



Hacettepe Üniversitesi
Fen Fakültesi İstatistik Bölümü

Birliktelik Kuralları
Association Rules

Metin USLU

İST 405 Veri Madenciliği

Ankara 2013

1. BİRLİKTELİK KURALLARI

1.1 Tanım

Olayların birlikte gerçekleşme durumlarını çözümleyen veri madenciliği yöntemlerine **birliktelik kuralları** denir. Bu yöntemler, birlikte olma kurallarını belirli olasılıklarla ortaya koyar. Birliktelik kuralı, geçmiş verileri analiz edilerek bu veriler içindeki birliktelik davranışlarının tespiti ile geleceğe yönelik çalışmalar yapılmasını destekleyen bir yaklaşımdır.

Birliktelik kuralları analizi problemi ilk olarak Agrawal, Imielinski ve Swami tarafından 1993 yılında ele alınmış olup ve veri madenciliğinde kullanılan ilk tekniklerden birisidir (Agrawal vd., 1993).

Örneğin, bir marketten müşterilerin süt ve peynir satın alımlarının % 70'inde bu ürünler ile birlikte yoğurt da satın alınmıştır. Bu tür birliktelik örüntüsünün tespit edilebilmesi için, örüntü içinde yer alan ürünlerin birden çok satın alma hareketinde birlikte yer alması gerekir. Milyonlarca veri üzerinde veri madenciliği teknikleri uygulandığında, birliktelik sorgusu için kullanılan algoritmalar hızlı olmalıdır (Agrawal ve Srikant, 1995).

1.1.1. Pazar Sepeti Analizi

Birliktelik kuralı uygulamasına pazar sepeti analizi örnek verilebilir (Frawley vd., 1991). Birliktelik kuralındaki amaç; alışveriş esnasında müşterilerin satın aldıkları ürünler arasındaki birliktelik ilişkisini bulmak, bu ilişki verisi doğrultusunda müşterilerin satın alma alışkanlıklarını tespit etmektir. Bu sayede mağaza yöneticileri, müşterilerin birlikte satın almaya eğilimli oldukları ürünleri yakın raflara koyar. Keşfedilen bu birliktelik bağıntıları ve alışkanlıklar sayesinde etkili ve kazançlı pazarlama ve satış imkanına sahip olmaktadır.

1.2 Matematiksel Modeli

Birliktelik kuralının matematiksel modeli Agrawal, Imielinski ve Swami tarafından 1993 yılında sunulmuştur (Agrawal vd., 1993).

Bu modelde, $I = \{i_1, i_2, \dots, i_m\}$ kümesine "ürünler" adı verilmektedir. D, veri bütünlüğündeki tüm hareketleri, T ise ürünlerin her bir hareketini simgeler. TID ise, her harekete ait olan tek belirteçtir.

Birliktelik kuralı şu şekilde tanımlanabilir;

$$A_1, A_2, \dots, A_m \Rightarrow B_1, B_2, \dots, B_n$$

Bu ifadede yer alan, A_i ve B_j , yapılan iş veya nesnelerdir. Bu kural, genellikle " A_1, A_2, \dots, A_m " iş veya nesneleri meydana geldiğinde, sık olarak " B_1, B_2, \dots, B_n " iş veya nesnelerinin aynı olay veya hareket içinde yer aldığını belirtir (Zhu, 1998).

1.3 Avantajları

Birliktelik kurallarının avantajları aşağıdaki gibidir:

- ✓ Kolay ve anlaşılır sonuçlar üretir.
- ✓ Değişik boyutlardaki veriler üzerinde çalışılabilir.
- ✓ Her ne kadar kayıtların sayısı ve kombinasyon seçimine göre işlem adedi artsa da birliktelik analizi için her adımda gerekli olan hesaplamalar, diğer yöntemlere göre (genetik algoritmalar, yapay sinir ağları gibi) çok daha basittir.

1.4 Dezavantajları

Birliktelik kurallarının dezavantajları aşağıdaki gibidir:

- ✓ Sorunun boyutu büyüdükçe, gerekli hesaplamalar üstel olarak artar.
- ✓ Kayıtlarda çok az rastlanan ürünleri yok sayar.
- ✓ Destek ve güven eşik değerleri, üretilen kural sayısına sınırlama getirirler; fakat eşik değerlerinin çok düşük belirlendiği durumlarda kullanıcı, gerçekten ilgilendiği kuralları kaybetme tehlikesi ile karşı karşıya kalır.

1.5 Türleri

Birliktelik kuralları türleri aşağıdaki gibidir:

- ✓ Genelleştirilmiş birliktelik kuralları
- ✓ Uzaysal simetrik kuralları
- ✓ Nicel birliktelik kuralları
- ✓ Aralık veri kuralları
- ✓ Çoklu min-destek birliktelik kuralları
- ✓ Multimedya birliktelik kuralları
- ✓ Maksimal birliktelik kuralları

1.6 Uygulama Alanları

Birliktelik kurallarının uygulama alanları aşağıdaki gibidir:

- ✓ Pazar sepeti analizi
- ✓ Çapraz pazarlama
- ✓ Promosyon analizi
- ✓ Katalog ve yerleşim düzeni
- ✓ Eğitim
- ✓ Tıp
- ✓ Mühendislik
- ✓ Finans

- ✓ Telekom
- ✓ Pazarlama
- ✓ Bankacılık ve sigorta

1.7 Kullanılan Ölçütler

Birliktelik kurallarının bulunması iki temel adıma dayanmaktadır. Bu kurallar minimum destek ve minimum güven durumunu sağlamalıdır.

