

Büyük Veri Analitiği (Big Data Analytics)

M. Ali Akcayol
Gazi Üniversitesi
Bilgisayar Mühendisliği Bölümü

Bu dersin sunumları, "Mining of Massive Datasets, Jure Leskovec, Anand Rajaraman, Jeffrey David Ullman, Stanford University, 2011." kitabı kullanılarak hazırlanmıştır.

Konular

- Birliktelik Kuralları
- Birliktelik Kuralı Oluşturulması

Birliktelik Kuralları

- Birliktelik kuralları büyük veri analizlerinde önemli bir yere sahiptir.
- **Birliktelik kural madenciliği veri madenciliğinin temel çalışma alanlarından birisidir.**
- Birliktelik kuralları, üzerinde en çok çalışma yapılan ve yaygın uygulanan modeldir.
- **Birliktelik kuralları modelinin amacı, veri parçalarının birlikte gerçekleşme (co-occurrence) ilişkilerinin tümünü ortaya çıkarmaktır.**
- Agrawal ve arkadaşları* tarafından 1993 yılında önerilmesinden sonra bir çok alanda başarılı bir şekilde uygulanmıştır.

* R. Agrawal, T. Imielinski, and A. N. Swami. Mining Association Rules between Sets of Items in Large Databases. In *Proc. of the ACM SIGMOD Intl. Conf. on Management of Data (SIGMOD'93)*, pp. 207-216, 1993.

3

Birliktelik Kuralları

- Birliktelik kural madenciliğindeki klasik uygulama alanı market sepeti veri analizidir.
- Market sepeti analizinde, **müşterilerin aldıkları ürünler arasındaki ilişkiler ortaya çıkartılır.**

Peynir → Ekmek [support = 10%, confidence = 75%]

- Yukarıdaki kuralda, **tüm müşterilerin %10'u peynir ve ekmeği birlikte almıştır (support).**
- Yukarıdaki kuralda, **peynir alan müşterilerin %75'i ekmeği de almıştır (confidence).**

4

Birliktelik Kuralları

- Birliktelik kural madenciliği ürünlerin sırasını dikkate almaz.
- Sıralı örüntü madenciliği ise ürünlerin sırasıyla ilgilenir.
- Sıralı örüntü madenciliğinde “müşterilerin %5’i önce ekmek sonra peynir daha sonra zeytin almaktadır” şeklinde kural elde edilir.
- Bu tür kurallar Web kullanım madenciliğinde clickstream analizinde veya dilsel örüntülerin bulunması amacıyla kullanılır.

5

Birliktelik Kuralları

- Birliktelik kural madenciliği aşağıdaki gibi ifade edilir:

$$I = \{i_1, i_2, \dots, i_m\}$$

$$T = (t_1, t_2, \dots, t_n)$$

- Burada, I parçalar (items) kümesi, T işlemler (transactions) kümesidir.

- Her bir işlem bir grup parçadan oluşan kümedir ve

$$t_i \subseteq I$$

şeklinde ifade edilir.

- Birliktelik kuralı ise aşağıdaki gibi gösterilir:

$$X \rightarrow Y, \text{ where } X \subset I, Y \subset I, \text{ and } X \cap Y = \emptyset$$

- Burada X ve Y parça kümesi (itemset) olarak ifade edilir.

6

Birliktelik Kuralları

Örnek

- Bir müşteri I kümesindeki ürünlere sahip bir marketten üç tane ürün alıyor. Bu işlem aşağıdaki gibidir.
 $\{\text{Beef, Chicken, Cheese}\}$
- Bir birliktelik kuralı aşağıdaki gibi olabilir.
 $\text{Beef, Chicken} \rightarrow \text{Cheese}$
- Burada,
 $X = \{\text{Beef, Chicken}\}$
 $Y = \{\text{Cheese}\}$
şeklinde ifade edilir.

7

Birliktelik Kuralları

Support

- $t_i \in T$ transaction'ı, X itemset'ini kapsar (eğer X itemset'i t_i transaction'ının alt kümesi ise ($X \subset t_i$)).
- X itemset'inin **destek sayısı (support count)**, T içerisinde X 'i kapsayan transaction sayısıdır ve $X.count$ şeklinde gösterilir.
- $X \rightarrow Y$ kuralının **support değeri**, T tüm transaction'lar kümesinde $X \cup Y$ itemset'ini içeren **transaction'ların oranıdır**.
- Bir kuralın **support değeri, olma olasılığını gösterir** ve $\Pr(X \cup Y)$ şeklinde ifade edilir.
- $X \rightarrow Y$ kuralının support değeri aşağıdaki eşitlikle hesaplanır. Burada, n toplam transaction sayısını gösterir.

