Fırat Üniv. Müh. Bil. Dergisi
29(2), 57-65, 2017

Science and Eng. J of Fırat Univ.
29(2), 57-65, 2017

Nicel Birliktelik Kuralları İçin Çoklu Minimum Destek Değeri

Yalçın ATEŞ, Murat KARABATAK
Fırat Üniversitesi, Teknoloji Fakültesi, Yazılım Müh. Bölümü
y_ates@hotmail.com, mkarabatak@firat.edu.tr

(Geliş/Received: 14.11.2016; Kabul/Accepted: 31.07.2017)

Özet: Veri Madenciliği, sayısal ortamlarda biriken verilerden anlamlı ve faydalı bilgiler elde etmek için son yıllarda yoğun olarak kullanılan bir çalışma alanıdır. Birliktelik Kuralı ise veriler arasındaki ilişkileri ortaya çıkarmak için kullanılan Veri Madenciliği tekniklerinden biridir. Birliktelik kuralı problemi, ilk olarak 1993 yılında Agrawal ve arkadaşları tarafından ele alınmıştır. Birliktelik kuralı uygulamalarında kullanılan destek ve güven değerleri bu yöntemin en önemli iki parametresidir. Ancak nicel değerlere sahip veri setlerinde destek değerinin belirlenmesi sorun oluşturabilmektedir. Minimum destek değerinin uygun seçilememesi, veri setindeki ilginç ve değerli bazı kuralların ya üretilememesine ya da değerli olmayan çok sayıda gereksiz kural üretilmesine neden olmaktadır. Bu çalışmada, nicel veri setleri üzerinde destek değerinin seçilmesi sorununu giderebilmek için çoklu destek değeri kullanan yeni bir yaklaşım önerilmektedir.

Anahtar Kelimeler: Birliktelik Kuralı, Minimum Destek, Çoklu Minimum Destek.

Multiple Minimum Support Value For Quantitative Association Rules

Abstract: Data Mining is a field that has been used extensively in recent years to obtain meaningful and useful information from data accumulated in digital environments. Association Rule is one of the Data Mining techniques that are used to reveal relationships between data. The association rule problem was offered in 1993 by Agrawal et al. The support and confidence values used in association rules are the two most important parameters of this method. However, determining the value of support in quantitative data sets can be problematic. Failure to select the minimum support value causes some interesting and valuable rules in the dataset to either not be produced or to generate a large number of unnecessary rules that are not valuable. In this study, a new approach that uses multiple support values to overcome the problem of selecting support values on quantitative data sets is offered.

Keywords: Association Rule, Minimum Support, Multiple Minimum Support

1. Giriş

Francis Bacon tarafından 1597'de söylenen "Scientia potentia est - Bilgi güçtür" sözü, teknoloji ile beraber günümüzde önemini oldukça arttırmıştır. Bilgi ve iletişim çağının hızla geliştiği dünyada, teknoloji hayatın her alanında kullanılmaya başlanmıştır. Bu kullanımlar sonucunda oluşan veriler çok büyük veri yığınlarına sebep olmaktadır. Veri tabanı sistemlerinin günümüzde yaygın olarak kullanılması, elde edilen verilerin olağanüstü şekilde artmasını sağlamıştır. Bu durum, insanları bu verilerden maksimum sonuçlar çıkarmak için yeni yollar keşfetmeye yöneltmiştir. Bu nedenle Veri Tabanlarında Bilgi Keşfi adı altında her geçen gün yeni arayışlar ortaya çıkmaktadır. VTBK sürecinin en önemli aşamasını, modelin kurulması ve değerlendirilmesi süreçlerini de içeren veri madenciliği oluşturmaktadır [1].

Veri Madenciliği ile bir marketin veri analizini yapılarak her ürün için bir sonraki ayın satış öngörüleri çıkarılabilmekte, müşteriler satın aldıkları ürünlere bağlı olarak gruplanabilmekte, yeni bir ürün için potansiyel müşteriler belirlenebilmekte ve müşterilerin zaman içindeki hareketleri incelenerek onların davranışları ile ilgili öngörüler yapılabilmektedir. Bu bilgilere bağlı olarak da yeni kampanya ve satış stratejileri belirlenmektedir. Binlerce ürünün ve müşterinin olabileceği düşünüldüğünde bu analizin gözle ve elle yapılamayacağı, otomatik olarak yapılmasının gerektiği ortaya çıkmaktadır. Bu aşamada veri madenciliği yöntemleri devreye girmektedir. Kısaca veri madenciliği; büyük ölçekli veriler arasından 'değeri olan' bilgileri

elde etme işi olarak tanımlanmaktadır [2]. Bu sayede, veriler arasındaki ilişkilerin ortaya çıkarılması ve gerektiği zamanda da ileriye yönelik öngörülerde bulunulması mümkün olabilmektedir.

