{sim}(i, j) \geq 0$
 2. $\text{sim}(i, j) = \text{sim}(j, i)$

www.cs.itu.edu.tr/~gunduz/courses/verimaden/

48

İkili Değişkenler Arası Benzerlik

- İkili bir değişkenin 0 veya 1 olarak iki değeri olabilir.
- Bir olasılık tablosu oluşturulur:

		Nesne j	
		0	1
Nesne i	0	M_{00}	M_{01}
	1	M_{10}	M_{11}

M_{00} : i nesnesinin 0, j nesnesinin 0 olduğu niteliklerin sayısı
 M_{10} : i nesnesinin 1, j nesnesinin 0 olduğu niteliklerin sayısı
 M_{01} : i nesnesinin 0, j nesnesinin 1 olduğu niteliklerin sayısı
 M_{11} : i nesnesinin 1, j nesnesinin 1 olduğu niteliklerin sayısı

- Yalın uyum katsayısı** (simple matching coefficient): ikili değişkenin simetrik olduğu durumlarda

$$sint(i, j) = \frac{M_{11} + M_{00}}{M_{00} + M_{01} + M_{10} + M_{11}}$$

- Jaccard katsayısı** (İkili değişkenin asimetrik olduğu durumlar):

$$d(i, j) = \frac{M_{11}}{M_{01} + M_{10} + M_{11}}$$

www.cs.itu.edu.tr/~gunduz/courses/verimaden/

49

Örnek

$$p = 1000000000$$

$$q = 000001001$$

$$M_{01} = 2$$

$$M_{10} = 1$$

$$M_{00} = 7$$

$$M_{11} = 0$$

Yalın Uyum Katsayısı:

$$(M_{11} + M_{00}) / (M_{01} + M_{10} + M_{11} + M_{00}) = (0 + 7) / (2 + 1 + 0 + 7) = 0.7$$

Jaccard Katsayısı:

$$(M_{11}) / (M_{01} + M_{10} + M_{11}) = 0 / (2 + 1 + 0) = 0$$

www.cs.itu.edu.tr/~gunduz/courses/verimaden/

50

Kosinüs Benzerliği

- d_1 ve d_2 iki doküman. Cosine benzerliği
 $\cos(d_1, d_2) = d_1 \bullet d_2 / ||d_1|| ||d_2||$
 $d_1 \bullet d_2$: iki dokümanın vektör çarpımı
 $||d_i||$: d_i dokümanının uzunluğu

- Örnek**

$$d_1 = 3205000200$$

$$d_2 = 1000000102$$

$$d_1 \bullet d_2 = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5$$

$$||d_1|| = (3^2 + 2^2 + 0^2 + 5^2 + 0^2 + 0^2 + 0^2 + 2^2 + 0^2 + 0^2)^{0.5} = (42)^{0.5} = 6.481$$

$$||d_2|| = (1^2 + 0^2 + 0^2 + 0^2 + 0^2 + 0^2 + 0^2 + 1^2 + 0^2 + 2^2)^{0.5} = (6)^{0.5} = 2.245$$

$$\cos(d_1, d_2) = .3150$$

www.cs.itu.edu.tr/~gunduz/courses/verimaden/

51