TÜBİTAK

DFIS- Çoklu Destek Eşiklerinde Dinamik Sık Kümeler Madenciliği ve Gizleme Platformu

Program Kodu: 3501

Proje No: 114E779

Proje Yürütücüsü:

Doç. Dr. Belgin Ergenç BOSTANOĞLU

Bursiyerler:

Ahmet Cumhur ÖZTÜRK

Sadeq DARRAB

Nourhan ABUZAYED

NİSAN 2018

İZMİR

Bu proje kapsamında, çoklu destek eşiklerine dayalı ilişki kuralları madenciliği ve duyarlı ilişki kurallarının gizlenmesi araştırmalarının yapılabildiği sınama platformunun oluşturulması hedeflenmiştir. Platformun tüm bileşenleri veri büyüklüğü, veri dinamizmi ve çoklu destek eşik değerleri gereksinimlerini dikkate almaktadır. Tüm bileşenler özgün algoritmalardan oluşmaktadır ve bu algoritmaların rakip algoritmalarla karşılaştırmaları yapılarak yüksek lisans tezleri, bilimsel dergi makaleleri ve uluslararası konferans bildirileri olarak araştırma dünyası ile paylaşılmıştır. Proje TUBİTAK ARDEB 3501 programı kapsamında desteklenmiştir.

TÜBİTAK

İÇİNDEKİLER

1. GİRİŞ 1
2. LİTERATÜR TARAMASI 4
2.1 Sık Kümeler Madenciliği 5
2.2 Dinamik Sık Kümeler Madenciliği 6
2.3 Sık Kümeler Gizlemesi 8
3. YÖNTEM 11
3.1 DFIS Sınama Platformu KısaTanıtım 11
3.2 DFIS Platform Geliştirme Süreci 14
4. DFIS PLATFORMU 20
4.1 DFIS Bileşenleri 20
4.2 DFIS Yazılım Önyüzü 25
5. SONUÇ 32
EKLER 34
KAYNAKLAR 35

ŞEKİLLER

Şekil 1. DFIS- Çoklu destek eşiklerinde dinamik sık kümeler madenciliği ve gizleme platformu 12

Şekil 2. DFIS Platformunun arka planı.....13

Şekil 3. DFIS Platformunun dinamik katmanı 14

Şekil 4. DFIS Projesi önyüzü 26

Şekil 5. Çoklu destek eşiklerinde sık kümeler madenciliği bileşeni önyüzü.....27

Şekil 6. Çoklu destek eşiklerinde sık kümeler gizleme bileşeni önyüzü 28

Şekil 7. Çoklu destek eşiklerinde dinamik sık kümeler madenciliği bileşeni önyüzü.....29

Şekil 8. Çoklu destek eşiklerinde dinamik sık kümeler gizleme bileşeni önyüzü 31

TÜBİTAK

TABLOLAR

Tablo 1. Sık kümeler madenciliği algoritmalarının karşılaştırması 6

Tablo 2. Dinamik sık kümeler madenciliği algoritmalarının karşılaştırması 8

Tablo 3. Sezgisel gizleme algoritmalarının karşılaştırması 9

Tablo 4 Proje iş paketlerinde yıllara göre gerçekleştirilenler 15

Tablo 5. DFIS Platform bileşenleri 23

Tablo 6. DFIS Platform hazır veri setleri 24

ÖZET

Veri madenciliği, büyük veri tabanlarından istatistik fonksiyonları ile çıkarılamayacak gizli örüntüleri çıkarmaya hedefler. Temel veri madenciliği görevlerinin başında ilişki kuralları madenciliği gelmektedir. İlişki kuralları madenciliği, veri tabanlarındaki ilginç ilişkileri ortaya koymayı amaçlar. İlişki kuralı bulma süreci, iki aşamadan oluşur; ilk aşamada veri içinde kullanıcı tarafından belirlenen eşik değerinden fazla tekrarlanan nesne kümeleri (”itemset”) bulunur, ikinci aşamada ise bu kümeler arası ilişki kuralları belirlenir. Büyük veri tabanlarında ilk aşama olan sık kümelerin bulunması, kayıt adedi ve nesne çeşitliliği fazla olduğu için karmaşıktır. Dolayısıyla araştırmalar, sık kümeleri bulmaya yoğunlaşır hatta çoğunlukla sık kümeler madenciliği ifadesi, ilişki kuralı madenciliği yerine kullanılır. Sık kümeler madenciliği gerçek dünya uygulamalarının birçoğunda kullanılmaktadır; tık sellerinin (“click stream”) analizinde, web linklerinin analizinde, genom analizinde, ilaç tasarımlarında, ürünlerin satışı arttıracak şekilde raflara ya da kataloglara yerleştirilmesinde, çapraz satış çalışmalarında, etkin web sitesi tasarımında, sahtekârlık tespitinde, teknik bağımlılık analizinde kullanımı akla gelen örneklerdir.

Sık kümeler madenciliği ve sık kümelerin gizlenmesi madalyonun iki yüzünde yer alan ve organizasyonların gereksinimi olan işlevlerdir. Madencilik organizasyon içinde kullanılabilecek stratejik bilgilerin çıkarılmasında, gizleme ise veri paylaşılırken veri sahiplerinin duyarlı buldukları bilgilerin gizlenmesinde kullanılmaktadır.

Bu proje kapsamında, sık kümeler madenciliğinin veri büyüklüğü, veri dinamikliği ve çoklu destek eşikleri gibi zorluklar ile uğraşan her bir işlevi özgün algoritmalardan oluşan bir sınama platformu geliştirilmiştir. Platform 4 temel işlev grubunu içermektedir; arka planda sık kümeler madenciliği ve duyarlı sık kümelerin gizlenmesi işlevleri, ön planda ise dinamik sık kümeler madenciliği ve dinamik sık kümeler gizlenmesi işlevleri bulunmaktadır. Tüm işlevler çoklu destek eşiklerini esas alır, etkin veri yapıları kullanarak veri büyüklüğü ile baş etmeye çalışır ve ayrıca veri devingenliğini dikkate alırlar. Tüm işlevler için önerilen algoritmaların başarım değerlendirmeleri kendi aralarında ve rakip algoritmalarla yapılarak uluslararası konferans bildiri kitapçıklarında bildiri ve bilimsel dergilerde makale olarak yer almıştır

ABSTRACT

The focus of the data mining is to discover unforeseen patterns which cannot be captured by statistical functions from big databases. .One of the basic tasks of data mining is association rule mining. Association rule mining targets to reveal out interesting relationships. The process of generating association rules is composed of two steps; the first step is generation of frequent patterns (“itemsets”) and second step is generation of association rules from frequent itemsets. The complexity of finding frequent itemsets is high with large number of transactions and items. For this reason the majority of the research is clustered around the generation of frequent itemsets and the terms association rule mining or itemset mining are used interchangeably. Frequent itemset mining is used in many real life situations; click-stream analysis, web-link analysis, genome analysis, drug designs, cross-sales works, efficient web site designs, intrusion detection is few examples.

Frequent itemset mining and frequent itemset hiding is two sides of the coin and the organizations need both functions. Mining is used to discover strategical knowledge from data in the organization and hiding is required to clean sensitive knowledge when the data is being shared with third parties.

The test bed that is developed within the context of this project is composed of novel algorithms that face the challenges as the size of the data, dynamicity of the data and multiple itemset support thresholds. There are four main functions in the subject test bed; frequent itemset mining and frequent itemset hiding at the background, dynamic frequent itemset mining and dynamic itemset hiding in the fore ground. All of the functions are designed for multiple itemset thresholds. Similarly all functions consider size and dynamicity of the data. The algorithms proposed for the main functions are tested with each other and other competitor algorithms anda are published in the proceedings of international conferences and journals.

1. GİRİŞ

Veri madenciliği, büyük veri tabanlarından istatistik fonksiyonları ile çıkarılamayacak gizli örüntüleri çıkarmaya hedefler. Temel veri madenciliği görevleri sınıflama, kümeleme ve ilişki kuralları madenciliğidir. Bu görevlerden ilişki kuralları madenciliği, veri tabanlarındaki ilginç ilişkileri ortaya koymayı amaçlar; bu alanın adını duyuran, en çok kullanılan ve dolayısı ile en çok araştırılanıdır. İlişki kuralı bulma süreci, iki aşamadan oluşur; ilk aşamada veri içinde "sık" tekrarlanan nesne kümeleri ("itemset") bulunur, ikinci aşamada ise bu kümeler arası ilişki kuralları belirlenir. Nesne kümesinin tüm veri tabanında kaç kere tekrarlandığının ölçütü "destek" olarak tanımlanır, nesne kümesinin destek değeri kullanıcı tarafından verilen destek eşiğini aşıyorsa "sık" olarak tanımlanır. Büyük veri tabanlarında sık kümelerin bulunması, kayıt adedi ve nesne çeşitliliği fazla olduğu için çeşitli güçlükler içerir. İkinci aşama olan sık kümelerden ilişki kuralları bulma işlemi daha basittir. Dolayısıyla araştırmalar, sık kümeleri bulmaya yoğunlaşır hatta çoğunlukla sık kümeler madenciliği ifadesi, ilişki kuralı madenciliği yerine kullanılır. Sık kümeler madenciliği gerçek dünya uygulamalarının birçoğunda kullanılmaktadır; tık sellerinin ("click stream") analizinde, web linklerinin analizinde, genom analizinde, ilaç tasarımlarında, ürünlerin satışı arttıracak şekilde raflara ya da kataloglara yerleştirilmesinde, çapraz satış çalışmalarında, etkin web sitesi tasarımında, sahtekârlık tespitinde, teknik bağımlılık analizinde kullanımı akla gelen örneklerdir.

Sık Kümeler Madenciliği algoritmalarının temel zorluğu veri büyüklüğünden kaynaklanmaktadır. Erken çalışmaların yetersizliği, Apriori algoritmasının (Agrawal vd., 1993) aday küme üreten ve bu aday kümelerin sıklığını kontrol etmek için veri tabanını çoklu tarayan yaklaşımından kaynaklanmaktadır. İzleyen çalışmalar, aday küme yaratmadan, ağaç ya da matris veri yapılarında kayıt imzalarını tutarak ve veri tabanını 2 kere tarayarak sık kümeleri bulurlar; Apriori tabanlı yaklaşımlardan daha başarılıdırlar. Ancak bu çalışmaların da gerçek dünya kullanımında eksiklikleri vardır; örneğin veri dinamizmi ile baş etme, çoklu "sık" tanımı ve veri paylaşımı durumunda organizasyon için duyarlı olan bilginin çıkarımının engellenmesi gereksinimlerini göz ardı ederler. Bu zorluklara tek tek odaklanan araştırmalar vardır; örneğin çoklu destek eşiklerinde sık kümeler madenciliği önerileri Jayasudha (2013); Kanimozhi ve Tamilarasi (2009) tarafından, dinamik sık kümeler madenciliği çözümleri Cheungt vd. (1996), Cheung ve Zaïane (2003), Taha vd. (2011), Oğuz vd. (2013) tarafından, sık kümelerin gizlemesi yöntemleri Abul (2009), Sun ve Yu (2005, 2007), Menon vd. (2005), Boora vd. (2009) ve Vaidya (2001) tarafından ele alınmıştır. Ancak bu gereksinimlere aynı işlev çinde yanıt vermeye çalışan araştırma bulunmamaktadır.