- I. Tüm sık geçen ürün kümelerinin bulunması
- II. Sık geçen ürün kümelerinden güçlü ilişki kurallarının yaratılması

Market Sepeti Analizinde, satılan ürünler arasındaki ilişkileri ortaya koymak için:

- I. Destek (Support)
- II. Güven (Confidence)

ölçütlerinden yararlanılır.

Destek ve Güven ölçütleri ne kadar büyükse birliktelik kurallarının da o kadar güçlü olduğuna karar verilir. Bu ölçütlerin hesaplanmasında **destek sayısı** adı verilen bir değer kullanılır.

1.7.1 Destek Sayısı

A ve B ürün gruplarını birlikte içeren alışveriş sayısını gösterir. **sayı(A,B)** şeklinde gösterilir.

1.7.2 Birliktelik Kuralı

A ürün grubunu alanların B ürün grubunu da alma durumu; yani birliktelik kuralı $A \rightarrow B$ biçiminde gösterilir.

1.7.3 Kural Destek Ölçütü

Bir ilişkinin tüm alışverişler içinde hangi oranda tekrarlandığını belirler.

$$\text{destek}(A \rightarrow B) = (\text{sayı}(A, B)) / N$$

N = Tüm alışverişlerin sayısı

1.7.4 Kural Güven Ölçütü

A ürün grubunu alan müşterilerin B ürün grubunu da alma olasılığını ortaya koyar.

$$\text{güven}(A \rightarrow B) = (\text{sayı}(A, B)) / (\text{sayı}(A))$$

$\text{sayı}(A)$ = A ürününü satın alan müşteri sayısı

1.8 Algoritmalar

Birliktelik Kurallarında yaygın olarak kullanılan algoritmalar aşağıdaki gibidir:

- ✓ Apriori
- ✓ Carma
- ✓ Sequence
- ✓ GRI

Ayrıca şu algoritmalar da kullanılmaktadır:

I. Sıralı Algoritmalar

- ✓ AIS
- ✓ SETM
- ✓ APRIORI-TID
- ✓ APRIORI-HYBRID
- ✓ OCD
- ✓ PARTITIONING
- ✓ SAMPLING
- ✓ DIC
- ✓ FP-GROWTH

II. Paralel Algoritmalar

- ✓ CD
- ✓ PDM
- ✓ DMA
- ✓ CCPD
- ✓ DD

2 APRIORI ALGORİTMASI

2.1 Tanım

Apriori Algoritmasının ismi, bilgileri bir önceki adımdan aldığı için “prior” anlamında Apriori’dir (Agrawal ve Srikant, 1994). Bu algoritma temelinde iteratif (tekrarlayan) bir niteliğe sahiptir (Han ve Kamber, 2006) ve hareket bilgileri içeren veritabanlarında sık geçen öge kümelerinin keşfedilmesinde kullanılır.

Apriori algoritmasına göre, eğer k-öge kümesi (k adet elemana sahip öge kümesi) minimum destek ölçütünü sağlıyorsa, bu kümenin alt kümeleri de minimum destek ölçütünü sağlar.

Birliktelik kuralı madenciliği, tüm sık geçen öğelerin bulunması ve sık geçen bu öğelerden güçlü birliktelik kurallarının üretilmesi olmak üzere iki aşamalıdır.

Birliktelik kuralının ilk aşaması için kullanılan Apriori Algoritması, sık geçen öğeler madenciliğinde kullanılan en popüler ve klasik algoritmadır. Bu algoritmada özellikler ve veri, Boolean ilişki kuralları ile değerlendirilir (Gao, 2004). Birliktelik kuralları çıkarmak için en çok kullanılan algoritmadır.

Gereksinimler

Apriori algoritması kullanılacak;

Veri seti tabular ve transaction yapıda olmalı,

Veri seti kategorik yapıda olmalı,

Giriş (input) ve çıkış (output) alanları sembolik olmalı,

Veri setindeki değişkenler yönleri (direction) in, out ya da both olarak tanımlanmalıdır.

Apriori Artıları

- Büyük problemlerde GRI algoritmasına göre daha hızlıdır.
- Apriori algoritmasında kural sayısında limit yoktur.
- Apriori Algoritması esnek bir yapıya sahiptir ve eldeki probleme uygun olarak 5 farklı eğitim yöntemi (different training) bulunmaktadır.
- Apriori algoritması ile model tahmin oluşturmak için veri akışına sokulabilir.

2.2 Tanımlamalar

Item Set(Nitelik Kümesi): Bir veya birden fazla elemanın bir koleksiyonudur. Verilen d adet eleman için 2^d adet olası aday itemset vardır.

Destek (Support): Bir ilişkinin tüm alışverişler içinde hangi oranda tekrarlandığını belirtir.

Güven (Confidence): X ürün grubunu alan müşterilerin Y ürün grubunu alma olasılığını belirtir.