$$support = \frac{(X \cup Y).count}{n}$$

8

Birliktelik Kuralları

Confidence

- $X \rightarrow Y$ kuralının **confidence değeri**, T tüm transaction'lar kümesinde X ile Y itemset'lerini **birlikte içeren transaction'ların**, X' i **bulunduran transactionlar içerisindeki oranıdır**.
- Bir kuralın **confidence değeri**, **şartlı olasılık ile ifade edilir** ve $\Pr(Y | X)$ şeklinde gösterilir.
- $X \rightarrow Y$ kuralının confidence değeri aşağıdaki eşitlikle hesaplanır. Burada, n toplam transaction sayısını gösterir.

$$confidence = \frac{(X \cup Y).count}{X.count}$$

- **Confidence bir kuralın öngörülebilirliğini (predictability) belirler.**

9

Birliktelik Kuralları

- **Birliktelik kural madenciliği**, verilen bir T transaction'lar kümesinde, **önceden belirlenmiş support ve confidence değerlerine eşit veya büyük olan birliktelik kurallarının elde edilmesini amaçlar**.
- Kullanıcı tarafından belirlenmiş olan support ve confidence değerlerine **minimum support (minsup)** ve **minimum confidence (minconf)** denilmektedir.

10

Birliktelik Kuralları

Örnek

- Aşağıda 7 transaction'a sahip olan T kümesi verilmiştir.
- I kümesi markette satılan tüm parçaları içermektedir.
- **minsup = 30%** ve **minconf = 80%** için kurallar bulalım.

t₁: Beef, Chicken, Milk
t₂: Beef, Cheese
t₃: Cheese, Boots
t₄: Beef, Chicken, Cheese
t₅: Beef, Chicken, Clothes, Cheese, Milk
t₆: Chicken, Clothes, Milk
t₇: Chicken, Milk, Clothes

- Transaction'lar için aşağıdaki kurallar **support = 42,86% (>30%)** ve **confidence = 100% (>80%)** değerine sahiptir ve geçerlidir.

Chicken, Clothes → Milk [sup = 3/7, conf = 3/3]

Clothes → Milk, Chicken [sup = 3/7, conf = 3/3]

11

Birliktelik kuralları

- Frequent itemset'lerin genellikle if-then kuralları ile gösterilmesi istenir.
- Bu kurallara **birliktelik kuralları** (association rules) denilmektedir.
- I item kümesi ve j bir item olmak üzere aşağıdaki gibi gösterilir:

$$I \rightarrow j$$

- I item kümesindeki tüm item'lar bir sepette varsa, j 'de bu sepette vardır.
- I item kümesindeki tüm item'lar ve j item'ının birlikte bulunma oranı destek (support) değeridir.
- I item kümesindeki tüm item'ların bulunduğu sepetlerde j item'ının da bulunma oranı güven (confidence) değeridir.

12

Birliktelik kuralları

1. {Cat, and, dog, bites}
2. {Yahoo, news, claims, a, cat, mated, with, a, dog, and, produced, viable, offspring}
3. {Cat, killer, likely, is, a, big, dog}
4. {Professional, free, advice, on, dog, training, puppy, training}
5. {Cat, and, kitten, training, and, behavior}
6. {Dog, &, Cat, provides, dog, training, in, Eugene, Oregon}
7. {"Dog, and, cat", is, a, slang, term, used, by, police, officers, for, a, male-female, relationship}
8. {Shop, for, your, show, dog, grooming, and, pet, supplies}

- {cat, dog} → and, kuralının **support** değeri 3/8, **confidence** değeri 3/5 tir.
- {cat} → kitten, kuralının **support** değeri 1/8, **confidence** değeri 1/6 dir.

13

Birliktelik kuralları

- $I \rightarrow j$ birliktelik kuralı için **interest** değeri, kuralın **confidence** değeri ile j 'nin **support** değerinin farkıyla hesaplanır.

- $Interest = (I \rightarrow j).conf - j.sup$
- Sıfıra yakın değerler $I \rightarrow j$ kuralının önemli olmadığını gösterir (rastgele dağılımda sıfır olur).
- Yüksek negatif değerler I 'nin j ile birlikte olmadığını (önemli) gösterir.
- Yüksek pozitif değerler $I \rightarrow j$ kuralının önemli olduğunu gösterir.
- {dog} → cat, kuralının interest değeri = $5/7 - 6/8 = -0,036$ dir.
- {dog} → cat, birliktelik kuralı önemli değildir.
- {cat} → kitten, kuralının interest değeri = $1/6 - 1/8 = 0,042$ dir.