Veri madenciliğinde bilgilerin verimli kullanılabilmesi için birçok teknik kullanılmaktadır. Bu tekniklerden biri de birliktelik kuralıdır. Birliktelik kuralı; veri tabanı içinde yer alan kayıtların birbiriyle olan ilişkilerini inceleyerek, hangi olayların eş zamanlı olarak birlikte gerçekleşebileceklerini ortaya koymaya çalışan veri madenciliği yöntemlerinden biridir. Birliktelik kuralı, nesnelerin veya niteliklerin bir arada olma durumlarını belirlemekte ve birçok alanda kullanılabilmektedir. Birliktelik kuralı bulma işlemi, yoğun nesne kümesi hesaplamaya dayalı bir işlem olup büyük veri tabanları üzerinde uygulanması oldukça pahalı bir işlemdir. Bu nedenle daha önceden tespit edilen birliktelik kurallarının korunması da oldukça önemli bir konu olmaktadır.

Genellikle büyük süpermarketlerde oluşan satış verilerine, market sepet verisi adı verilmektedir. Birçok kuruluş market sepet verilerinin önemini kullanarak bu verilerden büyük faydalar sağlamayı amaçlamaktadır. Market sepet verisi üzerinde birliktelik kuralı problemi ilk olarak 1993 yılında ele alınmıştır [3]. Sepet analizinde amaç, nitelikler (ürün satışları) arasındaki ilişkiyi bulmaktır. Bu ilişkilerin bilinmesi, bir şirketin gelecekte elde edebileceği kârını arttırmak için kullanılabilecektir. Sepet analizi günlük işlemler sonucu elde edilen verilerden anlamlı bağıntılar çıkarmada kullanılır.

Birliktelik kuralı, birçok araştırmada ele alınmış ve birliktelik kuralı üreten algoritmalar geliştirilmiştir. Agrawal ve Srikant'ın 1994 yılında yaptıkları çalışmada önerdikleri Apriori algoritması bu algoritmalar arasında en fazla bilinen ve en yaygın kullanıma sahip algoritmalardan biridir [4]. Yine aynı çalışmada, Apriori_TID isimli bir başka birliktelik kuralı algoritması da önerilmiştir. Önerilen Apriori_Hybrid algoritması, Apriori ve Apriori_TID algoritmalarının her ikisini beraber kullanan melez bir algoritmadır. Ayrıca Manila ve arkadaşları 1994'te çevrimdışı aday nesne kümesi belirleme algoritmasını, Houtsma ve Swami 1995'te SETM algoritmasını, Savasere ve arkadaşları 1995'te Partition algoritmasını, Holsheimer ve arkadaşları ise 1995'te MONET algoritmasını önermiş ve bu gibi birçok birçok algoritma literatürde yerini almıştır [5-8].

Birliktelik kuralı üreten algoritmalar sadece ürün birlikteliklerini dikkate almayıp, ürünlere ait; miktar, ağırlık ve hiyerarşik bilgi düzeni gibi özellikleri de göz önüne almaktadır. Bu kapsamda hiyerarşik birliktelik kuralı [9], sınırlandırılmış birliktelik kuralı [10], nicel birliktelik kuralı [11], sıralı örüntü keşfi [12], ağırlıklandırılmış birliktelik kuralı [13] ve negatif birliktelik kuralı [7] gibi algoritmalar geliştirilmiştir.

Birliktelik kuralı üretilecek veri setine ait niteliklerin, nicel değerlerden oluşması birliktelik kuralı algoritmaları için problem oluşturmaktadır. Bu problemleri çözmek için literatürde çeşitli çalışmalar bulunmaktadır. Chen ve arkadaşları tarafından 2016 yılında yapılan çalışmada, nitelikler için geçici bulanıklaştırma yöntemi uygulanarak nicel veriler üzerinde birliktelik kuralı uygulaması gerçekleştirmiştir [14]. Bir diğer çalışmada Gosain ve Maneela, gerçekte birçok verinin nicel değerlerden oluştuğunu dile getirmiş ve birliktelik kuralı probleminin nicel veriler üzerinde çalışmasına yönelik farklı algoritmalar önermiştir [15]. Ouyang ise 2012'de yaptığı çalışmada geleneksel birliktelik kuralında karşılaşılan problemlere değinmiş ve bu problemin çözümüne yönelik bulanık çoklu destek değer kullanan yeni bir algoritma önermiştir [16]. Yine Lee ve arkadaşları tarafından yapılan bir diğer çalışmada, çoklu minimum destek değeri uygulayan bir çalışma yapılarak bu probleme bir çözüm önerisi sunulmuştur. Ancak, yaptıkları bu çalışmada çoklu destek değerini ürünler bazında gerçekleştirmişlerdir [17].

Bu çalışmada da, birliktelik kurallarındaki minimum destek değerinin nicel veriler üzerinde oluşturmuş olduğu sorunlar irdelenmiş ve bu sorunların giderilmesine yönelik bir çözüm önerisi sunulmuştur. Elde edilen birliktelik kurallarının değerli ve ilginç olması çok önemli bir beklentidir. Çok sayıda kural üretilmesindense bu kuralların ne derece ilginç olduğu önemli bir parametre olup bu ilginçlik "*lift*" değeri ile

ölçülmektedir [18]. Bu çalışmada yine çoklu minimum destek değerlerini nitelikler bazında kullanarak, nicel birliktelik kurallarında ilginç ve değerli kuralların belirlenmesi hedeflenmektedir.