Örneğin X ve Y ürünleri için destek = %2 ve güven = %60 olsun;

Destek değeri; analiz edilen tüm alışverişlerin %2’sinde X ile Y ürünlerinin birlikte satıldığını,

Güven değeri; X ürününü satın alan müşterilerin %60'ının aynı alışverişte Y ürününü de satın aldığını gösterir.

2.3 Algoritma Adımları

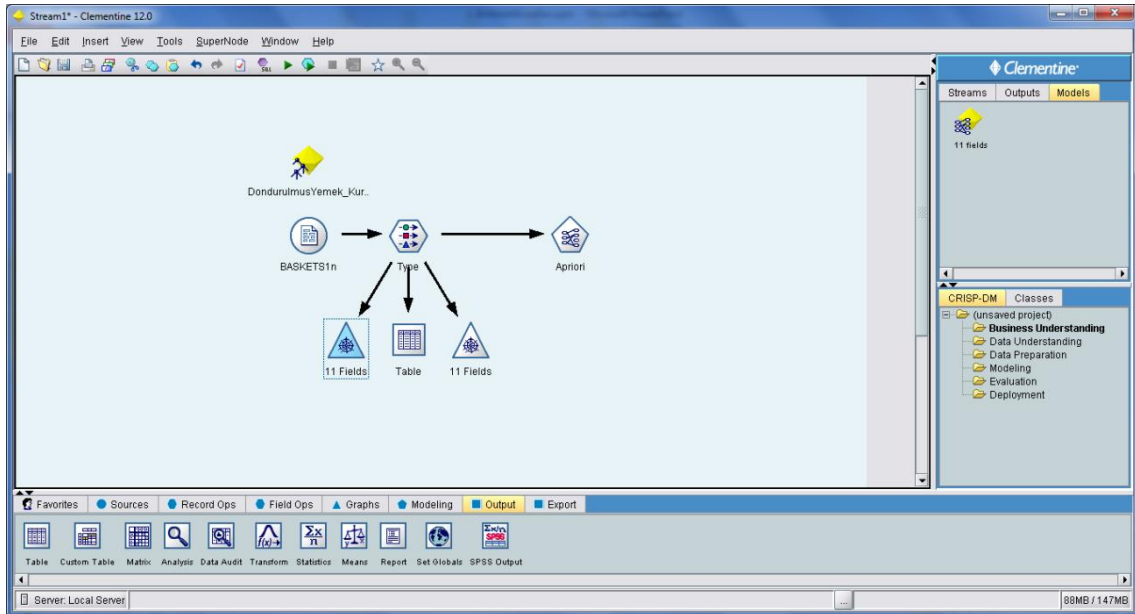
1. Minimum destek değeri ve minimum güven değerinin yani eşik değerlerinin belirlenmesi
2. Öğe kümeler içerisindeki her bir öğenin destek değerinin bulunması
3. Minimum destek değerinden düşük olan öğe kümelerin çıkartılması
4. Elde edilen tekli birliktelikler dikkate alınarak ikili birlikteliklerin oluşturulması
5. Minimum destek değerinden düşük olan öğe kümelerin çıkartılması
6. Elde edilen ikili birliktelikler dikkate alınarak üçlü birlikteliklerin oluşturulması
7. Üçlü birlikteliklerden minimum destek değerinden düşük olan öğe kümelerin çıkarılması
8. Üçlü birlikteliklerden birliktelik kurallarının çıkarılması

2.4 Uygulama

Veri Seti: baskets1n

- ✓ 11 değişken {MeyveSebze, Taze Et, Süt Ürünleri, KonserveSebze, KonserveEt, Dondurulmuş Yemek, Bira, Şarap, İçecek, Balık, Şekerleme}
- ✓ 1000 gözlemden oluşmaktadır.