Bu proje kapsamında, sık kümeler madenciliği alanının madencilik ve gizleme işlevlerinin yapılabildiği bir platform geliştirilmesi hedeflenmiştir. Bu iki işlev madalyonun iki yüzü gibidir; madencilik işlevi organizasyonların kendileri için stratejik önem taşıyan örüntüleri bulmalarını sağlar. Gizleme işlevi ise organizasyon verisini paylaşırken gereklidir. Veriler küresel fayda sağlamak için paylaşılmak istenebilir ancak paylaşılan veride organizasyon için hassas olan örüntülerin gizlenmesi gerekir. Söz konusu platformun adı **DFIS** (Çoklu Destek Eşiklerinde Dinamik Sık Kümeler Madenciliği ve Gizleme Platformu) olarak belirlenmiştir. Platform madencilik ve gizleme işlevlerini ortak bir ortamda yapılmasına olanak verirken, platformu oluşturan tüm bileşenler sık kümeler madenciliği alanının şu 3 temel zorluğu ile baş etmeye çalışmaktadırlar; veri büyüklüğü, veri dinamizmi, sık kümelerin tanımında çoklu destek eşikleri. Platform, arka planda madencilik ve hassas veri gizleme işlevlerini, ön planda ise bu işlevlerin dinamik sürümlerini içermektedir

DFIS platformunun yenilikçi yönü, birden çok sık kümeler madenciliği zorluğuna aynı platform içinde çözüm getirmek olduğu kadar platform bileşenlerinin de daha önce yapılan araştırmalardan farklı olmasından kaynaklanmaktadır. Platform bileşenleri 4 temel işlev altında kümelenmiştir. Her bir küme aynı işlevi yapan, proje kapsamında önerilen veya rakip algoritmaları içermektedir. Birinci işlev *Çoklu Destek Eşiklerinde Sık Kümeler Madenciliği* işlevidir. Bu kümedeki algoritmalar ile organizasyonların istedikleri çoklu destek değerlerini aşan örüntüleri bulmaları sağlanır. İkinci işlev *Çoklu Destek Eşiklerinde Dinamik Sık Kümeler Madenciliğ*i işlevidir. Bu kümedeki algoritmalar ise büyüyen veride her defasında madenciliğin baştan yapılmasına gerek kalmadan, gene çoklu destek eşik değerlerinde sık kümelerin güncel tutulmasında kullanılabilir. Üçüncü işlev kümesinde *Çoklu Destek Eşiklerinde Sık Kümeler Gizlemesi* algoritmalarını içermektedir. Bu algoritmalar veri paylaşılırken kullanıcının kendisi için hassas (“sensitive/private”) olan örüntülerini gizlemesinde kullanılırlar. Dördüncü işlev *Çoklu Destek Eşiklerinde Dinamik Sık Kümeler Gizlemesi* algoritmalarından oluşmaktadır. Bu algoritmalar büyüyen veride saklamanın her defasında baştan yapılması yerine sadece artımlara odaklanarak yapılmasını sağlar.

DFIS sınama platformunun tüm bileşenleri her biri kendi başlarına yenilikçi bilimsel çalışmalar olarak doktora ve yüksek lisans tezlerinde yer almaktadırlar. Aynı şekilde her bir bileşen uluslararası saygın konferans bildirisi veya bilimsel dergi makalesi haline getirilmiştir. Bu sonuç raporunun ekleri olarak verilecek söz konusu tez, bildiri ve makale şeklindeki yayınlarda görüleceği gibi bileşenlerin güncel rakip çalışmalarla detaylı karşılaştırmaları yapılmıştır. Platform bileşenleri projenin son bölümünde entegre bir yazılım halinde birleştirilmişlerdir. Yayınlar ve internette bir servis olarak açılması planlanan yazılım ile proje çıktılarının alandaki diğer araştırma çalışmalarında kullanılması olası olacaktır.

Bu rapor şu bölümlerden oluşmaktadır. İkinci bölümde sık kümeler madenciliği ve sık kümeler gizlemesi alanlarındaki rakip çalışmalar anlatılmaktadır. Üçüncü bölümde proje kısaca tanıtılmakta ve proje süresince izlenen yol ve gerçekleştirilenler açıklanmaktadır. Dördüncü bölüm **DFIS** platformunu tüm bileşenleri ile tanıtmaya ayrılmıştır. Dördüncü bölümde aynı zamanda entegre edilen platform bileşenlerinin kullanılmasına olanak veren yazılım tanıtılmaktadır. Beşinci bölümde sonuç değerlendirmelerine yer verilmektedir.

2. LİTERATÜR TARAMASI

Günümüzde özellikle internetin yaygınlaşması ile elektronik ortamda astronomik miktarda veri bulunmaktadır ve her geçen gün bu miktar artmaktadır. Bu veriler üzerinde sorgulama araçları ile arama, özetleme veya raporlama gibi işlemler yapılabilir. Ancak veri madenciliği bize sorgulama araçları ile elde edilebilecek yanıtların ötesinde sonuçlar verir. Veri madenciliği önceden tahmin edilemeyen çıkarımların ve verilerin eğiliminin ortaya çıkarılmasını sağlar. Tanım olarak veri madenciliği büyük çaptaki verilerden bilgi çıkarımı işlemidir (Han, 2011). Ortaya çıkarılan bilgilerin pazar analizi, iş yönetimi ve karar desteği de içeren birçok alanda kullanılabilmesinden dolayı veri madenciliği bilgi endüstrisinde çok dikkat çekmektedir. Veri madenciliği işlemleri arasında öne çıkanlar sınıflandırma, kümeleme ve birliktelik kurallarının bulunmasıdır.

Birliktelik kuralları madenciliği, geniş boyuttaki veri tabanlarında veriler arasındaki ilgi çeken ilişkileri belirlemek için kullanılan en popüler metodtur ve ilk olarak Agrawal vd. (1993) tarafından tanıtılmıştır. Kısaca birliktelik kuralı, nesnelerin birlikte olmaları ya da olayların birlikte gerçekleşmesi durumudur. X=>Y olarak gösterilir, bu ifadede X ve Y nesne kümeleridir. Böyle bir kuralın anlamı: bir veritabanında (D) yer alan kayıtların (T) içinde, X' i içerenlerde Y' nin de olmasıdır. Birliktelik kural madenciliği fikri, bir sürü ürünün satın alındığı pazar sepeti verisini analiz ederken, bir müşteri X_1, X_2,.., X_n ürünlerini satın almışsa Y_1, Y_2,.., Y_n ürünlerini de %c 'lik bir oranla satın alacaktır gibi kuralları ortaya koymaya yarar (Hipp vd., 2000). Birliktelik kuralları madenciliğinde iki temel ölçü vardır, bunlar destek(s) ve güven(c)' dir. Bu ölçülerin kullanılması gereklidir çünkü üzerinde madencilik yapılan veri miktarı çok büyük olduğundan genelde birçok anlamsız kural ortaya çıkabilir. Anlamsız kurallar, kullanıcının önceden belirlediği destek ve güven eşik değerleri ile elenebilir. Bir kümenin destek değeri o kümeyi içeren harekelerin veri tabanındaki oranıdır. Güven değeri veritabanında hem X hem de Y kümelerini içeren hareket sayısının X'i içeren hareket sayısına oranıdır. Birliktelik kuralları madenciliği çözümleri iki aşama içerir; ilkinde verinin içinde sık ve birlikte destek değerinden fazla geçen sık kümeler bulunur, ikinci aşamada ise bu sık kümelerden belirlenen güven değerinin üzerinde bulunan birliktelik kuralları oluşturulur. İkinci aşama basitdir ve tüm çözümler aynı yöntemi kullanır. Dolayısıyla çözümlerin birbirinden farkı ilk aşamada kullandıkları sık kümeleri bulunması yöntemlerinden gelir. Yine bu nedenle ilişki kuralları madenciliği yerine sık kümeler madenciliğini ifadesini duyarız. İzleyen alt başlıklar altında ilkin, sık kümeler madenciliği, sonra dinamik sık kümeler madenciliği ve en son ise veri paylaşılırken ihtiyaç duyulan sık kümeler gizlemesi alanındaki araştırmalar karşılaştırmalı olarak anlatılacaktır.

2.1 Sık Kümeler Madenciliği

Büyük veri tabanlarında sık kümeler madenciliği yapabilmek için birçok çalışma yapılmıştır. Tablo 1 de sık kümeleri bulmaya odaklanmış bu çalışmalar karşılaştırılmışlardır. İkinci sütun algoritmayı algoritmik yaklaşımına göre sınıflandırmada kullanılır; algoritmanın "Birleştirme odaklı", "Ağaca dayalı" ya da "Dikey" olarak çalıştığını gösterir. Üçüncü sütun algoritmanın kullandığı temel veri yapısını göstermektedir; kıyım ağacı, Fp-Tree ya da özel ağaçlar kullanılan veri yapıları arasındadır. Dördüncü sütun algoritmanın veri yapısını dolaşım şeklini ifade etmektedir; genişlik önce, derinlik önce ya da dikey seçenekleri vardır. Son sütunda "Tek" olması madencilik algoritmasının sık kümeleri tek bir destek eşiğini dikkate alarak bulduğunu, "Çoklu" olması ise algoritmanın sık kümeleri bulurken çoklu destek eşik değerlerini dikkate alabildiğini gösterir.

Birleştirme odaklı algoritmalardan Apriori ve onun optimizasyonları grubundakiler verilen eşik değerini aşan sık kümeleri bulurken genişlik önce dolaşım şekli ile çalışırlar (Agrawal vd., 1993; Agrawal ve Imielinski, 1994; Agrawal ve Han, 2014; Nhan vd., 2010; Ghanem ve Sallam, 2011; Park vd., 1997). Her k seviyesinde bulunan sık küme, k+1 seviyesindeki aday kümeyi oluşturmada kullanılır. Her seviyede veri tabanına ulaşılarak bu adayların destek değereelrinin belirlenen eşik değerini aşıp aşmadığı belirlenir. Veri tabanına çoklu erişim ve aday-yarat-test-et yaklaşımı bu grup algoritmanın bellek ve I/O karmaşıklığını arttırır. Bir diğer grup birleştirme odaklı algoritma kümesi Matrix Apriori grubudur (Pavon vd., 2006; Yıldız ve Ergenç, 2010). Matrix Apriori kayıt imzalarını matrisde tutarak veritabanına çoklu erişim yapılması gereksinimini ortadan kaldırılır. Eclat grubundaki algoritmalar ise verinin dikey temsilini kullanarak sık kümeleri kesişim işlemleri ile bulurlar (Zaki, 2010; Thieme, 2004).

Tablonun dördüncü satırında görülen FP-growth ve onun optimizasyonları grubu bir "pattern-growth tree" (Han, 2000) kullanarak tüm kayıtların içeriklerini tutmaktadır. Bu ağaç kayıtların derinlik öncelikli dolaşımla taranmasına olanak vermektedir. Bu algoritmalarda (Grahne ve Zhu, 2005; Jalan vd., 2009), bir kümenin alt ağacı sadece küme sık ise dolaşılmaktadır. Bu yaklaşımla sık kümelerin bulunması için veri tabanına ulaşılmasına gerek kalmamaktadır. Derinlik öncelikli algoritmalar çoğunlukla genişlik öncelikli dolaşım yapan algoritmalardan daha hızlıdır. Şimdiye kadar sözü edilen algoritmaların tümü sık kümeleri tek bir eşik değerine göre bulmaktadır. Bu durumda, doğru bir eşik değeri tespit edebilmek son derece önemli hale gelmektedir çünkü eşik değerinin düşük olması bir sürü anlamsız sık küme üretilmesine yol açarken eşik değerinin yüksek olması anlamlı ve değerli sık kümelerin bulunamamasına yol açabilmektedir (Liu vd., 1999).

Tablo 1. Sık kümeler madenciliği algoritmalarının karşılaştırması

Algoritma	Algoritmik Yaklaşım	Veri Yapısı	Dolaşma Şekli	Destek Eşiği
Apriori ve onun optimizasyonları	Birleştirme odaklı	Kıyım ağacı	Genişlik önce	Tek
Matrix Apriori	Birleştirme odaklı	Matris	Genişlik önce	Tek
Eclat	Dikey	Kıyım ağacı	Dikey	Tek
FP-growth ve onun optimizasyonları	Ağaca dayalı	FP-Tree	Derinlik önce	Tek
MSapriori ve onun optimizasyonları	Birleştirme odaklı	Kıyım ağacı	Genişlik önce	Çoklu
FP-ME	Ağaca dayalı	FP-ME , Diffest	Dikey	Çoklu
CFP-growth ve onun optimizasyonları	Ağaca dayalı	FP-Tree	Derinlik önce	Çoklu

Tablo 1 in son kolonunda “Çoklu” ifadesi gördüğümüz algoritmalar sık kümeleri bulurken çoklu destek eşiklerini dikkate almaktadır. Kümenin sıklığı kümeyi oluşturan elemanların destek eşiklerine göre belirlenmektedir; küme ancak sıklığı kümeyi oluşturan elemanlardan en düşük desteklisi de verilen küme destek eşik değerini aşıyorsa “sık”tır. MSapriori ve onun optimizasyonu grubunda bulunan algoritmalar (Liu vd., 1999; Xu ve Dong, 2013) birleştirme odaklı bir yaklaşımla derinlik öncelikli dolaşım yapmaktadır ve Apriori algortmasını (Agrawal ve Srikant, 1994) temel almaktadır. FP-ME algoritması Gan vd. (2016) da önerilmiş olan ve çoklu destek eşik değerlerini dikkate alan bir diğer algoritmadır.