2. Veri Madenciliği

Günümüzde bilgiye ulaşmak kadar var olan koşullar dâhilinde eldeki verilerden yeni ve faydalı bilgiler üretmek de önemli bir konu haline gelmiştir. Devasa şekilde büyüyen sayısal veriler arasından yararlı ve de gerekli bilgilerin çıkarılması çabaları hızla sürmektedir. Bu aşamada, veri madenciliği kavramı göze çarpan en önemli olgulardan biridir. Frawley veri madenciliğini "*Daha önceden bilinmeyen ve potansiyel olarak yararlı olma durumuna sahip verinin keşfedilmesi*" olarak tanımlamıştır. Berry ve Linoff, 1997 yılında bu kavrama "*Anlamlı kuralların ve örüntülerin bulunması için geniş veri yığınları üzerine yapılan keşif ve analiz işlemleri*" şeklinde bir açıklama getirirken Sever ve Oğuz çalışmalarında veri madenciliği hakkında "*Önceden bilinmeyen, veri içinde gizli, anlamlı ve yararlı örüntülerin büyük ölçekli veritabanlarından otomatik biçimde elde edilmesini sağlayan veri tabanlarında bilgi keşfi süreci içerisinde bir adımdır.*" tanımını kullanmışlardır [19, 20]. Kısaca veri madenciliği; geniş veri tabanlarındaki veriler arasından bilgi çıkarma işlemi olarak tanımlanmaktadır.

2.1. Veri madenciliği modelleri

Veri madenciliği modelleri, temel olarak iki başlık altında incelenmektedir. Bunlardan birincisi, eldeki örüntülerden sonuçları bilinmeyen verilerin öngörüsü için kullanılan model, diğeri ise eldeki verinin tanımlanmasını sağlayan tanımlayıcı modeldir [1].

Öngörü yapan modellerde, sonuçları bilinen veriler kullanılarak bir model geliştirilir. Oluşturulan bu model kullanılarak sonuçları bilinmeyen veri kümeleri için sonuç değerleri ile ilgili öngörü yapılması amaçlanmaktadır. Örneğin bir banka, daha önceki dönemlerde müşterilerine verdiği tüm kredilerle ilgili bilgilere sahiptir. Bu bilgileri kullanarak daha sonraki dönemlerde müşterilere vereceği kredinin geri dönüp dönmeyeceğini müşteri bilgilerini kullanarak öngörüde bulunabilir.

Tanımlayıcı modeller ise karar vermeye rehberlik etmede kullanılabilecek mevcut verilerdeki örüntülerin tanımlanmasını sağlamaktadır. Belirli özelliklere sahip insanların bazı davranışlarının birbirine benzerlik göstermesi tanımlayıcı modele bir örnek olabilir.

Veri madenciliği modellerini gördükleri işlevlere göre ise üç ana başlık altında incelemek mümkündür [2]. Bunlar;

- Sınıflama ve Regresyon,
- Kümeleme,
- Birliktelik kuralları ve ardışık zamanlı örüntülerdir.

2.2. Birliktelik kuralı

Günümüzde, birçok alandaki veriler bulut sistemleri gibi verinin yoğun olduğu teknolojilerde, sunucu bilgisayarlarda veya kişisel bilgisayarlarda, veri tabanları üzerinde saklanmaktadır. Bu verilerden, istenilen ve kayda değer bilgilere ulaşmak için kullanılan tekniklerden biri de birliktelik kurallarıdır. Birliktelik kuralı, birçok alanda yaygın kullanım alanına sahiptir ve nesnelerin veya niteliklerin bir arada olma durumlarını belirlemede kullanılmaktadır. Birliktelik kuralı bulma işlemi, yoğun nesne kümesi hesaplamaya dayalı bir işlem olup büyük veri tabanları üzerinde uygulanması oldukça pahalı bir işlemdir. Bu nedenle daha önceden tespit edilen birliktelik kurallarının korunması da oldukça önemli bir konu olmaktadır [1].

Son zamanlarda, otomatik tanıma ve veri toplama uygulamalarındaki gelişmeler sayesinde firmaların satış noktalarında barkod sistemleri kullanımı yaygınlaşmaya başlamıştır. Bu gelişmeler ile beraber bir markete ait satış verilerinin elektronik ortamlara aktarılmasına olanak sağlanmıştır. Genellikle büyük süpermarketlerde oluşan bu tür verilere market sepet verisi adı verilmektedir. Birçok kuruluş, market sepet verilerini kullanarak bu verilerden büyük faydalar sağlamayı amaçlamaktadır.

Market sepet verisi üzerinde birliktelik kuralı problemi, ilk olarak 1993 yılında ele alınmıştır.

[3]. Sepet analizinde amaç, nitelikler (ürün satışları) arasındaki ilişkiyi bulmaktır. Bu ilişkilerin bilinmesi şirketin kârını arttırmak için kullanılabilir. Eğer A ürün veya ürün grubunu alan müşterilerin B ürün veya ürün grubunu da çok yüksek bir olasılıkla aldıkları biliniyorsa o müşterilerin potansiyel bir B müşterisi olduğu öngörülebilmektedir. Sepet analizi günlük işlemler sonucu elde edilen verilerden anlamlı bağıntılar çıkarmada kullanılır. "Eğer A ürününü alıyorlarsa %x ihtimalle B ürününü de almaya da yatkındırlar." şeklinde bir sonuç A ürününü satan bir mağaza için çok faydalı bir bilgi olabilmektedir. Sepet analizi uygulamaları; çapraz satış (cross-selling), mağaza raflarının düzenlenmesi (layout), katalog tasarımı ve fiyatlandırma (pricing) gibi alanlarda kullanılmaktadır.