Stream Yapısı

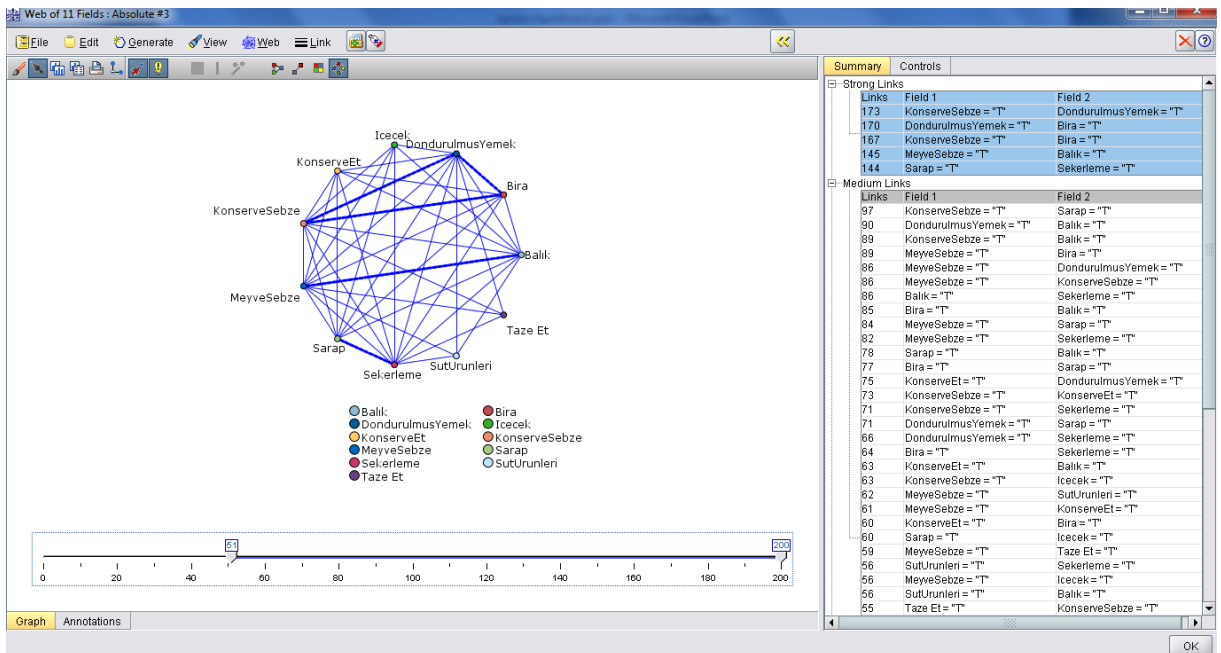


Type Nodu

Field	Type	Values	Missing	Check	Direction
MeyveSebze	Flag	T/F		None	Both
Taze Et	Flag	T/F		None	Both
SutUrunleri	Flag	T/F		None	Both
KonserveSebze	Flag	T/F		None	Both
KonserveEt	Flag	T/F		None	Both
Dondurulmus...	Flag	T/F		None	Both
Bira	Flag	T/F		None	Both
Sarap	Flag	T/F		None	Both
Icecek	Flag	T/F		None	Both
Balik	Flag	T/F		None	Both
Sekerleme	Flag	T/F		None	Both

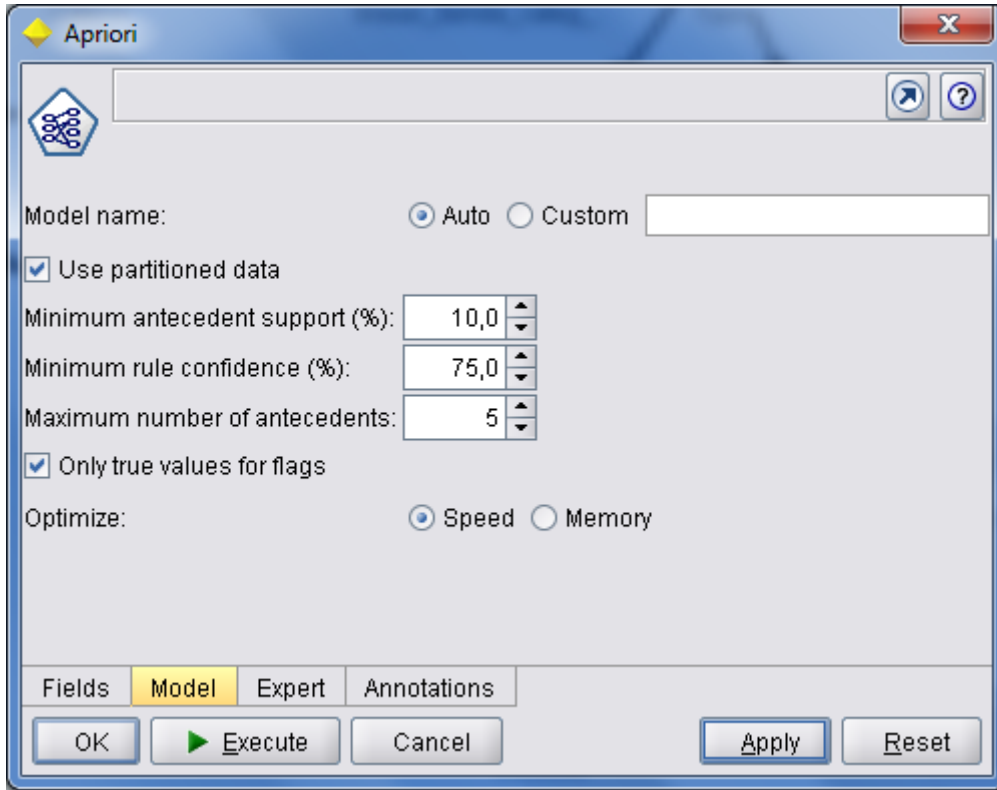
Burada, 11 değişkene ait tanımlamalar, Type(Tip) kısmı iki oldukları için Flag, Direction(Yön) ise birliktelik kuralının hem sağ hem de sol tarafında olabilmesi için Both olarak seçilmiştir.

Web Nodu



Yorum: Birlikte sıklıkla satın alınan ürünler Konserve Sebze, Dondurulmuş Yemek, Meyve Sebze, Şarap, Şekerleme, Bira olduğu görülmektedir. Görsel olarak elde edilen ilişkiler, Birliktelik Kuralları analizi ile kural seti olarak ifade edilebilir.