Çoklu destek eşiklerini dikkate alan ve Apriori tabanlı algoritmaların çoklu veri tabanı tarama gereksinimini ortadan kaldıran bir grup algoritma tablonun son satırında gördüğümüz CFP-growth ve onun optimizasyonlarıdır (Hu ve Chen, 2006; Sinthuja vd., 2011; Kiran vd., 2011). Bunlar da “pattern-growth tree” (Han, 2000) kullanmaktadırlar ve derinlik öncelikli etkin dolaşım stratejisi ile sık kümeleri bulmaktadırlar. Genişlik öncelikli dolaşım yapan algoritmalardan daha etkin olmalarına karşılık ağaç oluşturma, ağaç budama ve ağaç birleştirme gibi işlem zamanı ve bellek kullanımı açısından karmaşık bir dizi işleme gereksinim duymaktadırlar.

2.2 Dinamik Sık Kümeler Madenciliği

Dinamik sık kümeler madenciliği algoritmaları, büyüyen veriye yönelik çözümleri, farklı ilgi odakları ile oluşturmaktadır. Bu algoritmaların temel problemi veri tabanına güncellemeler

geldiğinde sık küme madenciliği işlevini baştan çalıştırmadan sadec güncellemeyi kullanarak sık kümelerin son halini belirleyebilmektir. Bu alanda şimdiye kadar geliştirilmiş olan algoritmalardan öne çıkanlar, çeşitli özellikleri ile Tablo 2 de karşılaştırılmaktadırlar. Algoritmanın tipi ilk kolonda gösterilmektedir. Algoritmalar “Apriori” algoritmasını ya da “FP-Growth” algoritmasını temel alabilir, “Border” ise sınırı takip ederek sık kümeleri güncel tutmaya çalışabilirler, “Other” ise farklı veri yapıları ile dinamik çözüm üretebilirler. 4-7 sütünlar algoritmaların güncellemelerde ekleme, silme, yeni eleman ve destek değişikliği işlevlerine sahip olup olmadığını göstermektedir. Sekizinci sütunda algoritmanın sık kümeleri üretirken aday üretip üretmediği ifade etmektedir. Son sütün algoritmanın çoklu destek eşiklerinde çalışma yeteneğini göstermektedir.

Apriori algoritmasını temel alan algoritmalar “apriori” prensibini kullanarak, yinelemeli aday-üret-test-et yaklaşımı ile sık kümeleri bulurlar. Bu yaklaşımda veri tabanının çoklu taranması gereklidir ve çalışma zamanları yüksektir. Bu algoritmaların ikinci dezavantajı çalışmaları esnasında destek değişimize izin vermeyişleridir. İkinci grup algoritma FP-Growth algoritmasını temel alırlar ve ilk gruptaki algoritmaların zayıf yönlerine ilişkin çözümleri vardır. Çoklu destek eşiklerini tüm dinamik algoritmalar içinde sadece 2 algoritma desteklemektedir; Incremental Tuning Tree (Hoque vd., 2011) ve Dynamic Matrix with MIS (Chaudhary, 2014).

Üçüncü grup dinamik algoritma “Border” tipli algoritmalardır. Sınırdaki sık kümelerin gelen güncellemelerle değişip değişmediğinin sürekli kontrol edildiği bu yaklaşımla tüm veri tabanı taranmadan sadece güncelleme dikkate alınarak sık kümelerin bulunabilmesi olasıdır. Dördüncü grup “Other” farklı veri yapılarını kullanarak hızlı bir şekilde güncellenen veri tabanlarında sık kümeleri bulmaktadırlar. Çalışmalar incellendiğinde güncellenen veri tabanlarında, çoklu destek eşiklerinde, esnek ve etkin bir yaklaşımla sık kümeleri bulabilen algoritmanın olmadığı görülmektedir.

Tablo 2. Dinamik sık kümeler madenciliği algoritmalarının karşılaştırması

Algoritma	Tip *	Ekleme	Silme	Destek Değişikliği	Yeni Eleman	Aday Üretimi	Destek
FUP (Cheung vd., 1996]	Apriori	+			+	+	Single
DELI (Lee vd., 1998)	Apriori	+	+		+	+	Single
UWEP (Ayan vd., 1999)	Apriori	+			+	+	Single
DB-tree & PotFp-tree (Ezeife ve Su, 2002)	FP-Growth	+	+		+		Single
FELINE (Cheung vd., 2003)	FP-Growth	+	+	+	+		Single
PROMISING (Amornchewin ve Kreesuradej, 2007)	Apriori	+			+	+	Single
Incremental FP-Tree, Alhajj ve Barker, 2008)	FP-Growth	+	+		+		Single
DARM (Taha vd., 2011)	Border		+			+	Single
Incremental Tuning Tree (Hoque vd., 2011)	FP-Growth	+	+	+			Multiple
IMA (Oğuz ve Ergenç, 2012)	Other	+		+	+		Single
DMA (Oğuz vd., 2013)	Other	+	+	+	+		Single
Dynamic Matrix with MIS (Chaudhary, 2014)	Other	+	+	+	+		Multiple

2.3 Sık Kümeler Gizlemesi

Sık küme gizlemesi problemi ilk kez Attalah vd. (1999) yayınında ortaya atılmıştır; bu yayında yazarlar sezgisel bir sık küme gizlemesi algoritması önermişler ve problemin NP-hard olduğunu ıspatlamışlardır. İzleyen dönemde, bu problemin çözümü için çok miktarda çalışmalar yapılmıştır; çalışmaları 4 sınıf altında inceleyebiliriz: 1) “Border Based” yaklaşımlar (Moustakides ve Verykios, 2008; Stavropoulos vd., 2016; Sun ve Yu, 2005; Sun ve Yu, 2007) sık ve sık olmayan kümeleri ayıran bir sınır olduğunu ve gizlemenin bu sınırın güncellenmesi olarak yapılabileceğini gösterirler. Bu şekilde hassas olayan sık kümelere diğer (“heuristic” ve “reconstruction based”) yaklaşımlardan daha az zarar vermelerine karşılık en uygun çözümü bulamayabilirler. 2) “Exact” yaklaşımlar (Ayav ve Ergenç, 2015; Gkoulalas-Divanis ve Verykios, 2006; Gkoulalas-Divanis ve Verykios, 2008; Gkoulalas-Divanis ve Verykios, 2009; Menon vd., 2005) problemi önce “CSP-constraint based satisfaction problem” olarak formüle ederler ve ardından “linear” programlama yaklaşımı ile çözerler. Bu yaklaşımlar en uygun çözümü bulmayı garanti ederler. En uygun çözüm veri tabanında en az bozma yapan çözümdür. Bu yaklaşımlar “linear” programlamadan dolayı yüksek hesaplama sürelerine sahiptir bu nedenle diğer yaklaşımlara göre daha az kullanılmaktadırlar. 3) Sezgisel yaklaşımlar (Keer ve Singh, 2012; Oliveria ve Zaiane, 2002;

Oliveria ve Zaiane, 2003; Verykios vd., 2004; Pontikakis vd., 2004; Wu vd., 2007) gizleme sürecinde bazı tespitlere dayanarak yan etkiyi azaltmaya odaklanırlar. Diğer yaklaşımlara göre daha fazla yan etki üretebilirler ancak pratiktirler. 4) "Reconstruction" tabanlı yaklaşımlar (Boora vd., 2009; Guo, 2007; Lin ve Liu, 2007) önce veri tabanında madencilik yaparlar, tüm sık kümeleri bulurlar, bunların arasından duyarlı kümeleri çıkarırlar, buna göre veri tabanını yeniden oluştururlar. Bu yeniden oluşturma esnasında sık olmayan kümelerin eklenmesi nedeniyle orijinal veri tabanından çok farklılaşmış bir veri tabanının üretilebilmesi mümkündür.

Tablo 3. Sezgisel gizleme algoritmalarının karşılaştırması

<table>
<tr><th>Algoritma</th><th>Gizleme</th><th>Nesne Seçimi</th><th>Kayıt Seçimi</th><th>Çakışma</th><th>Hassas Küme Eşik</th><th>Ortam</th></tr>
<tr><td>RHID (Jadav vd., 2014)</td><td>Rule</td><td>Weight</td><td>Weight</td><td rowspan="9">Overlap</td><td rowspan="3">Multiple</td><td>Incremental</td></tr>
<tr><td>SPITF (Dai ve Chiang, 2010)</td><td>Itemset</td><td>Degree</td><td>Degree</td><td rowspan="12">Static</td></tr>
<tr><td>TTBS (Kuo vd., 2008)</td><td>Itemset</td><td>Degree</td><td>Degree</td></tr>
<tr><td>SWA (Oliveria ve Zaiane, 2003)</td><td>Itemset</td><td>Support</td><td>Length</td><td rowspan="11">Single</td></tr>
<tr><td>HSARWI (Dehkordi ve Dehkordi, 2016)</td><td>Rule</td><td>Weight</td><td>Weight</td></tr>
<tr><td>FHSAR (Weng vd., 2008)</td><td>Rule</td><td>Degree</td><td>Weight</td></tr>
<tr><td>MICF (Li vd., 2007)</td><td>Itemset</td><td>Degree</td><td>Weight</td></tr>
<tr><td>Hybrid (Amiri, 2007)</td><td>Itemset</td><td>Greedy</td><td>Greedy</td></tr>
<tr><td>IGA (Oliveria ve Zaiane, 2002)</td><td>Itemset</td><td>Degree</td><td>Degree</td></tr>
<tr><td>RelevanceSorting (Cheng vd., 2016)</td><td>Rule</td><td>Support</td><td>Weight</td><td rowspan="8">Disjoint</td></tr>
<tr><td>EDSR (Norafkan vd., 2015)</td><td>Itemset</td><td>None</td><td>Length</td></tr>
<tr><td>HRR (Garg vd., 2014)</td><td>Rule</td><td>Support</td><td>All</td></tr>
<tr><td>SIF-IDF (Hong vd., 2013)</td><td>Itemset</td><td>Support</td><td>Weight</td></tr>
<tr><td>Algorithm 2.b (Verykios vd., 2004)</td><td>Itemset</td><td>Support</td><td>Length</td></tr>
<tr><td>PDA (Pontikakis vd., 2004)</td><td>Rule</td><td>Greedy</td><td>Weight</td></tr>
<tr><td>MaxFIA, MINFIA (Oliveria ve Zaiane, 2002)</td><td>Itemset</td><td>Support</td><td>Degree</td></tr>
</table>

Sık kümeler gizlemesi araştırmalarının çoğunluğu çalışma zamanlarının kısalığı ve uygulama pratiklikleri nedeniyle sezgisel yaklaşımlar etrafında kümelenmişlerdir. Table 3 de bu algoritmaları farklı özelliklerine göre karşılaştırılmış olarak görüyoruz. Gizleme sütünü algoritmanın sık küme ("itemset") ya da kural ("rule") gizlemesi yaptığını, nesne seçimi sütunu algoritmanın silinecek nesneyi belirleme stratejisini göstermektedir. "Cover" seçilecek nesnenin kaç tane hassas kümede bulunduğuna göre, "Support" seçilecek nesnenin veri

tabanındaki sıklığına göre, "Greedy" seçilecek nesneye denemelerle karar verildiğini gösterirken, "None" seçim yapılmadığını tüm kaydın silindiğini göstermektedir. Kayıt seçimi sütunu algoritmanın silme yapılacak kayda nasıl karar verdiğini göstermektedir; "Length" algoritmanın kayıt uzunluğuna göre, "Degree" içinde geçen hassas küme adedine göre, "Greedy" ise denemelerle karar verildiğini ifade etmektedir. Ortam sütunu saklamanın hangi veri tabanları için planlandığını göstermektedir; "Static" güncellemelerin dikkate alınmadığını "Incremental" ise veri tabanının güncellemelerine çözüm olduğunu göstermektedir. Hassas küme eşiği sütunu algoritmanın çoklu küme eşik değerine olanak verip vermediğini söylemektedir. Çakışma sütunu algoritmanın gizleme yapılacak hassas kümelerde ortak elemana izin verip vermediğini söylemektedir.