Birliktelik kuralı sorgusunun matematiksel modeli Agrawal ve ark. tarafından tanımlanmıştır. Bu modelde $I = \{i_1, i_2, .., i_m\}$ ürün kodlarını; D, tüm hareketleri; T, bir hareketteki ürün kodlarını; ($T \subseteq I$), tid her harekete ait birincil anahtarı, k-nesne kümesi, k adet ürünü içeren kümeyi temsil etmektedir. X bir ürün kümesi olmak üzere; T hareketi X ürün kümesini ancak ve ancak $X \subseteq T$ şartını sağladığı zaman içermektedir. X ve Y, I ürün kümesinin bir alt kümesi ve $X \cap Y = \varnothing$ olduğu zaman, bir birliktelik kuralı $X \rightarrow Y$ biçimindeki bir bağımlılık ifadesi ile gösterilmektedir. Bu ifade ile X, Y'yi belirler veya Y, X'e bağımlıdır denir. Hareket numaraları gruplandırılarak elde edilen ürünler arasındaki bağımlılık ilişkisinin yüzde yüz doğru olması beklenemez. Benzer şekilde, çıkarsama yapılan kuralın eldeki hareketler kümesinin önemli bir kısmı tarafından desteklenmesi istenir. Bu nedenlerden dolayı, $X \rightarrow Y$ eşleştirme kuralı kullanıcı tarafından minimum değeri belirlenmiş güven (c:confidence) ve destek (s:support) eşik değerlerini sağlayacak biçimde üretilir. $X \rightarrow Y$ eşleştirme kuralına, c güven ölçütü ve s destek ölçütü eklenir ve $\Phi(D) = <X \rightarrow Y, c, s>$ ile gösterilir. Burada D, örneklemi; $X \rightarrow Y$ ifadesi birliktelik kuralını; *c* eşik değeri, ilgili kuralın minimum güvenirliğini; *s* eşik değeri ise ilgili kuralın minimum destek değerini göstermektedir. Sepet analizinde ürünler arasındaki bağıntı, destek ve güven kriterleri ile hesaplanmaktadır. Destek ve güven kriterlerinin tanımları aşağıdaki gibi yapılmaktadır.

$$\text{Destek} : P(X \rightarrow Y) = \frac{\text{X ve Y nin birlikte satın alındığı durum sayısı}}{\text{Toplam müşteri sayısı}}$$

$$\text{Güven} : P(X/Y) = \frac{P(X \wedge Y)\ \{\text{X ve Y satın alanların sayısı}\}}{P(Y)\ \{\text{Y satın alanların Sayısı}\}}$$

Destek kriteri, veri içerisinde bulunan nesneler arasındaki bağıntının ne kadar sık olduğunu, güven kriteri ise Y ürününü almış olan bir kişinin hangi olasılıkla X ürününü alacağını göstermektedir. İki ürün arasında elde edilen bağıntının önemli olabilmesi için hem destek değerinin, hem de güven kriterinin olabildiğince yüksek olması gerekmektedir. Ancak bu iki değerin de yüksek olması her zaman önemi yüksek ve ilginç kuralların elde edileceği anlamı taşımamaktadır. Bu nedenle, bir kuralın ne derece ilginç olduğunun tespitine yönelik olarak *lift* değeri kullanılmaktadır [18]. Lift değeri;

$$\mathbf{Lift(X \rightarrow Y)} = \frac{\mathbf{destek(X \wedge Y)}}{\mathbf{destek(X).destek(Y)}} = \frac{\mathbf{P(Y \backslash X)}}{\mathbf{P(Y)}}$$

denklemi ile hesaplanmaktadır. Lift ölçütünün "1" değerin alması ilginçliğin olmadığını 1'den büyük veya küçük değerler alması ise ilginçliğin arttığını göstermektedir.

2.3. Nicel birliktelik kuralları

Orijinal birliktelik kuralı problemi, niteliklerin var/yok (boolean) değeri ile

ilgilenmektedir. Ancak niteliklerin aldığı değerler birçok problemde ikiden fazla olabilmektedir. Bu durumlarda, niteliklerin sayısal veya kategorik değerleri ile kural üretmek için Srikant ve Agrawal tarafından 1996'da etkili bir algoritma önerilmektedir [11]. Örneğin, Yaş:30..39 ve Evli: Evet → Arabaların Sayısı: 2 (%40, 90%) şeklinde verilen bir kural nicel birliktelik kuralı olarak değerlendirilmektedir. Bu kural, yaşı 30 ile 39 arasında olan ve evli olan insanların %90'nın iki araba sahibi olduğu anlamına gelmekte ve veri tabanı üzerindeki destek değeri %40'tır. Tablo 1'de Nitel değerlere sahip bir veri seti görülmektedir.