Apriori Nodu



Apriori

Model name: ☒ Auto ☐ Custom

☒ Use partitioned data

Minimum antecedent support (%):

Minimum rule confidence (%):

Maximum number of antecedents:

☒ Only true values for flags

Optimize: ☒ Speed ☐ Memory

Fields Model Expert Annotations

OK Execute Cancel Apply Reset

Apriori Algoritmasının Eşik Değerleri

Minimum Önsel Destek Değeri: %10

Minimum Kural Güven Değeri: %75 ve

Maksimum Önsel Değeri: 5 olarak belirtilmiştir.

11 fields								
File Generate								
Sort by: Confidence %								
Consequent	Antecedent	Rule ID	Instances	Support %	Confidence %	Rule Support %	Lift	Deployability
DondurulmusYemek	Bira	2	167	16,7	87,425	14,6	2,895	2,1
KonserveSebze	KonserveSebze							
KonserveSebze	Bira	1	170	17,0	85,882	14,6	2,834	2,4
DondurulmusYemek	DondurulmusYemek							
Bira	DondurulmusYemek	3	173	17,3	84,393	14,6	2,88	2,7
KonserveSebze	KonserveSebze							

Sonuç olarak 3 adet birliktelik kuralının oluştuğu görülmüştür;

1- Instances: Bira ve Dondurulmuş Yemek birlikte 170 kere görülmüştür. Veri setinde görülme sıklığıdır.

2- Support: Bira ve Dondurulmuş Yemek ürünlerinin birlikte görülme olasılığı %17 olduğu görülmüştür. $P(A, B)$ olarak formüle edilir.

3- Confidence: Bira ve Dondurulmuş Yemek ürünlerini alan bir kişinin %85,8 olasılıkla Konserve Sebze de aldığı görülmüştür. $P(B/A)$ olarak formüle edilir.

4- Rule Confidence: Bira, Dondurulmuş Yemek ve Konserve Sebze ürünlerinin bir fiş hareketinde görülme olasılığı %14,6 dır.

Rule Confidence = $\frac{P(A,B)}{N}$ olarak formüle edilir.

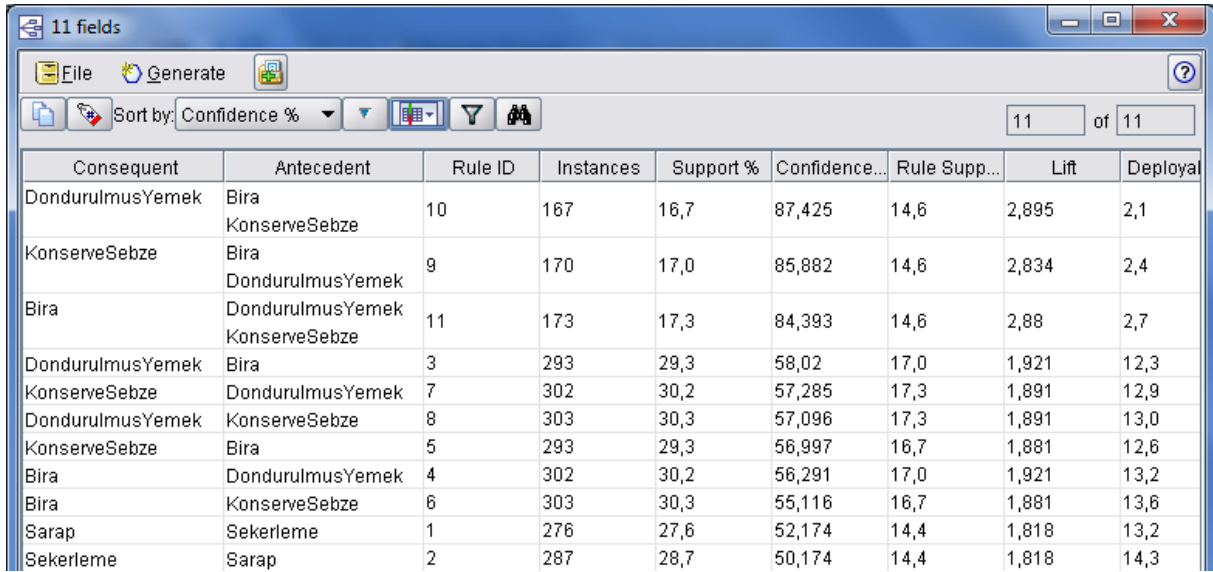
5-Lift: Lift değeri korelasyon analizinde kullanılan bir ölçüttür. Burada 2,834 değeri 1'den büyük olduğu için Antecedent(Önsel) ile Consequent (Sonsal) arasında pozitif yönlü korelasyon olduğu görülmüştür.

Lift(A,B)= $\frac{P(B/A)}{P(B)}$ olarak formüle edilir.

6-Deployability: Veri seti üzerinde öncüllerin ne kadarının karşılandığının fakat sonuculların ne kadarının karşılanmadığının ölçüsüdür. Yani müşterilerin ne kadar öncüllere sahip olduğunu ancak henüz soncul satın almadığı anlamına gelir.