Tablo 3 den anlaşılacağı gibi, sık küme gizleme algoritmaları sık kural gizleme algoritmalarından fazladır. Algoritmalar silinecek nesne ve kayıtların tespitini farklı sezgilere dayanarak yapmaktadırlar. Çoğunluk çalışma ortak elemana ya da kayıt ağırlığına göre karar vermektedirler. Çoklu destek eşiğinde gizleme yapabilen algoritma sayısı çok azdır. Dinamik veri tabanlarında gizleme yapabilen sadece 2 tane algoritma bulunmaktadır; SPITF (Dai ve Chiang, 2010) ve RHID (Jadav vd., 2014).

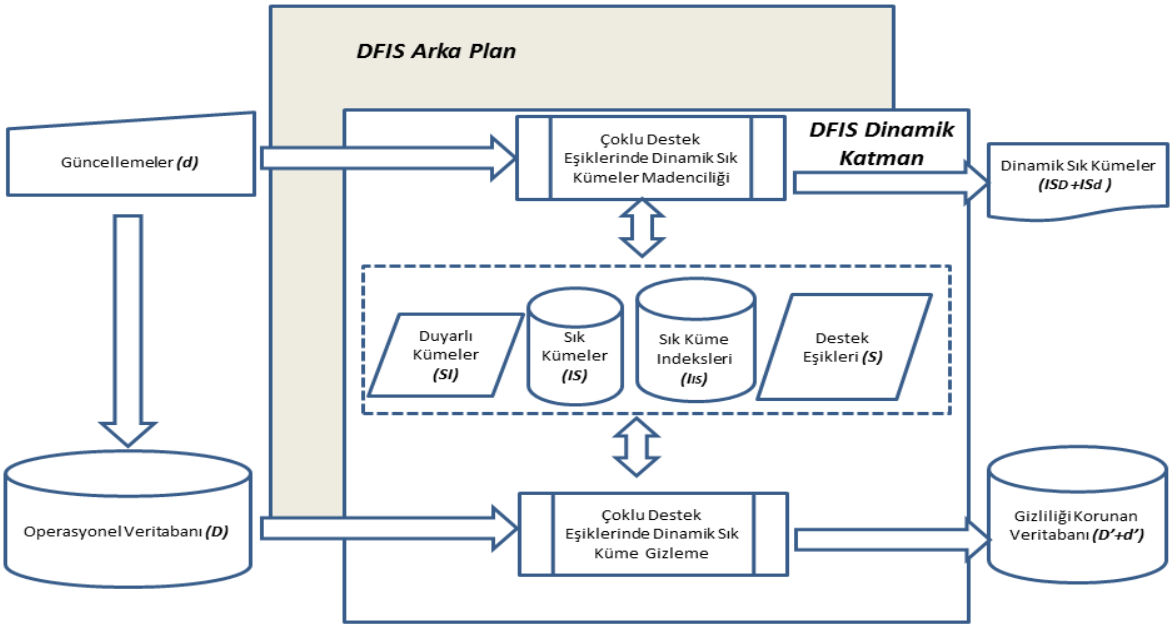
3. YÖNTEM

Projede, sürekli güncellenen veri tabanları üzerinde çalışan, çoklu destek eşiklerini dikkate alan, dinamik olarak sık kümeler madenciliği yapabilen ve duyarlı sık kümeleri yine dinamik olarak gizleyebilen bir sınama platformu oluşturulmasını amaçlanmıştır. Sık kümeler madenciliği alanının farklı zorluklarına çözüm bulmayı amaçlayan **DFIS** ("Dynamic Frequent Itemset Mining and Hiding under Multiple Support Thresholds") sınama platformu, veri tabanındaki sık kümeleri tek bir destek eşiğinde değil, kümelere özel olarak belirlenen çoklu destek eşiklerinde bulabilecek bileşenlere sahiptir. Veri tabanında güncelleme yapıldığında, güncel sık kümeler baştan bulunmayarak sadece veri tabanı güncellemesi ve eldeki sık kümeleri dikkate alarak bulunmaktadır. Odaklanılan alan güçlüklerden bir diğeri ise sık kümelerin dinamik olarak gizlenebilmesidir; veri tabanı istenildiğinde duyarlı sık kümeler (çoklu destek eşiklerinde) gizleme algoritması baştan çalıştırılmadan sadece veri tabanı güncellemesi ve eldeki sık kümeler kullanılarak gizlenebilmektedir. Bu bölümde ilk olarak **DFIS** platform katmanları kısaca tanıtılacaktır, ardından platform oluşturma çalışmaları planlanan iş paketleri kapsamında detaylı olarak anlatılacaktır.

3.1 DFIS Sınama Platformu KısaTanıtım

Şekil 1'de platformun katmanları, temel girdi ve çıktıları görülmektedir. Arka plan dinamik olmayan algoritmaları ve karşılaştırma çalışmalarında kullanılan algoritmaları içermektedir. Çoklu Destek Eşiklerinde Sık Kümeler Madenciliği yapan temel algoritma arka plandadır. Bu algoritma verilen veri tabanı ve çoklu destek eşiklerine göre sık kümeleri üretebilmektedir. Arka planda yer alan bir diğer algoritma ise Çoklu Destek Eşiklerinde Sık Küme Gizleme algoritmasıdır. Bu algoritma verilen veri tabanı, çoklu destek eşikleri ve hassas kümelere göre hassas kümelerin en az yan etki ile gizlenmesini yaparak paylaşıma hazır hale getirebilmektedir. Arka planda karşılaştırma çalışmalarında kullanılacak algoritmalar da çalıştırılabilir kılınmıştır.

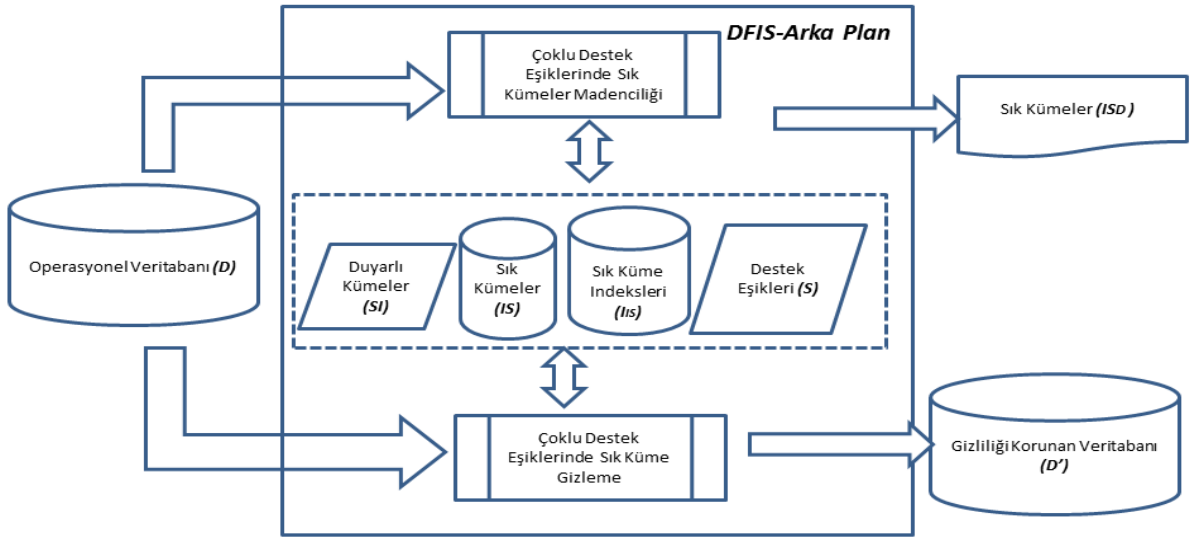
Dinamik katman ise gelen güncellemeleri ve daha önceki sonuçları kullanarak, artımlı olarak sonuç üretebilen katmandır. Bu katmanın da iki temel algoritması bulunmaktadır. İlk algoritma, Dinamik Sık Kümeler Madenciliği algoritmasıdır. Bu algoritma sürekli gelen veri tabanı güncellemelerini dikkate alarak sık kümeleri güncel tutmaktadır. Dinamik katmanın ikinci algoritması ise Çoklu Destek Eşiklerinde Dinamik Sık Kümeler Gizleme algoritmasıdır. Güncellemelere göre hassas kümeler-kayıtlar indekslerini güncel tutan bu algoritma istendiğinde gizliliği korunmuş veri tabanını üretebilmektedir.



Şekil 1. DFIS- Çoklu destek eşiklerinde dinamik sık kümeler madenciliği ve gizleme platformu

Şekil 2'de **DFIS** platformunun arka plan katmanı görülmektedir. Bu katmanın ilk algoritması, çoklu destek eşiklerinde sık kümeleri bulmayı sağlayan, Çoklu Destek Eşiklerinde Sık Kümeler algoritmasıdır. Bu algoritma girdi olarak veri tabanını (D) ve çoklu destek eşiklerini (S) alır; sık kümeleri (IS) bulur ve sık kümelerin hangi kayıtlarda bulunduğunu gösterir indeksleri (I_{IS}) oluşturur.

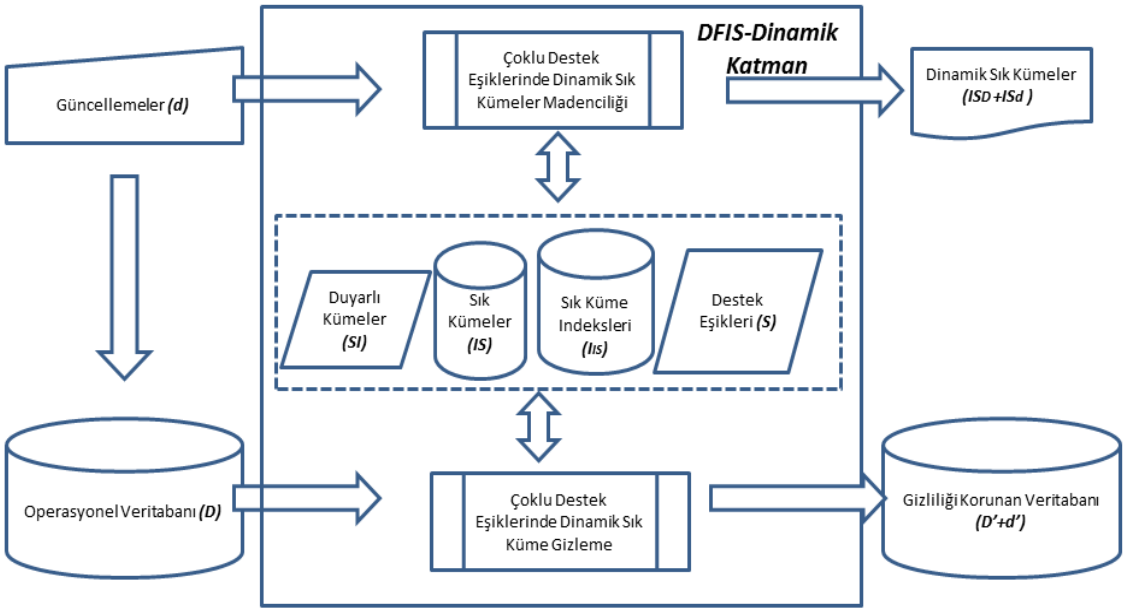
Şekil 2'de görüldüğü gibi, **DFIS** sınama platformunun arka planında bulunan bir diğer algoritma de Çoklu Destek Eşiklerinde Sık Kümeler Gizlemesi algoritmasıdır. Bu algoritma girdi olarak veri tabanını (D), sık kümeler (IS), destek eşikleri (S), duyarlı sık kümeleri (SI) alır. Duyarlı sık kümelerin gizlendiği veri tabanını (D') çıktı olarak üretir. Çalışması esnasında sık kümeler (IS) ve sık kümelerin veri tabanında bulunduğu kayıtları gösterir sık küme indeksleri (I_{IS}) güncellenir. Bunun dışında Şekil 2'de gösterilmeyen ama geliştirilen algoritmalar, başarım değerlendirme çalışmalarında kullanılan rakip araştırmalara ilişkin algoritmalardır.



Şekil 2. DFIS Platformunun arka planı

Proje kapsamında geliştirilecek **DFIS** sınama platformunun dinamik katmanı Şekil 3'de görülmektedir. Yakından bakıldığında, iki temel algoritma ve iki algoritma tarafından ortak kullanılan veri yapıları dikkat çekmektedir. İşlemlerden bir tanesi Çoklu Destek Eşiklerinde Dinamik Sık Kümeler Madenciliği yapan algoritmadır. Bu algoritma girdi olarak veritabanı güncellemelerini (d) alır; mevcut sık kümeleri (IS), destek eşikleri (S) ve güncellemeyi içeren veritabanı parçacığını kullanarak yeni sık kümeleri bulur (IS_d). İşlem çalışması esnasında Sık Kümeler (IS) ve sık kümelerin hangi kayıtlarda bulunduğunu gösterir indeksleri (I_{IS}) günceller.