Tablo 1. Nicel/nitel değerler içeren örnek veri seti

ID	Cinsiyet	Gelir	Yaş
1	Bayan	Düşük	yaş <18
2	Erkek	Orta	18<yaş<25
3	Erkek	Orta	25<yaş<35
4	Erkek	Düşük	Yaş>35
5	Erkek	Yüksek	Yaş>35

3. Yöntem

3.1. Problemin tanımı

Birliktelik kuralı, minimum destek değerini parametre olarak kullanmakta ve veri setini tarayarak bu değeri sağlamayan nitelikleri her adımda elemektedir. Veri setine ait tüm niteliklerin aldığı değerler eşit aralıkta olduğunda sabit bir minimum destek değeri kullanılması sorun teşkil etmemektedir. Ancak nicel veya nitel değerler alan ve niteliklerin aldığı değerlerin farklılıklar oluşturduğu veri setlerinde sabit minimum destek değeri uygun sonuçlar üretmemektedir. Veri setlerindeki farklı niteliklerin aynı öneme, sıklığa veya yapıya sahip olmamaları veya bazı niteliklerin sık olmasa da daha büyük öneme sahip olmaları çeşitli sorunlar ortaya çıkarabilmektedir [17].

- Minimum destek değeri yüksek olduğunda, bazı ilginç ve değerli kurallar sık tekrarlanmadığı için elde edilememekte,

- Minimum destek değeri düşük olduğunda ise elde hem kural sayısı aşırı derecede artmakta hem de elde edilen kuralların önemi ve ilginçliği azalmaktadır.

Örneğin bir veri tabanında cinsiyet niteliği sadece iki değer (bay, bayan) alabilirken, öğrenim durumu niteliği yedi farklı değer (ilk, orta, lise, ön lisans, lisans, yüksek lisans, doktora) alabilmektedir. Böyle durumlara sahip bir veri seti üzerinde kullanılacak olan sabit bir minimum destek değeri, elde edilen kuralların ilginçliğini ve değerini düşürecektir. Destek değerinin yüksek seçilmesi, öğrenim durumu ile ilgili hiçbir kuralın üretilmemesine, destek değerinin düşük verilmesi ise cinsiyet niteliğinin elde edilen tüm kurallar içerisinde yer almasına, anlamlı veya anlamsız birçok kural oluşmasına sebep olacaktır.

3.2. Önerilen yöntem

Yukarıda verilen problemin çözümü için Veri madenciliği problemlerinde yaygın olarak normalizasyon işlemi kullanılmaktadır. Ancak, Birliktelik Kuralı problemimde normalizasyon işlemi çözüm üretmemektedir. Bu nedenle, yukarıda verilen problemin çözümüne yönelik olarak, veri setindeki her nitelik için ayrı ayrı minimum destek değeri kullanan yeni bir yaklaşım önerilmiştir. Tablo 2'de, bir okuldaki öğrencilerle ait bir veri seti örneği görülmektedir. Bu tabloya göre birliktelik kuralları oluşturulmaya çalışıldığında, minimum destek değerinin seçimi nitelikler arasındaki ölçek farkından dolayı sorun oluşturabilmektedir. Seçilecek minimum destek değerinin yüksek olması durumunda, "gelir durumu" ile ilgili hiçbir kuralın elde edilememesi sorunu ortaya çıkabilmektedir. Minimum destek değerinin düşük olduğu durumlarda ise "cinsiyet" ve "internet kullanımı" gibi sadece iki farklı değer alan niteliklerin her kuralda yer alması ve ilginç olmayan değersiz kuralların elde edilmesi durumu ortaya çıkmaktadır. Çünkü veri setinde "internet", "cinsiyet" ve "başarı durumu" nitelikleri sadece iki farklı değer alırken "yaşam" niteliği dört, "gelir durumu" niteliği ise üç farklı değer almaktadır. Nitelikler arasındaki bu ölçek farklılığı, nicel veya nitel birliktelik kuralları için problem oluşturmaktadır.

Tablo 2. Öğrenci yaşam ve başarı durumu

ID	Cinsiyet	Gelir	Internet	Yaşam	Başarı
1	Bayan	Düşük	Yok	Yurt	Başarısız
2	Erkek	Orta	Var	Akraba	Başarılı
3	Erkek	Orta	Var	Arkadaş	Başarısız
4	Erkek	Düşük	Var	Yurt	Başarısız
5	Erkek	Yüksek	Yok	Arkadaş	Başarılı
6	Bayan	Yüksek	Yok	Akraba	Başarısız
7	Bayan	Orta	Yok	Aile	Başarılı
8	Bayan	Düşük	Var	Yurt	Başarısız
9	Erkek	Düşük	Var	Yurt	Başarılı
10	Bayan	Yüksek	Var	Arkadaş	Başarısız

Bu problemin çözümü için, algoritmaya girilen minimum destek değeri parametresini her nitelik üzerinde farklı olacak şekilde kullanan çoklu destek değeri kullanılması önerilmektedir. I = $\{I_1, I_2, \ldots I_n\}$ veri setine ait nitelikler olmak üzere her bir niteliğe ait minimum destek sayısı Denklem 1'deki formül ile hesaplanması önerilmektedir.

$$Minsup(i) = \frac{|T|.s.2}{100.\varphi(i)} \qquad (1)$$

formülde,

Minsup(i)= *i* niteliği için minimum destek sayısı,
|T| = Kayıt sayısı,
s = % olarak destek değeri,
φ(i) = *i* niteliğin aldığı farklı değer sayısı olarak verilmektedir.