Deployability = (Support – Rule Support)*100

Dondurulmuş Gıda için Kural Seti Oluşturma



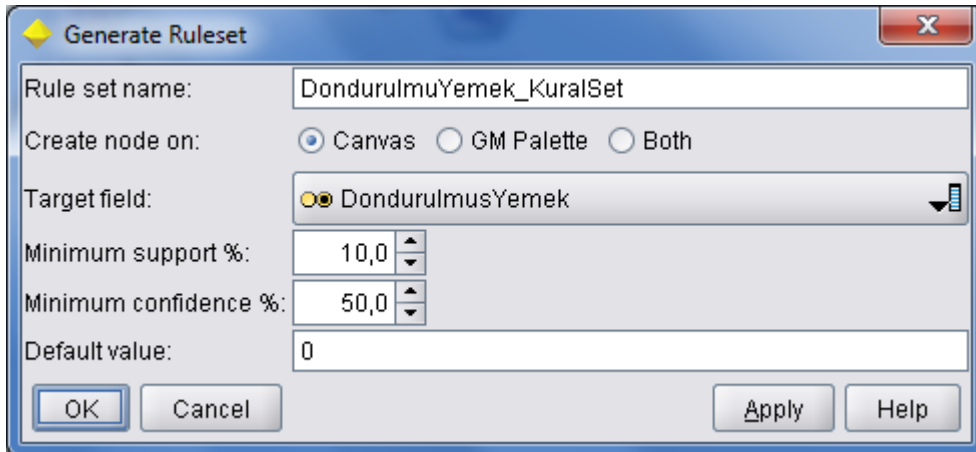
Consequent	Antecedent	Rule ID	Instances	Support %	Confidence...	Rule Supp...	Lift	Deployal
DondurulmusYemek	Bira KonserveSebze	10	167	16,7	87,425	14,6	2,895	2,1
KonserveSebze	Bira DondurulmusYemek	9	170	17,0	85,882	14,6	2,834	2,4
Bira	DondurulmusYemek KonserveSebze	11	173	17,3	84,393	14,6	2,88	2,7
DondurulmusYemek	Bira	3	293	29,3	58,02	17,0	1,921	12,3
KonserveSebze	DondurulmusYemek	7	302	30,2	57,285	17,3	1,891	12,9
DondurulmusYemek	KonserveSebze	8	303	30,3	57,096	17,3	1,891	13,0
KonserveSebze	Bira	5	293	29,3	56,997	16,7	1,881	12,6
Bira	DondurulmusYemek	4	302	30,2	56,291	17,0	1,921	13,2
Bira	KonserveSebze	6	303	30,3	55,116	16,7	1,881	13,6
Sarap	Sekerleme	1	276	27,6	52,174	14,4	1,818	13,2
Sekerleme	Sarap	2	287	28,7	50,174	14,4	1,818	14,3

Burada ise Min. Antecedent Support: 10 ve Min. Rule Confidence: 50 olan bir birliktelik kuralı görülmektedir. Bu birliktelik kuralları içinden sadece DondurulmuşYemek için bir kural seti oluşturalım.

Dondurulmuş Yemek ürünü için kural seti oluşturalım;

Minimum Support %10

Minimum Confidence %50 olarak oluşturalım.