Dinamik katmanın ikinci algoritması ise çoklu destek eşiklerinde dinamik sık kümeler gizlemesini gerçekleştirir. Bu algoritma da girdi olarak veritabanı güncellemesi (d), sık kümeler (IS), destek eşikleri (S), duyarlı sık kümeler (SI), ve temizlenmis ("sanitized") veritabanını alır. Duyarlı sık kümelerin gizlendiği güncel veritabanını (D') çıktı olarak üretir. Çalışması esnasında sık kümeler (IS) ve sık kümelerin veritabanında bulunduğu kayıtları gösterir sık küme indeksleri (I_{IS}) güncellenir.



Şekil 3. DFIS Platformunun dinamik katmanı

DFIS sınama platformunda bulunan ve 2 katmanın da ortak kullandığı veri yapıları Şekil 2 ve Şekil 3'de gösterilmiştir. Kullanıcı tarafından girilmesi gereken 2 tip veri vardır; bunlardan ilki duyarlı kümeler (SI), ikincisi ise sık kümelerin destek eşikleridir (S). Diğer veriler- sık kümeler (IS) ve sık kümelerin veri tabanın hangi kayıtlarında bulunduğunu gösterir indeksler (I_{IS}) platform algoritmaları tarafından üretilmekte ve güncellenmektedirler.

3.2 DFIS Platform Geliştirme Süreci

DFIS (Çoklu Destek Eşiklerinde Dinamik Sık Kümeler Madenciliği ve Gizleme Platformu) projesi, 6 iş paketinden oluşacak şekilde planlanmış ve oluşturulmuştur. Tablo 4 de gösterildiği gibi bunlar; ***Temel*** (Çoklu destek eşiklerinde çalışan temel sık kümeler madenciliği işlevi çalışmaları), ***Genel*** (tüm bileşenlerin kullanacağı veri yapılarının belirlenmesi), ***Dinamik*** (Çoklu destek eşiklerinde çalışan dinamik sık kümeler işlevi çalışmaları), ***Gizleme*** (Çoklu destek eşiklerinde sık küme gizleme işlevi çalışmaları), ***Dinamik Gizleme*** (Çoklu destek eşiklerinde dinamik sık kümelerin gizlenmesi işlevi çalışmaları) ve ***Entegrasyon*** (Platform bileşenlerinin bir arada çalışabilir hale getirilmesi çalışmaları) iş paketleridir. Tablo 4 de aynı zamanda her bir paket kapsamında yıllar içinde neler yapıldığı gösterilmektedir. İş paketlerine göre yapılanlar aşağıda anlatılmaktadır.

Tablo 4 Proje iş paketlerinde yıllara göre gerçekleştirilenler

	1. YIL	2.YIL	3. YIL
	15/4/2015- 15/4/2016	**15/4/2016 – 15/4/2017)**	**15/4/2017 – 15/4/2018**
Temel	• Literatür taraması, • MISFP-Growth algoritmasının tasarımı, kodlaması ve başarım değerlendirmesi, • Rakip algoritma CFP-Growth++'ın kodlanması • Konferans yayını hazırlanması (Darrab ve Ergenç, 2016-[Ek-1])	• Konferans yayın sunumu (Wseas2016) • Yüksek lisans tezi yazımı ve savunulması [Ek-2] • Yeni algoritma MIS-eclat tasarımı, kodlaması ve başarım değerlendirmesi • Konferans yayını hazırlanması (Darrab ve Ergenç, 2017- [Ek-3])	• Konferans yayın sunumu (KES2017)
Genel	• Platformda kullanılacak tüm ortak veri yapılarının belirlenmesi		
Dinamik	• Literatür taraması, • Dynamic MIS1 ve Dynamic MIS2 (Dynamic CFP-Growth) algoritmalarının tasarımı, kodlamasıve başarım değerlendirmesi	• Konferans yayını hazırlığı (Abuzayed ve Ergenç, 2016-[Ek-4]) • Yüksek lisans tezi yazımı ve savunulması [Ek-5] • Konferans yayın sunumu • Yeni konferans yayın hazırlığı (Abuzayed ve Ergenç, 2017-[Ek-6])	• Konferans yayın sunumu (IDEAS2017)
Gizleme	• Literatür taraması • PGBS algoritmasının tasarımı ve kodlanması • Rakip algoritma TTBS'in kodlanması	• TTBS ve SWA rakip algoritmalarının kodlanması • Konferans yayını hazırlığı (Öztürk ve Ergenç, 2017-[Ek-7])	• Konferansta yayının yayınlanması (KMIS 2017) • İki yeni gizleme algoritmasının tasarımıı (TPGBS ve IPGBS)
Dinamik Gizleme	• Literatür taraması	• Dynamic PGBS algoritmasının tasarımı, kodlaması ve başarım değerlendirmesi • SPITF ve RHID rakip algoritmaların kodlanması	• Makale hazırlığı ve dergide yayınlanması (Öztürk ve Ergenç, 2018a-[Ek-8])
Entegrasyon			• Entegrasyon çalışmaları • Tüm gizleme algoritmalarını içeren yayın hazırlığı (Öztürk ve Ergenç, 2018b-[Ek-9]) • Doktora tezi yazımı

Temel *(Çoklu destek eşiklerinde çalışan temel sık kümeler madenciliği algoritmasının oluşturulması):* Bu algoritma tüm platform için önemli olan bir algoritmadır. Platformun tüm bileşenleri bu algoritmayı ya da bu algoritmanın istenen işleve uyarlanmış sürümlerini

kullanmaktadır Bu algoritma, bölüm 3.1 de platformun arka planında çalışan CokluDestekEşiklerindeSıkKümeler bileşeni olarak tanıtılmıştır.

Tablo 4'de görüldüğü gibi bu paket kapsamında, ilk 18 ay içinde literatür taraması yapılmış iki önemli çıktı verilmiştir, bunlar; 1) Multiple Item Support Frequent Pattern (*MISFP-growth*) algoritması ismi verilerek tasarlanan, kodlanan, başarım değerlendirmesi ile bir konferans yayını olarak Wseas (World Scientific and Engineering Academy and Society) 16th International Conference on Applied Computer Science konferansında, 16 Nisan 2016'da, bursiyer Yüksek Lisans öğrencisi Sadeq Darrab tarafından sunumu yapılan yayın (Darrab ve Ergenç, 2016-[Ek-1]) ve 2) Bursiyer Sadeq Darrab'ın yüksek lisans tezidir [Ek-2].

Mevcut konferans yayınını daha etkili bir konferansta sunulabilecek bir bildiri haline getirebilmek ve MISFP-Growth algoritmasının başarımını karşılaştırmalı olarak değerlendirebilmek için, projenin ikinci yılında yeni bir algoritma daha tasarlanmış, kodlanmış, başarım değerlendirmesi yapılarak tüm çalışma konferans bildirisi haline getirilmiştir. Algoritmanın adı Multiple Item Support Eclat (MIS-eclat) dır, hazırlanan bildiri 21st International Conference on Knowledge Based and Intelligent Information Engineering-KES2017'ye gönderilmiştir.

Projenin üçüncü yılında bildirinin konferansa kabulü üzerine, gerekli sunum hazırlanarak 6-8 Eylül 2017 de Marsilya-Fransa'da yapılan ve bildiri kitapçığı Web of Science Konferans indeksi tarafından taranan KES 2017 de sunulmuştur (Darrab ve Ergenç, 2017-[Ek-3]).

Genel *(tüm bileşenlerin kullanacağı veri yapılarının belirlenmesi):* Bu paket kapsamında, ilk 8 ayda platformda kullanılacak ana veri yapıları belirlenmiştir. Buna göre platformun arka planında madencilik yapacak CokluDestekEşiklerindeSıkKümeler platform bileşeninin ağaç ve tablolarla çalışması uygun bulunmuştur. Aynı şekilde arka planda gizleme yapacak ÇokluDestekEşiklerindeSıkKümeGizleme platform bileşeninin ise bu ağacı sözde-çizge (pseudo graph) haline getirerek kullanmasının uygun olacağına karar verilmiştir. Bunun nedeni; 1) madencilik bileşeni için sık kümelerin tutulduğu ağaç yapısı yeterlidir, 2) gizleme bileşeninin gizleme yaparken kullanmak üzere tüm sık kümelerin geçtiği kayıt numaralarını tutması gerekmektedir, 3) sözde-çizge veri yapısı sık kümeleri ve kayıt numaralarını tutmaya uygundur. Platformun dinamik katmanındaki madencilik ve gizleme bileşenleri ise bu ağaç ve sözde-çizgeyi devingen hale getirerek kullanmışlardır.

Dinamik *(Çoklu destek eşiklerinde çalışan sık kümeler algoritmasının dinamik hale getirilmesi):* Bu iş paketi kapsamında, gelen güncellemeleri işleyebilen bölüm 3.1 de tanıtılan ÇokluDestekEşiklerindeDinamikSıkKümeler adlı işlevin hedeflenmiştir.

Tablo 4'de de görüldüğü gibi bu paket kapsamında ilk 18 ay içerisinde iki önemli çıktı verilmiştir, bunlar; 1) ***Dynamic MIS*** olarak isimlendirilen algoritmanın tasarımı, kodlanması ve başarım değerlendirmesi ile bir konferans bildirisi haline getirilerek, 11-14 Aralık 2016 Çin Macau'da yapılan ve bildiri kitapçığı Web of Science Konferans indeksi tarafından taranan "The 2nd International Conference on Fuzzy Systems and Data Mining- FSDM2016" konferansında sunulan bildiri (Abuzayed ve Ergenç, 2016-[Ek-4]) ve 2) Bursiyer Nourhan Abuzayed'in yüksek lisans tezidir [Ek-5].

Projenin ikinci yılında ayrıca Dynamic MIS algoritmasının, bursiyerin yüksek lisans tezinde yer alan Dynamic CFP-Growth++ ile karşılaştırıldığı yeni bir çalışma yapılmıştır. Farklı veri setleri kullanılarak, iki algoritmanın çalışma zamanı ve bellek kullanım başarımları ölçülmüştür. Tüm çalışma yeni bir konferans bildirisi haline getirilmiştir. Bildiri 21st International Database Engineering & Applications Symposium-IDEAS2017 konferansına gönderilmiştir (Abuzayed ve Ergenç, 2017-[Ek-6]].Bildirinin konferansa kabulü üzerine, projenin son yılında hazırlanan sunum University of the West of England, Bristol, England tarafından düzenlenen IDEAS2017 konferansında, 12-14 Temmuz 2017 de sunulmuştur.

Gizleme *(Çoklu destek eşiklerinde sık küme gizleme algoritması):* Bu paket kapsamında platform arka katmanında kullanılan, bölüm 3.1 de tanıtılan ÇokluDestekEşiklerindeSıkKümeGizleme bileşeninin oluşturulmuştur.

Tablo 4'de görüldüğü gibi bu paket kapsamında, ilk 18 ay içerisinde literatür taraması yapılmış, ***PGBS*** (Pseudo Graph based Sanitization) algoritması tasarlanmış ve kodlanmıştır. Algoritmanın başarım değerlendirmesi yapılabilmesi için rakip çalışmalardan TTBS ve SWA algoritmaları da kodlanmıştır. Algoritmanın başarım değerlendirmesi iki rakip algoritmayla yapılarak genişletilmiştir.

Projenin ikinci yılı içerisinde, çalışmanın tamamı bildiri haline getirilerek bildiri 9th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K) konferansına gönderilmiştir. Bildiri konferansa kabul edilmiş ve bildiri kitapçığında yer almıştır (Öztürk ve Ergenç, 2017-[Ek-7]). Proje seyahat bütçesi yetersiz kaldığı için bildirinin 1-3 Kasım 2017 tarihlerinde Madeira Portekiz'de yapılan söz konusu konferansta sunumu yapılamamıştır.