Örneğin Tablo 2'de verilen veri seti için "yaşam" niteliğine ait destek sayısı;

$$Minsup(yaşam) = \frac{10.40.2}{100.4} = 2$$

olarak hesaplanmaktadır.

Tablo 3'te her bir niteliğin aldığı değerler ve her bir nitelik için kullanılacak minimum destek sayısı verilmektedir. Ayrıca veri setine dönüşüm uygulanarak, Apriori algoritması uygulanabilecek mantıksal formata dönüştürülmüştür. Bu şekilde her bir niteliğin aldığı farklı değerler o niteliğin bir alt niteliğini oluşturmakta ve alt niteliklere de o niteliğe ait minimum destek değeri uygulanmaktadır.

Tablo 3. Nitelikler ve %40 destek değeri için her bir niteliğin destek sayıları (d).

ID	Cinsiyet d=4		Gelir d=2			Internet d=4		Yaşam d=2				Başarı d=4	
	Bayan	Erkek	Düşük	Orta	Yüksek	Yok	Var	Aile	Arkadaş	Akraba	Yurt	Başarısız	Başarılı
	A	B	C	D	E	F	G	H	I	J	K	L	M
1	1	0	1	0	0	1	0	0	0	0	1	1	0
2	0	1	0	1	0	0	1	0	0	1	0	0	1
3	0	1	0	1	0	0	1	0	1	0	0	1	0
4	0	1	1	0	0	0	1	0	0	0	1	1	0
5	0	1	0	0	1	1	0	0	1	0	0	0	1
6	1	0	0	0	1	1	0	0	0	1	0	1	0
7	1	0	0	1	0	1	0	1	0	0	0	0	1
8	1	0	1	0	0	0	1	0	0	0	1	1	0
9	0	1	1	0	0	0	1	0	0	0	1	0	1
10	1	0	0	0	1	0	1	0	1	0	0	1	0

3.3. Kuralların oluşturulması

Tablo 3'te elde edilen 10x13 boyutundaki veri setine, her niteliğe ayrı destek değeri uygulanacak şekilde Apriori algoritması uygulanmaktadır. Algoritmanın birinci adımında, her niteliğin ayrı ayrı destek değerleri hesaplanmakta ve minimum destek değerini sağlamayan nitelikler veri setinden elenmektedir. Tablo 4'te, her bir alt niteliğe ait destek değerleri ile o niteliğin minimum destek değerleri karşılaştırılmalı olarak verilmektedir. Burada H alt niteliği minimum destek değerini sağlamadığı için aday nesne kümesinden elenmektedir.

Tablo 4. C1-Bir elemanlı aday nesne kümesi

Nitelik	A	B	C	D	E	F	G	H	I	J	K	L	M
Destek Sayısı	5	5	4	3	3	4	6	1	3	2	4	6	4
Minimum Destek	4	4	2	2	2	4	4	2	2	2	2	4	4

Apriori algoritmasının ikinci adımında H alt niteliği elendikten sonra kalan alt niteliklerin tüm ikili birliktelikleri üretilmektedir. Daha sonra, oluşturulan bu ikili birliktelikler in destek değerleri hesaplanmaktadır. Bu destek değerleri daha sonra minimum destek değerleri karşılaştırıldığında, minimum destek değerini sağlamayan ikili birliktelikler aday kümeden çıkarılır. İkili ve daha sonra elde edilecek birliktelikler minimum destek değeri ile karşılaştırılırken alt niteliklere ait birden fazla minimum destek değeri elde edildiği için bu destek değerlerinden küçük olan değer minimum destek değeri olarak kabul edilmektedir. Örneğin ikili birliktelikte yer alan (A-E) ikili birlikteliği için, "A" alt niteliğe ait minimum destek değeri 4 ve "E" alt niteliğine ait minimum destek değeri 2 olduğu görülmektedir. Bu iki niteliğin birlikteliğinde minimum destek değeri olarak "E" niteliğin destek değeri daha küçük olduğu için 2 kabul edilerek algoritmaya yürütülmektedir. Minimum destek değerinin altında kalan alt nitelikler elenerek bu şekilde algoritma sürdürülmektedir.

Tablo 5. C2-İki elemanlı aday nesne kümesi

Nitelik	Destek Sayısı	Min. Destek	Nitelik	Destek Sayısı	Min. Destek	Nitelik	Destek Sayısı	Min. Destek
A-B	*0*	*4*	C-E	0	2	E-M	1	2
A-C	**2**	**2**	C-F	1	2	F-H	0	4
A-D	1	2	**C-G**	**3**	**2**	F-I	1	2
A-E	**2**	**2**	C-I	0	2	F-J	1	2
A-F	3	4	C-J	0	2	F-K	1	2
A-G	2	4	**C-K**	**4**	**2**	F-L	2	4
A-I	1	2	**C-L**	**3**	**2**	F-M	2	4
A-J	1	2	C-M	1	2	**G-I**	**2**	**2**
A-K	**2**	**2**	D-E	0	2	G-J	1	2
A-L	**4**	**4**	D-F	1	2	G-K	3	2
A-M	1	4	**D-G**	**2**	**2**	**G-L**	**4**	**4**
B-C	**2**	**2**	D-I	1	2	G-M	2	4
B-D	**2**	**2**	D-J	1	2	I-J	0	2
B-E	1	2	D-K	0	2	I-K	0	2
B-E	1	4	D-L	1	2	**I-L**	**2**	**2**
B-G	**4**	**4**	**D-M**	**2**	**2**	I-M	1	2
B-I	**2**	**2**	**E-F**	**2**	**2**	J-K	0	2
B-J	1	2	E-G	1	2	J-L	1	2
B-K	**2**	**2**	**E-I**	**2**	**2**	J-M	1	2
B-L	2	4	E-J	1	2	K-L	**3**	**2**
B-M	3	4	E-K	0	2	K-M	1	2
C-D	0	2	**E-L**	**2**	**2**	M-L	0	4