1. {Cat, and, dog, bites}
2. {Yahoo, news, claims, a, cat, mated, with, a, dog, and, produced, viable, offspring}
3. {Cat, killer, likely, is, a, big, dog}
4. {Professional, free, advice, on, dog, training, puppy, training}
5. {Cat, and, kitten, training, and, behavior}
6. {Dog, &, Cat, provides, dog, training, in, Eugene, Oregon}
7. {"Dog, and, cat", is, a, slang, term, used, by, police, officers, for, a, male-female, relationship}
8. {Shop, for, your, show, dog, grooming, and, pet, supplies}

14

Konular

- Birliktelik Kuralları
- Birliktelik Kuralı Oluşturulması

15

Birliktelik kuralı oluşturulması

- Bazı uygulamalarda **frequent itemset'leri kullanarak tüm birliktelik kurallarının oluşturulması gereklidir.**
- Bir f frequent itemset için **kural oluştururken f kümesinin tüm altkümeleri kullanılır.**
- **Confidence** değerine bağlı olarak bir frequent itemset için kural aşağıdaki gibi elde edilebilir:

$$(f - \alpha) \rightarrow \alpha, \text{ if}$$

$$confidence = \frac{f.count}{(f - \alpha).count} \geq minconf,$$

- Burada $f.count$ ile $(f - \alpha).count$ tüm transaction kümesi T 'deki sayılarını gösterir.
- f frequent item için support değeri ise $f.count / n$ ile gösterilir.

16

Birliktelik kuralı oluşturulması

- Eğer f frequent itemset ise, Apriori algoritması **boş küme hariç** tüm altkümelerinin de frequent itemset olmasını gerektirir.
- Eğer $(f-\alpha) \rightarrow \alpha$ kuralı geçerliyse, $(f-\alpha_{sub}) \rightarrow \alpha_{sub}$ kuralları da geçerlidir.
- Eğer $\{A, B, C, D\}$ itemset için $(A, B \rightarrow C, D)$ kuralı varsa, $(A, B, C \rightarrow D)$ ve $(A, B, D \rightarrow C)$ kuralları da **kesinlikle geçerlidir**.
 $\{A, B, C\}.count \leq \{A, B\}.count$, $\{A, B, D\}.count \leq \{A, B\}.count$
 $(A, B, C \rightarrow D).conf \geq (A, B \rightarrow C, D).conf$
- Eğer $\{A, B, C, D\}$ itemset için $(A, B, C \rightarrow D)$ kuralı varsa, $(A, B \rightarrow C, D)$ ve $(A, C \rightarrow B, D)$ kuralları **geçerli olmayabilir**.
 $\{A, B\}.count \geq \{A, B, C\}.count$, $\{A, C\}.count \geq \{A, B, C\}.count$
 $(A, B, C \rightarrow D).conf \leq (A, B \rightarrow C, D).conf$

17

Birliktelik kuralı oluşturulması

- Öncelikle 1 elemana sahip ardıl kısımlardan oluşan kurallar elde edilir.

Algorithm genRules(F) // F is the set of all frequent itemsets

- 1 for each frequent k -itemset f_k in F , $k \geq 2$ do
- 2 output every 1-item consequent rule of f_k with confidence $\geq minconf$ and support $\leftarrow f_k.count / n$ // n is the total number of transactions in T
- 3 $H_1 \leftarrow \{ \text{consequents of all 1-item consequent rules derived from } f_k \text{ above} \};$
- 4 ap-genRules(f_k, H_1);
- 5 endfor

Son elemanlardan oluşan küme
 $H_1 = \{\{Chicken\}, \{Milk\}\}$

1 elemanlı ardıl kısımlara sahip kurallar oluşturulur.

Rule 1: Chicken, Clothes $\rightarrow$ Milk [sup = 3/7, conf = 3/3]
 Rule 2: Chicken, Milk $\rightarrow$ Clothes [sup = 3/7, conf = 3/4]
 Rule 3: Clothes, Milk $\rightarrow$ Chicken [sup = 3/7, conf = 3/3].

minsup = %30 ve minconf = %80

18

Birliktelik kuralı oluşturulması

$F_3: \{\{\text{Chicken, Clothes, Milk}\}:3\}$
 $H_1 = \{\{\text{Chicken}\}, \{\text{Milk}\}\}$