Projenin üçüncü yılında mevcut gizleme algoritması PGBS'e ilave olarak iki yeni algoritma daha hazırlanmıştır; TGBS ve IPGBS. Bunlardan ilki TPGBS (Transaction Oriented Pseudo Graph based Sanitization) algoritmasıdır. Bu algoritma da PGBS algoritması gibi veri setini sözde çizge ("pseudo graph") halinde tutmaktadır. Aynı şekilde düğümler tekli kümeleri tutarken, kenarlar kayıt kodlarını tutmaktadır. Farklı olarak tutulan kayıt kodları, veri

tabanının tüm kayıtlarının değil sadece duyarlı kayıtlar (“sensitive transaction”) ınındır. Bu şekli ile sözde çizgenin bellekte kaplayacağı yerin azaltılması ve bu algoritmanın Dinamik Gizleme paketi kapsamında da kullanılması hedeflenmiştir. PGBS ve TPGBS silinecek tekli kümeleri tespit ederken benzer stratejiyi izlerler; duyarlı kümelerin (“sensitive itemsets”) ortak elemanlarını silerler. Bu şekli ile paylaşılacak ve gizlemenin yapıldığı veri tabanında mümkün olduğunca az bozma yapılmasını sağlamayı hedeflerler. TGBS algoritmasının özelleştirilmiş bir versiyonu olan DynamicPGBS algoritması 2 rakip dinamik algoritmayla da başarım karşılaştırılması yapılarak bir dergi yayını haline getirilerek SCIE indeksi tarafından International Journal of Data Warehousing and Mining dergisine gönderilmiştir. Kabul edilerek derginin 2018 ikinci çeyrek sayısında yayınlanmıştır (Öztürk ve Ergenç, 2018a-[Ek-8]).

Projenin üçüncü yılında Dinamik Gizleme ve Entegrasyon paketleri çalışmaları sürdürülürken yeni bir gizleme algoritması fikri daha ortaya atılmıştır. Proje planlaması içinde yer almamakla birlikte platformda yer almasının, özellikle ortak elemanı olmayan duyarlı kümelerin, yoğun veri tabanlarında, gizlenmesinde fayda sağlayacağına karar verilmiştir. Algoritmanın adı IPGBS (Itemset Oriented Pseudo Graph Based Sanitization] dır. Bu algoritma da önceki gizleme algoritmalar gibi veri tabanını sözde çizge (“peudo graph” olarak tutmaktadır. Farklı olarak sözde çizgenin düğümleri duyarlı kümeler (“”sensitive itemsets”) i tutmaktadır. Kenarlar TGBS’de olduğu gibi sadece duyarlı kayıt kodlarından “sensitive transactions” oluşmaktadır. IPGBS, projenin önceki 2 gizleme algoritmasından farklı olarak, ortak eleman taşımayan duyarlı kümelerin gizlenmesinde yan etkiyi azaltmaya odaklanmaktadır. Bu üç gizleme algoritmasının (PGBS, TPGBS ve IPGBS) başarım çalışmaları Entegrasyon paketi kapsamında hazırlanan dergi yayınında yer almıştır (Öztürk ve Ergenç, 2018b-[Ek-9]). Bu yayın Turkish Journal of Electrical Engineering and Computer Science dergisine gönderilmiştir, hakemlerin kararı beklenmektedir.

Dinamik Gizleme *(Çoklu destek eşiklerinde sık kümelerin dinamik olarak gizlenmesi):* Bu iş paketi kapsamında platformun dinamik katmanının bölüm 3.1 de tanıtılan ÇokluDestekEşiklerindeDinamikSıkKümeGizleme bileşeninin oluşturulması hedeflenmiştir. İlk 18 ayda bu konudaki benzer çalışmaların literatür taraması yapılmıştır.

Tablo 4’den anlaşılacağı gibi projenin ikinci yılında dinamik gizleme algoritmasının tasarımı ve kodlaması yapılmıştır. Veriyi saklama ve gizleme stratejileri Gizleme paketi algoritmalarından PGBS ve TPGBS’e benzeyen dinamik algoritmanın adı DynamicPGBS (Dynamic Pseudo Graph based Sanitization) dir. Bu dinamik algoritma 4 ayrı süreçten oluşmaktadır. Bunlar “Initialization”-Sözde çizge (“pseudo graph) nin oluşturulması, “Increment Handling”-İlave yönetimi, “Hiding”-Gizleme ve “Publish Database”-Gizlenmiş veri

tabanının yayınlanması. Gizleme süreci bir önceki paket kapsamında anlatılan TGBS algoritmasına benzemektedir. Tüm süreçlere ilişkin alt algoritmaların tasarımı ve kodlanması yapılmıştır

Projenin üçüncü döneminde DynamicPGBS algoritmasının başarım değerlendirmelerinde kullanılmak üzere iki dinamik rakip algoritmanın da kodlaması yapılmıştır; bunlar SPITF and RHID algoritmalarıdır. Dinamik Gizleme paketinin tüm çalışmaları bir dergi yayını haline getirilmiş ve SCIE indeksi tarafından International Journal of Data Warehousing and Mining dergisine gönderilmiştir. Yayın kabul edilerek derginin 2018 ikinci çeyrek sayısında yayınlanmıştır (Öztürk ve Ergenç, 2018a-[Ek-8]).

Entegrasyon *(Tüm proje bileşenlerinin bir araya getirilmesi):* Bu iş paketi kapsamında projenin tüm bileşenleri bir araya getirilmiştir. Bir yazılım ön yüzü hazırlanarak 4 temel test senaryosu ile denenmiştir. Bu senaryolar; çoklu destek eşiklerinde sık kümeler madenciliği, çoklu destek eşiklerinde sık küme gizlemesi, çoklu destek eşiklerinde dinamik sık kümeler madenciliği ve çoklu destek eşiklerinde dinamik sık küme gizlemesidir. Senaryolar çalıştırılarak projenin tüm bileşenlerinin ortak veri setleri, veri yapıları, girdi parametreleri, çıktı dosyaları kullanarak çalıştıkları görülmüştür. Platformun entegre edildiği şekli ile tüm bileşenleri ve hazırlanan yazılımın ön yüzü bir sonraki bölümde detaylı olarak anlatılmıştır.

Projenin üçüncü yılında daha önce de belirtildiği gibi, bu paket kapsamında yeni bir yayın hazırlığı da yürütülmüştür. Bu yayın daha önce hiçbir yayında yer almayan gizleme algoritmasını ortaya koymakta bu algoritmayı önceki 2 gizleme algoritması ile karşılaştırmaktadır. Yayının başarım değerlendirme çalışmalarında, platformun entegre yazılım versiyonu kullanılmış ve aynı zamanda platformun detaylı testi yapılabilmiştir. Hazırlanan yeni yayın (Öztürk ve Ergenç, 2018b-[Ek-9]) Turkish Journal of Electrical Engineering and Computer Science dergisine gönderilmiştir, hakemlerin kararı beklenmektedir.

4. DFIS PLATFORMU

DFIS (“Dynamic Frequent Itemset Mining and Hiding under Multiple Support Thresholds”) sınama platformu ile sürekli güncellenen veri tabanları üzerinde, çoklu destek eşiklerini dikkate alarak, sık kümeler madenciliği yapılabilmesi ve duyarlı sık kümelerin gizlenebilmesi amaçlanmıştır. **DFIS** sık kümeler madenciliği alanın 3 zorluğuna çözüm getiren yazılımların çalıştırılabildiği ve karşılaştırılabildiği bir platformdur. Sözü geçen 3 zorluk çoklu destek eşikleri, veri dinamikliği ve veri büyüklüğüdür. Platform için geliştirilen algoritmalada bu üç zorluk dikkate alınmıştır. Her bir geliştirilen algoritmanın karşılaştırması rakip çalışmalarla yapılabilmesi için rakip algoritmalar da kodlanmış ve platforma eklenmiştir.

4.1 DFIS Bileşenleri

Platformun madencilik ve gizleme işlevleri projenin Temel ve Gizleme paketleri kapsamında geliştirilen algoritmalara, dinamik madencilik ve dinamik gizleme işlevleri de Dinamik ve Dinamik Gizleme paketi kapsamında geliştirilen algoritmalara karşılık gelmektedir. Proje başvurusunda **DFIS** arka planda yer alan işlevler madencilik ve gizleme algoritmalarından, **DFIS** dinamik katmanda yer alan işlevler ise dinamik madencilik ve dinamik gizleme algoritmalarından oluşmaktadır. Geliştirilen tüm algoritmaların başarım testleri rakip algoritmalarla da yapıldığı için rakip algoritmalar da platformda yer almaktadırlar.

Platform bileşenleri 4 grup altında toplanmışlardır. Tablo 5 de bu grupları görmek olasıdır. Buna göre gruplar “MIS IS Mining”-Çoklu Destek Eşiklerinde Sık Kümeler Madenciliği, “MIS IS Hiding”-Çoklu Destek Eşiklerinde Sık Küme Gizleme, “MIS Dynamik IS Mining”- Çoklu Destek Eşiklerinde Dinamik Sık Kümeler Madenciliği ve “ MIS Dynamic IS Hiding”- Çoklu Destek Eşiklerinde Dinamik Sık Küme Gizleme sütunları altında yer almaktadır. Tablo aynı zamanda her grup altında önerilerek geliştirilen algoritmaları, geliştirilen rakip algoritmaları, algoritmanın projenin hangi paketi kapsamında geliştirildiğini, algoritmanın girdilerini, çıktı olarak ürettiklerini ve algoritmanın detaylı anlatımının projenin hangi yayınında yer aldığını da göstermektedir. Bir örnekle açıklamak gerekirse “MIS IS Mining” grubu altında yer alan MIS-eclat algoritmasının rakip algoritması CFP-Growth++ algoritmasıdır, her ikisi de projenin Temel iş paketinde yer almıştır, algoritmanın girdisi LS ve Beta değişkenleridir. Algoritma çalıştırılmadan önce veri setinin seçilmesi gerekmektedir. Algoritma çalıştıktan sonra çoklu destek eşiklerini dikkate alarak sık kümeleri ve çalışmasına ilişkin başarım değerlerini çıktı olarak vermektedir. Bu algoritma (Darrab ve Ergenç, 2017) KES 2017 konferansında sunulan bir yayında detaylı olarak açıklanmıştır. İzleyen bölümlerde tüm yukarıda sözü geçen ve Tablo 5 de gösterilen gruplar detaylı olarak açıklanmaktadır.

“MIS IS Mining- Multiple Item Support Threshold based Itemset Mining” grubu: Bu grup bileşenlerin görevi, platformun arka planında yer alan çoklu destek eşiklerinde madencilik işlevlerini yerine getirmektir. Tablo 5 de görülen 3. ve 4. sütunlarda bu grup bileşenlerine ilişkin bilgiler yer almaktadır. Tablodan anlaşılabileceği gibi bu grubun çalışmaları projenin Temel iş paketi kapsamında ele alınmıştır. Bu kapsamda iki yeni algoritma önerilmiş ve geliştirilmiştir. Bu algoritmalar MISFP-Growth (Darrab and Ergenç 2016) ve MIS-eclat (Darrab ve Ergenç, 2017) algoritmalarıdır. İki yeni algoritmanın da başarım testlerinde kullanılmak üzere rakip algoritma olan CFP-Growth++ (Kiran vd., 2011) algoritması da geliştirilmiş ve platforma eklenmiştir.

Bu grup altında çalışan algoritmaların tamamı girdi olarak LS (en düşük destek eşiği) ve Beta (kontrol parametresi) değerlerini alırlar. Bu iki parametre çoklu destek eşiklerini kümelerin gerçek destek değerlerine bağlı olarak oluşturmakta kullanılır. Bu grubun ilk bileşeni olan MISFP-Growth (Darrab ve Ergenç, 2016) algoritması girdi veri setini “pattern-growth” ağacı halinde tutmakta ve en küçük eşik değeri altında desteği olan kümeleri erken budayarak rakip algoritma olan CFP-Growth++ algoritmasına işlem zamanı ve bellek gereksinimi açısından üstünlük sağlamaktadır. Bu algoritma projenin ilk yayını olan WSEAS 2016 konferansında detaylı olarak anlatılmıştır. Bu grubun ikinci algoritması olan MS-eclat (Darrab ve Ergenç, 2017) veri setini dikey gösterim şekli ile bellekte tutarak çalışmakta ve belli özellikte veri setlerinde MISFP-Growth++ ve CFP-Growth++ algoritmalarından daha iyi başarım göstermektedir. Bu algoritmaya dair proje çıktısı KES 2017 konferansında sunulmuştur. Bu grup altında çalışan tüm algoritmalar çıktı raporlarında çoklu destek eşiklerinde bulunan sık kümeleri ve girilen parametrelere karşılık gelen başarım sonuçlarını yayınlar.