Apriori algoritması bu şekilde yürütüldüğünde Tablo 6'de görülen 4 elemanlı yoğun nesne kümeleri birliktelik kuralı üretmek için elde edilen son yoğun küme olarak elde

edilmektedir. Tabloda görüldüğü üzere 4 elemanlı 3 farlı yoğun nesne kümesi elde edilebilmiştir. Bu durum daha fazla sayıda güven değeri yüksek ve lift değeri uygun kurallar üretilebilmesine olanak sağlamaktadır.

Tablo 6. L4 Dört elemanlı yoğun nesne kümesi

Nitelik	Destek Sayısı	Min. Destek
A-C-K-L	2	2
B-C-G-K	2	2
C-G-K-L	2	2

Tablo 7'de tekli destek değeri ile çoklu destek değeri kullanımı sonucunda elde edilen yoğun nesne kümeleri karşılaştırmalı olarak verilmektedir. Algoritma %40 minimum destek değeri ile işletildiğinde, önerilen yöntem Denklem 1'e göre her bir niteliğe ait destek sayısını ayrı ayrı hesaplamakta ve algoritma yürütülmektedir. Tablo 7'de harf kodları yerine veri setine ait gerçek değerler verilmektedir. Tablodan görüldüğü üzere destek değeri %40 olmasına rağmen destek sayısı 2 olan bazı nitelik değerleri de yoğun nesne kümelerine dâhil edilmiştir.

Tablo 7. Tekli ve çoklu minimum destek değerli yoğun nesne kümelerinin karşılaştırılması

Tekli Destek (%40)		Çoklu Destek (%40)	
Bayan-Başarısız	: Destek=4	Bayan-Gelir Düşük-Yurt-Başarısız	: Destek=2
Erkek-Int. Var	: Destek=4	Erkek-Gelir Düşük-Int. Var-Yurt	: Destek=2
Gelir Düşük-Yurt	: Destek=4	Gelir Düşük-İnt. Var-Yurt-Başarısız	: Destek=2
İnt. Var-Başarısız	: Destek=4		

Tablo 7'den görüldüğü üzere çoklu destek değeri ile önerilen algoritma, verilen veri seti üzerinde 4 elemanlı yoğun nesne kümeleri üretebilirken klasik birliktelik kuralı algoritması en fazla 2 elemanlı yoğun nesne kümeleri üretebilmiştir. Bu durum, nicel birliktelik kurallarında çoklu destek değeri kullanılarak ilginç ve değerli kuralların üretilebilmesinin mümkün olduğunu göstermektedir. Tablo 7'de verilen yoğun nesne kümeleri için %80 güven değeri ile kurallar elde edildiğinde her iki yöntem için de elde edilen kurallar Tablo 8'de görülmektedir. Çoklu destek değeri ile elde edilen kural sayısı fazla olup bu kurallardan sadece 4 tanesi Tablo 8'de örnek olarak verilmektedir.

Tablo 8. Tekli ve çoklu minimum destek değeri ile elde edilen kurallar

Tekli Destek (%40) Tüm Kurallar		Çoklu Destek (%40) Örnek Kurallar	
Bayan→Başarısız	: Güven= %80	Düşük-Başarısız→ Yurt	: Güven=%100 lift= 2.50
Erkek→İnt. Var	: Güven= %80	Bayan-Düşük-Yurt → Başarısız	: Güven=%100 lift= 2.00
Gelir Düşük→Yurt	: Güven=%100	Erkek-Düşük-Yurt→ Int. Var	: Güven=%100 lift= 1.67
Yurt→Gelir Düşük	: Güven=%100	Bayan-Yurt→Düşük-Başarısız	: Güven=%100 lift= 3.33
		…	

Tablo 8 incelendiğinde, tekli destek değeri ile minimum güven değerini sağlayan sadece 4 kuralın elde edildiği görülmektedir. Oysa çoklu destek değeri kullanıldığında hem güven değeri hem de ilginçliği yüksek çok güçlü kuralların elde edildiği görülmektedir. Bu kuralların klasik birliktelik kurallarından elde edilemediği tabloda açıkça görülmektedir. Bu kuralların elde edilebilmesi, klasik birliktelik kuralında destek değerinin %40 yerine %20 seçilmesi ile ancak mümkün olabilecektir. Bu durumda, destek değeri düşük olan çok sayıda ilginç ve değerli olmayan kural üretilecektir.