Generate Ruleset

Rule set name: DondurulmuYemek_KuralSet

Create node on: ☒ Canvas ☐ GM Palette ☐ Both

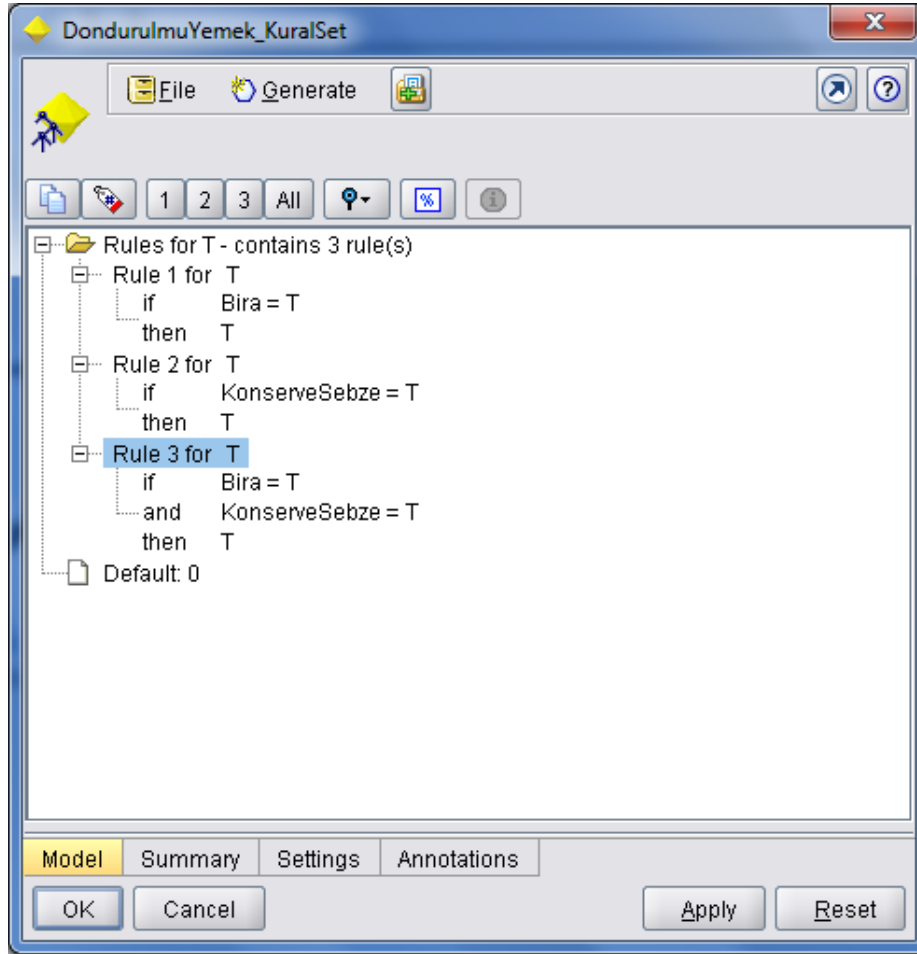
Target field: DondurulmusYemek

Minimum support %: 10,0

Minimum confidence %: 50,0

Default value: 0

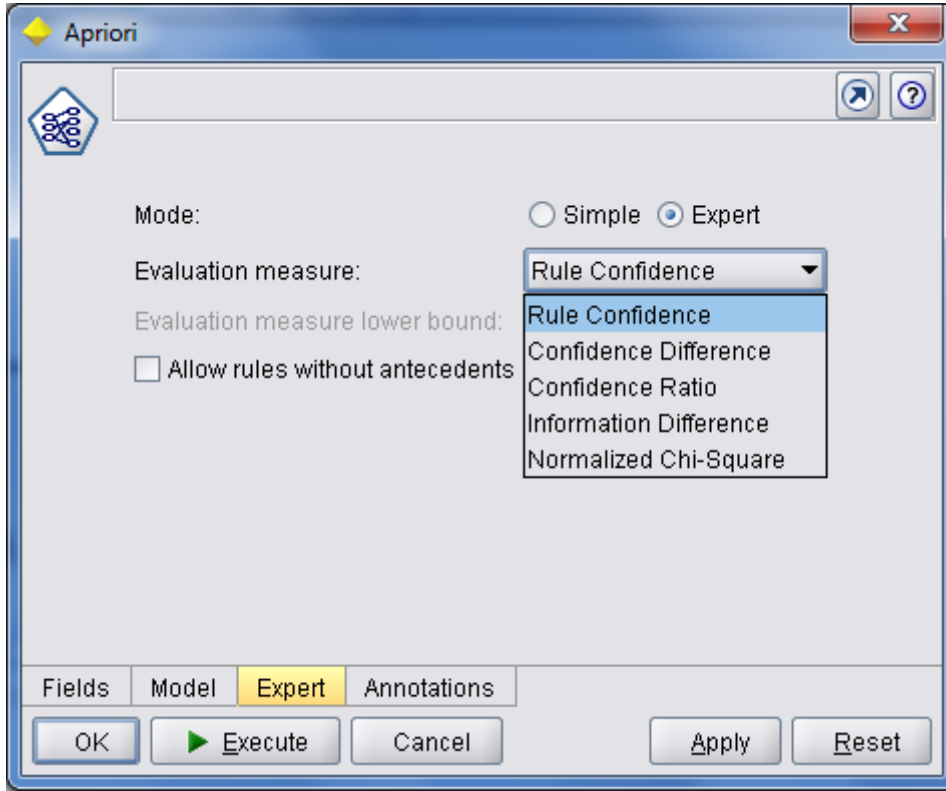
OK Cancel Apply Help



Dondurulmuş Yemek için oluşturulan kural kümesinde 3 tane kural oluştuğunu görülmüştür;

1. kural da Dondurulmuş Yemek satın alanların Bira da aldığını,
2. kural da Dondurulmuş Yemek satın alanların Konserve Sebze da aldığını,
3. kural da Dondurulmuş Yemek satın alanların Bira ve Konserve Sebze aldığı görülmüştür.

Değerlendirme Ölçütleri



1. **Rule Confidence (Kural Güveni):** Kural güveni ölçütü kuralları değerlendirmek için kuralın güvenliğini kullanılır ve varsayılan apriori değerlendirme metodudur.

Kural güveni ölçütü seçildiyse Evaluation measure lower bound spin control aktif hale gelir. Eğer kurallar sadece güven (confidence) veya doğru tahminlerden seçilmiş ise algoritma yalnızca doğru kuralları bulacaktır. Bu önerme gibi anlaşılrsa dahi doğru kuralların her zaman en enteresan ve kullanışlı olan kurallar olduğu doğru değildir. Boş bir kuralın güveni basitçe sonucun frekansıdır ve prior confidence (öncül güven) olarak adlandırılır. Bir veya birden fazla öncülü olan bir kuralın güveni posterior confidence(soncul güven) olarak adlandırılır ve oranına artırma ölçütü (lift measure) denir.

$$P = \frac{(A,B)}{N} \text{ olarak formüle edilir.}$$

2. **Confidence Difference (Güven Farkı):** Bu değerlendirme metodu Consequent(soncul) ve prior(öncül) confidence arasındaki mutlak farka dayanan çok basit bir ölçüttür. Evaluation measure lower bound (default olarak %10'a ayarlı) iki güven arasındaki en az farkı ayarlar.