“MIS IS Hiding- Multiple Item Support Threshold based Itemset Hiding” grubu: Bu grup bileşenlerin görevi platformun arka planında yer alan çoklu destek eşiklerinde duyarlı (“sensitive”) sık kümeleri gizleme işlevlerini yerine getirmektir. Tablo 5’de görülen 5., 6. ve 7. sütunlarda bu grup bileşenlerine ilişkin bilgiler yer almaktadır. Bu kapsamda üç yeni algoritma önerilmiş ve geliştirilmiştir. Bu algoritmalar PGBS (Öztürk ve Ergenç (2017)), TGBS (Öztürk ve Ergenç, 2018) ve IPGBS algoritmalarıdır. Yeni algoritmaların başarım testlerinde kullanılmak üzere rakip algoritmalar olan TTBS (Kuo vd., 2008), SWA (Oliveira ve Zaiane, 2003) ve RS (Cheng vd., 2016) algoritmaları da geliştirilmiş ve platforma eklenmiştir. Bu grubun tüm algoritmaları projenin Gizleme paketi kapsamında geliştirilmiştir.

Bu grup altında çalışan algoritmaların tamamı girdi olarak duyarlı (“sensitive”) sık kümeleri ve bunların eşik değerlerini alırlar. Bir diğer deyişle, gizlenmesi gereken sık kümeler ve bunların düşürülmesi beklenen eşik değerleri bu gruptaki algoritmaların girdileridir. Bu grupta önerilen

tüm algoritmalar veri setini sözde çizge (“pseudo graph”) halinde tutmaktadır. Bu çizgeyi oluşturma ve gizleme stratejileri farklılık göstermektedirler. Bu grubun ilk algoritması olan PGBS algoritması IC3K-KMIS 2017 konferansında sunulmuştur. Bu algoritmanın kullandığı sözde çizge (“pseudo graph”) tekli kümeleri ve tüm kayıt etiketlerini tutmaktadır. İkinci algoritma olan TPGBS algoritması projenin dergi yayındaki (Öztürk ve Ergenç, 2018a) Dinamik PGBS algoritmasının gizleme işlemini esas almaktadır. Bu algoritmadaki sözde çizge tüm tekli kümeler ve duyarlı kayıt etiketlerini tutmaktadır. Daha az verinin tutulduğu bu sözde çizge ile bellek gereksinimi azaltılmıştır. Her iki algoritma da gizleme yaparken, duyarlı kümelerin ortak tekli elemanlarına öncelik vererek, üretilecek olan gizlenmiş veri tabanında, duyarlı olmayan bilgideki kaybı azaltmayı hedeflenmiştir. Son algoritma olana IPGBS algoritması proje entegrasyon adımında ortaya atılabilmiş, önceki iki algoritmayla başarım değerlendirmesi yapılarak bir dergi yayını oluşturulmuş ve Tubitak “Turkish Journal of EE&CS” gönderilmiştir (Öztürk ve Ergenç, 2018b). Bu grup altında çalışan tüm algoritmalar çıktı olarak gizleme yapılmış veri tabanını ve girilen parametrelere karşılık gelen başarım sonuçlarını yayınlar.

“MIS Dynamic IS Mining-Multiple Item Support Threshold based Dynamic Itemset Mining” grubu: Bu grup bileşenlerinin görevi, platformun dinamik katmanda yer alan çoklu destek eşiklerinde sık kümeler madenciliği işlevlerini yerine getirmektir. Tablo 5 de görülen 8.ve 9. sütunlarda bu grup bileşenlerine ilişkin bilgiler yer almaktadır. Bu kapsamda iki yeni algoritma önerilmiş ve geliştirilmiştir. Bu algoritmalar Dynamic MIS (Abuzayed ve Ergenç, 2016) ve Dynamic CFP-Growth (Abuzayed ve Ergenç, 2017) algoritmalarıdır. Bu algoritmaların başarım testlerinde platforma eklenmiş olan CFP-Growth++ (Kiran vd., 2011) algoritması da kullanılmıştır. Bu grubun tüm algoritmaları projenin Dinamik paketi kapsamında geliştirilmiştir.

Bu grup altında çalışan algoritmaların tamamı girdi olarak LS (en düşük destek eşiği) ve Beta (kontrol parametresi) değerlerini alırlar. Bu iki parametre çoklu destek eşiklerini kümelerin gerçek destek değerlerine bağlı olarak oluşturmakta kullanılır. Bu grup algoritmalarının girdisi olan bir diğer parametre de veri setinin üzerine ilave olarak gelmesi beklenen veri büyüklüğüdür. Bu parametre için yüzde değer belirlenebilir ya da ilave gelecek veri seti bir dosya halinde verilebilir. Bu grup da önerilen her iki algoritma da “pattern growth” ağaçları kullanmaktadırlar. Dinamik MIS iki başlık tablosu kullanarak sık ve sık olmayan kümeleri ayırarak takip etmektedir. Bu bellek kullanımını arttırmakla beraber çalışma zamanını kısaltmaktadır. Dinamik MIS algoritması FSDM 2016 konferansında, Dinamik CFP-Growth algoritması ise IDEAS 2017 konferansında sunulmuştur. Bu grup altında çalışan tüm

algoritmalar çıktı raporlarında çoklu destek eşiklerinde bulunan sık kümeleri ve girilen parametrelere karşılık gelen başarım sonuçlarını yayınlar.

Tablo 5. DFIS Platform bileşenleri

DFIS Platform: Dynamic Frequent Itemsets for MIS									
		Background Layer					**Dynamic Layer**		
		MIS IS Mining		**MIS IS Hiding**			**MIS Dynamic IS Mining**		**MIS Dynamic IS Hiding**
	Algorithms	**MISFP-Growth**	**MIS-eclat**	**PGBS**	**TPGBS**	**IPGBS**	**Dynamic CFP-Growth++**	**Dynamic MIS**	**DynamicPGBS**
	Comp. Algos	**CFP-Growth++**		**TTBS, SWA and RS**			**CFP-Growth+**		**SPITF and RHID**
	Pr. Pack.	Temel	Temel	Gizleme	Gizleme	Gizleme	Dinamik	Dinamik	Dinamik Gizleme
Input	Sensitive Info.			Sensitive Itemsets	Sensitive Itemsets	Sensitive Itemsets			Sensitive Itemsets
	MIS Thresholds	LS (*) and Beta(*)	LS(*) and Beta(*)	Sensitive Support Thresholds	Sensitive Support Thresholds	Sensitive Support Thresholds	LS(*) and Beta(*)	LS(*) and Beta(*)	Sensitive Support Thresholds
	Increment						Increment size (%) or increment file	Increment size (%) or increment file	Increment size (%) or increment file
	Datasets	Ready (Tablo 6)/ prepared	Ready (Tablo 6)/ prepared	Ready (Tablo 6)/ prepared	Ready (Tablo 6)/ prepared	Ready (Tablo 6)/ prepared	Ready (Tablo 6)/ prepared	Ready (Tablo 6)/ prepared	Ready (Tablo 6)/ prepared
Output	Output	Frequent Itemsets	Frequent Itemsets	Sanitized Dataset	Sanitized Dataset	Sanitized Dataset	Frequent Itemsets	Frequent Itemsets	Sanitized Dataset
	Analysis	Frequent Itemsets & Performance Results	Frequent Itemsets & Performance Results	Performance Results	Performance Results	Performance Results	Frequent Itemsets & Performance Results	Frequent Itemsets & Performance Results	Performance Results
	Publication	WSEAS 2016 (Darrab ve Ergenç, 201)-[Ek-1])	KES 2017 (Darrab ve Ergenç, 2017-Ek-3])	KMIS 2017 (Öztürk ve Ergenç, 2017-[Ek-7])	IJDWM 2018 (Öztürk ve Ergenç, 2018a-[Ek-8])	Submitted to Turkish J. of EE & CS (Öztürk ve Ergenç, 2018b-[Ek-9])	FSDM 2016 (Abuzayed ve Ergenç, 2016-[Ek-4])	IDEAS 2017 (Abuzayed ve Ergenç, 2017)-[Ek-6])	IJDWM 2018 (Öztürk ve Ergenç, 2018a-[Ek-8])
*: LS: Least Minimum Item Support, Beta: Control Parameter									

“MIS Dynamic IS Hiding- Multiple Item Support Threshold based Dynamic Itemset Hiding” grubu: Bu grup bileşenlerin görevi platformun dinamik katmanında yer alan çoklu destek eşiklerinde dinamik olarak duyarlı (“sensitive”) sık kümeleri gizleme işlevlerini yerine getirmektir. Tablo 5 de 10. sütunda bu grup bileşenlerine ilişkin bilgiler yer almaktadır. Bu kapsamda bir yeni algoritma önerilmiş ve geliştirilmiştir. Bu algoritma Dinamik PGBS (Öztürk ve Ergenç, 2018) algoritmasıdır. Yeni algoritmanın başarım testlerinde kullanılmak üzere rakip algoritmalar olan SPITF (Dai ve Chiang, 2010) ve RHID (Jadav vd., 2014) algoritmaları da geliştirilmiş ve platforma eklenmiştir. Bu grubun tüm algoritmaları projenin Dinamik Gizleme paketi kapsamında geliştirilmiştir.

Tablo 6. DFIS Platform hazır veri setleri

Veri Seti	Tip	Kayıt Sayısı	Tekli Küme Adedi	Ortalama Kayıt Uzunluğu	Yoğunluk
Kosarak	Gerçek	990,002	41,271	8.1	0.002
Retail	Gerçek	88,126	16,470	10.3	0.006
Pumsb	Gerçek	49,047	2,113	75	3.6
Mushroom	Gerçek	8,124	119	23	19.4
BM-POS	Gerçek	515,597	1,657	6.5	0.39
BMSWebView2	Gerçek	75,512	3,340	5	0.15
Chess	Gerçek	3,196	75	37	49.4
Connect	Gerçek	67,557	129	43	33.4
T10I4D100K	Sentetik	100,000	870	10.1	1.15
T40I1D100K		88,162	942	40	4.25
SyntheticDense		29,166	99	43.09	43.5
SyntheticSparse		28,417	9,479	11.48	0.1212

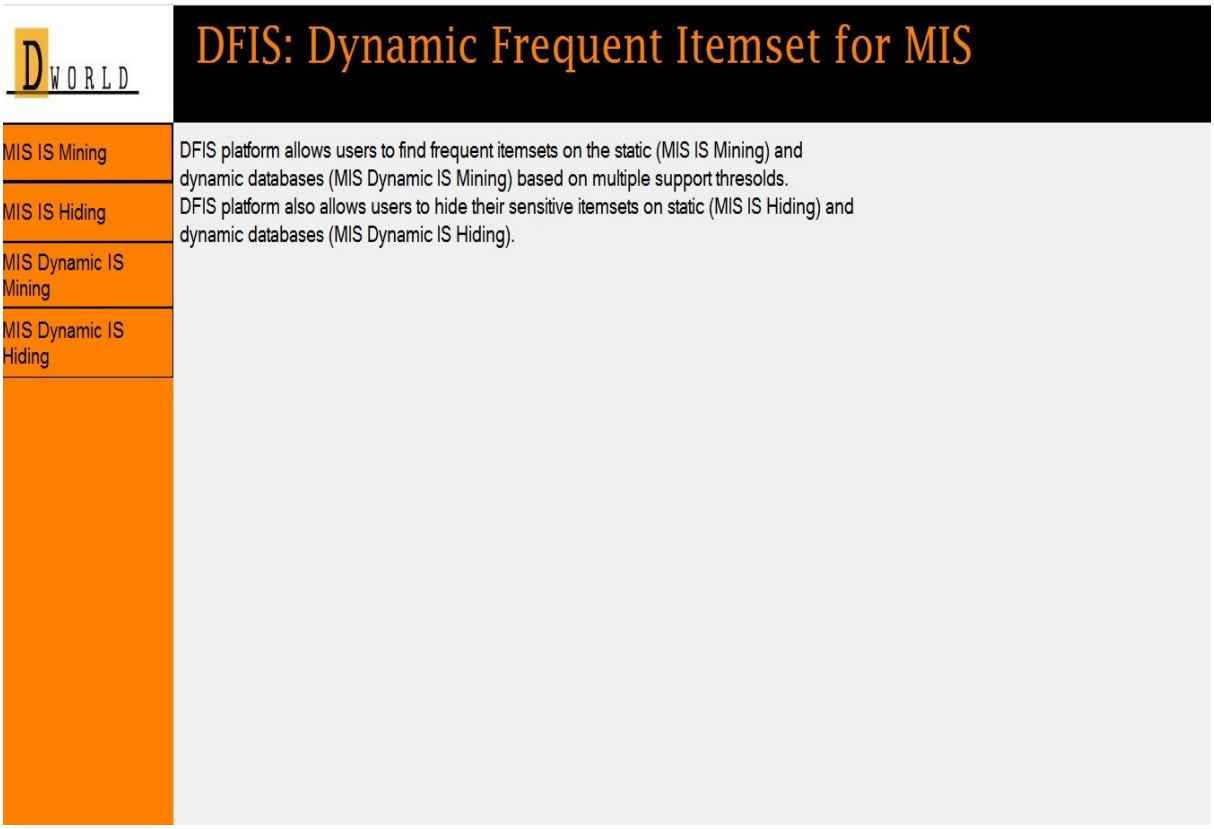
Bu grup altında çalışan algoritmaların tamamı girdi olarak duyarlı (“sensitive”) sık kümeleri ve bunların eşik değerlerini alırlar. Bu grup algoritmalarının girdisi olan bir diğer parametre de veri setinin üzerine ilave olarak gelmesi beklenen veri seti ile ilgilidir; yüzde cinsinden veri büyüklüğü verildiğinde orijinal veri setinin istenen büyüklükteki parçası ilave olarak alınır ya da ilave veri seti ayrı bir dosya olarak verilebilir. Bu grupta önerilen algoritma veri setini sözde çizge (“pseudo graph”) halinde tutmaktadır. Bu çizge veri setine gelen artımları bellekte tutmakta ve gizleme işleminin diske gitmeden çizge üzerinde yapılabilmesine olanak vermektedir. Veri tabanı paylaşılmak istendiğinde sadece diske erişilerek çizge ve gizleme çözümünü içeren tablo kullanılarak hassas bilginin gizlendiği veri tabanı hazırlanabilmektedir. Bu grubun algoritmasını ve rakiplerle karşılaştırmasını içeren detaylı çalışma bir makale