Procedure ap-genRules(f_k, H_m) // H_m is the set of m -item consequents

- 1 if ($k > m + 1$) AND ($H_m \neq \emptyset$) then
- 2 $H_{m+1} \leftarrow \text{candidate-gen}(H_m);$ $H_2 = \{\{\text{Chicken, Milk}\}\}$
- 3 for each h_{m+1} in H_{m+1} do
- 4 $\text{conf} \leftarrow f_k.\text{count} / (f_k - h_{m+1}).\text{count};$ $\{\{\text{Chicken, Clothes, Milk}\}.\text{count} / \text{Clothes.count}\}$
- 5 if ($\text{conf} \geq \text{minconf}$) then
- 6 output the rule ($f_k - h_{m+1}$) $\rightarrow h_{m+1}$ with confidence = conf and
support = $f_k.\text{count} / n$; // n is the total number of transactions in T
- 7 else
- 8 delete h_{m+1} from H_{m+1} ; $\text{Clothes} \rightarrow \text{Milk, Chicken}$
- 9 endfor
- 10 ap-genRules(f_k, H_{m+1});
- 11 endif

$F_3: \{\{\text{Chicken, Clothes, Milk}\}:3\}$
 $H_2 = \{\{\text{Chicken, Milk}\}\}$

19

Birliktelik kuralı oluşturulması

Örnek

- Aşağıdaki T kümesi için minsup = %30 (3/7) ve minconf = %80 (4/5) olsun.

t_1 : Beef, Chicken, Milk
 t_2 : Beef, Cheese
 t_3 : Cheese, Boots
 t_4 : Beef, Chicken, Cheese
 t_5 : Beef, Chicken, Clothes, Cheese, Milk
 t_6 : Chicken, Clothes, Milk
 t_7 : Chicken, Milk, Clothes

- Aşağıdaki frequent itemset'ler elde edilir.

$F_1: \{\{\text{Beef}\}:4, \{\text{Cheese}\}:4, \{\text{Chicken}\}:5, \{\text{Clothes}\}:3, \{\text{Milk}\}:4\}$
 $F_2: \{\{\text{Beef, Cheese}\}:3, \{\text{Beef, Chicken}\}:3, \{\text{Chicken, Clothes}\}:3, \{\text{Chicken, Milk}\}:4, \{\text{Clothes, Milk}\}:3\}$
 $F_3: \{\{\text{Chicken, Clothes, Milk}\}:3\}$

20

Birliktelik kuralı oluşturulması

Örnek – devam

- F_3 için kurallar aşağıdaki gibi elde edilebilir.
Rule 1: Chicken, Clothes $\rightarrow$ Milk [sup = 3/7, conf = 3/3]
Rule 2: Chicken, Milk $\rightarrow$ Clothes [sup = 3/7, conf = 3/4]
Rule 3: Clothes, Milk $\rightarrow$ Chicken [sup = 3/7, conf = 3/3].
- minconf = %80 gerektiğinden dolayı **genRules()** algoritmasının **2.satırında** 1. ve 3. frequent itemset'ler alınır.
- Böylelikle $H_1 = \{\{\text{Chicken}\}, \{\text{Milk}\}\}$ olarak elde edilir.
- F_3 ve H_1 için ap-GenRules() algoritması çağrılır. **Bu algoritmanın 2.satırındaki candidate-gen()** fonksiyonuyla $H_2 = \{\{\text{Chicken, Milk}\}\}$ elde edilir.

21

Birliktelik kuralı oluşturulması

Örnek – devam

- ap-GenRules() algoritmasının 2.satırında $H_2 = \{\{\text{Chicken, Milk}\}\}$ elde edildikten sonra aşağıdaki 4.kural oluşturulur:

Rule 4: Clothes $\rightarrow$ Milk, Chicken [sup = 3/7, conf = 3/3].
- $F_3 = \{\text{Chicken, Clothes, Milk}\}$ frequent itemset'i kullanılarak aşağıdaki 4 kural elde edilmiş olur.

Rule 1: Chicken, Clothes $\rightarrow$ Milk [sup = 3/7, conf = 3/3]
~~Rule 2: Chicken, Milk $\rightarrow$ Clothes [sup = 3/7, conf = 3/4]~~
Rule 3: Clothes, Milk $\rightarrow$ Chicken [sup = 3/7, conf = 3/3].
Rule 4: Clothes $\rightarrow$ Milk, Chicken [sup = 3/7, conf = 3/3].
- Bunlardan bir tanesi minconf değerini sağlamadığından silinir.

22



Ödev

- Stream'deki frequent itemset'lerin sayılması hakkında bir araştırma ödevi hazırlayınız.