Sonucu neredeyse her zaman doğru olan bir kural yalnızca öncül güveni düşük olduğunda seçilecektir. Hatta mutlak fark kullanıldığında bu ölçütü negatif kurallar dahi bulunacaktır.

$$CD = | \text{Confidence of Consequent} - \text{Confidence of Prior} | \text{ olarak formüle edilir.}$$

3. **Confidence Ratio (Güven Oranı):** İki güvenin (confidence) karşılaştırılmasında daha kolay bir yöntemdir. Bu değerlendirme ölçüsü 1 den çıkarılmış güven kuralının öncül(prior) güvene oranıdır. Güven fark ölçütü 'nün aksine bu yöntem ile oran kullanıldığından düşük güven bölgelerindeki oranlara daha duyarlıdır. Açıkçası güven oran ölçütü düşük öncül güvenleri olan kuralları ya da yüksek öncül güvenleri olan negatif kuralları seçmeye meyillidir. Güven farkı gibi bu metod düzensiz dağılımları hesaba katar. Bu özellik nadir olayları tahmin eden olayları bulmada iyi bir yoldur.

$$CR = \frac{(1 - \text{ConfidenceRules})}{\text{ConfidenceofPrior}}$$
 olarak formüle edilir.

4. **Information Difference (Bilgi Farkı):** Bu ölçü bilgi kazanım ölçüsüne dayanır. Bu bilgi farkı öncüllerin vermediği bilgi kazanımıyla sonucun önceki güveninin verildiği bilgi kazanım arasındaki farktır.

Bu değerlendirme ölçütü önceki iki ölçüte göre daha karmaşık ve bilgi kazanımı ölçütüne dayanır. Bilgi kazanımı ölçütü bilgi parçacıkları(bits) şeklinde hesaplanır. İki sebep arasındaki entropi değişimi 0 ile 1 arasında değişmesi için normalizasyonun ardından kural seçimi için kullanılır.

Değerlendirme ölçütü alt sınırı bitlerin yüzdelik dilimleri halinde yada aynı şekilde öncüller eklendiğinde entropi değişimindeki bir bitin yüzdeleri şeklinde verilir. Öyleyse değerlendirme ölçütü %50 alt sınırı bilgiyi 50 bit kadar artıran kuralın bulunmasına çalışıldığı manasına gelir.

Bilgi farkı ölçütü bir kuralın dayanağını dikkate almasının nedeni bu ölçüt altında aynı prior (öncül) ve posterior(soncul) güvenleri için kural büyük sayıdaki olaylarda geçerli olduğunda daha fazla ilgi görecektir. Bu yüzden başka algoritmaların bulduğu nadiren alakalı kuralları dikkate almaz.

$$H = - P(\text{productname}) * \log_2 P(\text{productname}) - P(\text{no productname}) * \log_2 P(\text{no productname})$$

olarak formüle edilir.

5. **Normalized Chi-square (Normalleştirilmiş Ki-Kare):** Bu ölçüt öncüller ve sonuçlar arasındaki birleşmesinin istatistiksel indeksidir. Klasik istatistikten çok iyi bilinen ki-kare testini kullanarak değişkenler arasındaki bağımlılıkları tespit eder. Bu ölçüt değerleri 0 ve 1 arasında olmak üzere normalleştirir. Ki-karenin değeri kayıt sayısının etkisini kaldırmak için normalize edilir. Bu normalizasyon ile ki-kare ölçütü 0 (ilişkisiz) ile 1 (mükemmel ilişki) arasında değerler alabilir. Değerlendirme ölçütü alt sınırı bu yüzden ilişkinin gücüne ya da ki-karenin değeridir.

$$\chi^2(R) = n \frac{(n_{01}n_{10} - n_{11}n)^2}{n_{01} (n - n_{01}) n_{10} (n - n_{10})}$$

olarak formüle edilir.

Kaynakça

Clementine ile İleri Modelleme. SPSS Clementine.

Clementine Help, Association Rules

Douglas, D. Data Mining with IBM SPSS Modeler 14.2 University of Arkansas.

Gürsoy, U. T. (2011). Uygulamalı Veri Madenciliği Sektörel Analizler. Pegem Akademi Yayıncılık.

Haider, S. Knowledge Discovery and Data Mining.

Han, J., & Kamber, M. Data Mining Concepts and Techniques. Morgan Kaufmann Publishers.

Takçı H. (,2008). Birliktelik Kuralları: Temel Kavramlar ve Algoritmalar

IBM SPSS Modeler 14.2 Modeling Nodes.

Online Association Rule Mining. <http://control.cs.berkeley.edu/carma.html#Hid98> adresinden alınmıştır

Özkan, Y. (2008). Veri Madenciliği Yöntemleri.

Şen, Z. (2004). Yapay Sinir Ağları İlkeleri. İstanbul: Su Vakfı Yayınları.

Tan, P.-N., Steinbach, M., & Kumar, V. Introduction to Data Mining. Addison-Wesley.

(Temmuz Ağustos 2013). Sosyal Güvenlik Dünyası.

Yaghini Masoud. (2009). Data Mining SPSS Clementine 12.06.AprioriAlgorithm