4. Sonuç

Bu çalışmada, market sepet analizi olarak bilinen birliktelik kuralı algoritmasının nicel veri setleri üzerinde kullanılan türü için yeni bir yaklaşım önerilmiştir. Birliktelik kuralı algoritmasındaki en önemli parametre olan minimum destek değerinin nicel birliktelik kurallarında uygulanması ile ortaya çıkan soruna

çözüm olarak çoklu destek değeri kullanılmıştır. Nicel birliktelik kuralı üretilen veri seti üzerinde her bir niteliğin aldığı değerler farklı ölçekte olduğunda tek bir destek değerinin kullanılması sorun oluşturmaktadır. Minimum destek değerinin yüksek olması veya düşük olması durumunda nitelikler arasındaki birliktelikler ya hiç üretilmemekte ya da değersiz çok sayıda kural ile karşı karşıya kalınmaktadır.

Çoklu destek değeri önerilen bu çalışmada tek destek değeri sorununa çözüm üretmek amacı ile her bir niteliğin aldığı değerlere bağlı olarak farklı minimum destek değerlerinin uygulanması ile bu sorunun çözüldüğü görülmektedir. Böylece birliktelik kuralı algoritmasına giriş olarak tek bir minimum destek değeri girilmekte ancak algoritma her bir niteliğe bu değeri farklı uygulayarak yoğun nesne kümelerini hesaplamaktadır. Bu sayede ölçek farklılığı olan nitelikler arasında ilginçliği yüksek birliktelik kuralların elde edilmesi mümkün kılınmaktadır.

5. Kaynaklar

1. Akpınar, H., (2000), Veri tabanlarında bilgi keşfi ve veri madenciliği, İ.Ü. İşletme Fakültesi Dergisi, **29**, 1-22.

2. Özkan, Y., (2008), Veri Madenciliği Yöntemleri, Papatya Yayıncılık, İstanbul.

3. Agrawal, R., Imielinski, T., Swami, A., (1993) Mining association rules between sets of items in large databases, In ACM SIGMOD Conf. Management of Data, **22**, 207-216.

4. Agrawal, R. and Srikant, R., (1994), Fast algorithms for mining association rules, Proceeding of the 20th Int. Conference on Very Large Database, VLDB, **1215**, 1994.

5. Manila, H., Toivonen, H., Verkamo, A.I., (1994) Efficient algorithms for discovering association rules. In Proceedings of AAAI'94 Workshop on Knowledge Discovery in Databases (KDD'94), 181-192, Seattle, Washington, USA.

6. Houtsma, M., Swami, A., (1995), Set-Oriented mining for association rules in relational databases, Proceedings of the 11th IEEE International Conference on DataEngineering, 25-34, Taipei, Taiwan,

7. Savasere, A., Omiecinski, E., Navathe, S.B., (1998), Mining for strong negative associations in a large database of customer transaction, In Proceedings of the 14nd International Conference on Data Engineering, 494-502, Orlando, Florida, USA.

8. Holsheimer, M., Kertsen, M., Manila, H., Toivonen, H., (1995), A perspective on databases and data mining, Proceeding of 1st International Conference on Knowledge and Data Mining, 150-155, Montreal, Kanada.

9. Han, J., Fu, Y., (1995), Discovery of multiple-level association rules from large Databases, Proceedings of the 21nd International Conference on Very Large Databases, 420-431, Zurich, İsviçre.

10. Srikant, R., Vu, Q., Agraval, R., (1997), Mining association rules with item constraints, In Proceeding of 3rd International Conference on Knowledge Discovery and Data Mining, California, USA.

11. Srikant, R., Agrawal, R., (1996), Mining quantitative association rules in large relational tables, In Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data, 1-12, Montreal, Quebec, Kanada.

12. Agrawal, R., Srikant, R., (1995), Mining sequential patterns, In Proceedings of the 11th IEEE International Conference on Data Engineering, Taipei, Taiwan, IEEE Computer Society Press.

13. Cai, C.H., Fu, A.W.C., Cheng, C.H., Kwong, W.W., (1998), Mining association rules with weighted items, In Proceedings of 1998 International Database Engineering and Applications Symposium, 68-77, Cardiff, Wales.

14. Chen, Chun-Hao, et al., (2016), Mining fuzzy temporal association rules by item lifespans. Applied Soft Computing, **41**, 265-274.

15. Gosain, A., Maneela B., (2013), A comprehensive survey of association rules on quantitative data in data mining, Information & Communication Technologies (ICT), IEEE Conference on.

16. Ouyang, W., (2012), Mining positive and negative fuzzy association rules with multiple minimum supports. In Systems and Informatics (ICSAI), 2012 International Conference on (pp. 2242-2246). IEEE.

17. Lee, Y. C., Hong, T. P., & Lin, W. Y., (2005), Mining association rules with multiple minimum supports using maximum constraints. International Journal of Approximate Reasoning, **40(1)**, 44-54.

18. Zhao, Y., Sourav S. B., (2015), Association Rule Mining with R. A Survey Nanyang Technological University, Singapore.

19. Berry, M. J., Linoff, G., (1997), Data mining techniques: for marketing, sales, and customer support. John Wiley & Sons, Inc..

20. Sever, H., Oğuz B., (2003), Veri Tabanlarında Bilgi Keşfine Formel Bir Yaklaşım Kısım I: Eşleştirme Sorguları ve Algoritmalar, Bilgi Dünyası **3(2)**, 173-204.