haline getirilmiş, dergide yayınlanmaya uygun bulunmuş ve yayınlanmıştır. Bu grup altında çalışan tüm algoritmalar çıktı olarak gizleme yapılmış veri tabanını ve girilen parametrelere karşılık gelen başarım sonuçlarını yayınlar.

Tablo 6 da **DFIS** sınama platform içinde hazır olarak bulunan veri setlerinin özellikleri gösterilmektedir. Tabloda bazı veri setleri gerçek bazıları ise sentetiktir. Gerçek veri setleri UCI Repository (Blake ve Merz, 1998) ya da Brijs vd. (1999) dan temin edilmişlerdir. Sentetik veri setleri ise "IBM quest dataset generator" (Bhalodiya, 2014) ile oluşturulmuşlardır. Veri setlerinin kayıt sayısı, tekli küme adedi, ortalama kayıt uzunluğu ve yoğunlukları farklılıklar göstermektedir. Veri setinin yoğunluğu, ortalama kayıt uzunluğunun tekli küme adedine bölünmesiyle bulunmaktadır. Veri setlerinin farklı özelliklerde olmaları sağlanarak her bir özelliğin madencilik ve gizleme algoritmalarının başarımlarına etkisinin ölçülebilmesi sağlanmıştır. Platformda hazır olarak bu veri setleri dışında da gerçek ya da sentetik veri setleri seçilerek kullanılabilir.

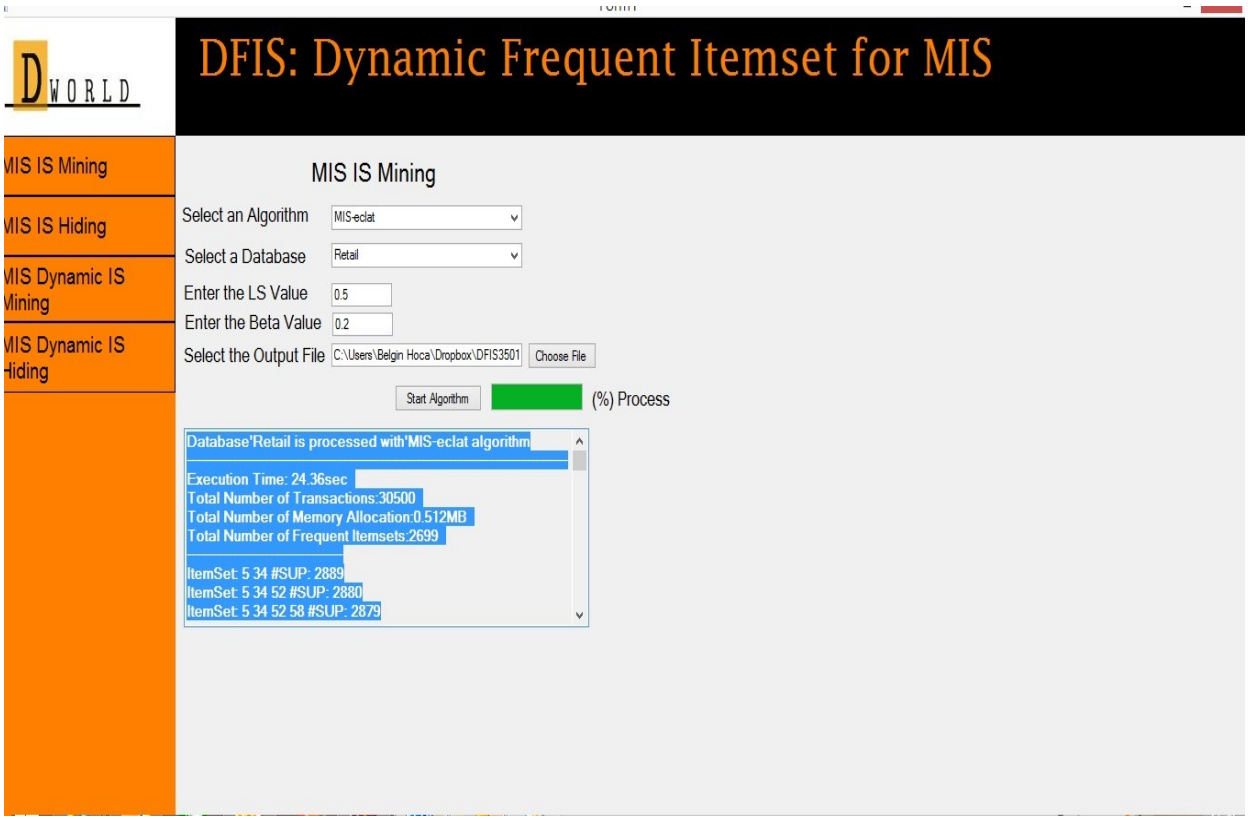
4.2 DFIS Yazılım Önyüzü

DFIS platformu için hazırlanan yazılımın ana menüsü Şekil 4 de görülmektedir. Menüdeki 4 sekme projenin ana işlevlerine ve Tablo 5 de açıklanan proje ana bileşen gruplarına karşılık gelmektedir. Bu sekmeler "**MIS IS Mining**"-Çoklu Destek Eşiklerinde Sık Kümeler Madenciliği, "**MIS IS Hiding**"-Çoklu Destek Eşiklerinde Sık Küme Gizleme, "**MIS Dynamic IS Mining**"- Çoklu Destek Eşiklerinde Dinamik Sık Kümeler Madenciliği ve "**MIS Dynamic IS Hiding**"- Çoklu Destek Eşiklerinde Dinamik Sık Küme Gizleme'dir. "MIS IS Mining" sekmesinden çoklu destek eşiklerinde madencilik işlevleri, "MIS IS Hiding" sekmesinden çoklu destek eşiklerinde gizleme işlevleri, "MIS Dynamic Mining" sekmesinden sürekli güncelleme alan veri setleri için çoklu destek eşiklerini dikkate alan madencilik işlevleri, "MIS Dynamic IS Hiding" sekmesinden ise sürekli güncelleme alan veri setleri için çoklu destek eşiklerini dikkate alan gizleme işlevleri çalıştırılabilmektedir. Aşağıda her bir sekmeden çağrılabilecek işlevlerin nasıl çalıştırılabildiğini gösteren ara-yüz formlar açıklanacaktır.



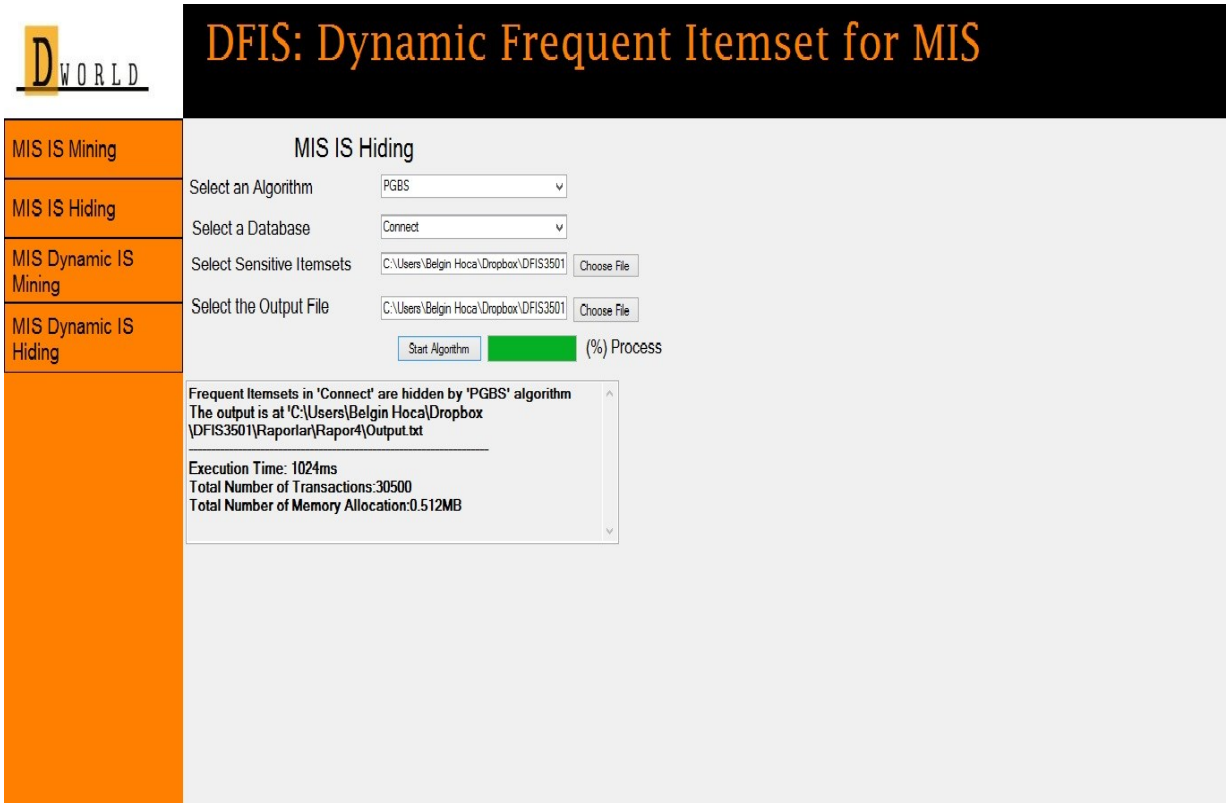
Şekil 4. DFIS Projesi önyüzü

Şekil 5 “MIS IS Mining” sekmesi ile çalışan ara-yüz formunu göstermektedir. Formda görüldüğü gibi ilkin çalıştırılmak istenen algoritmanın seçilmesi gerekmektedir. Bir önceki bölümde anlatıldığı gibi bu grup altında 3 algoritma bulunmaktadır; MISFP-Growth, MIS-eclat ve CFP-Growth++. MISFP-Growth (Darrab ve Ergenç, 2016) ve MIS-eclat (Darrab ve Ergenç, 2017) projenin Temel iş paketi altında önerilen ve yayın haline getirilerek iki ayrı konferansda sunulmuş algoritmalardır. Aynı paket içinde rakip algoritma olan CFP-Growth++ (Kiran vd., 2011) algoritması da geliştirilmiştir ve bu formda istenen algoritma olarak çalıştırılabilir. Formda gösterilen örnek senaryoda istenen algoritma olarak MIS-eclat seçilmiştir.



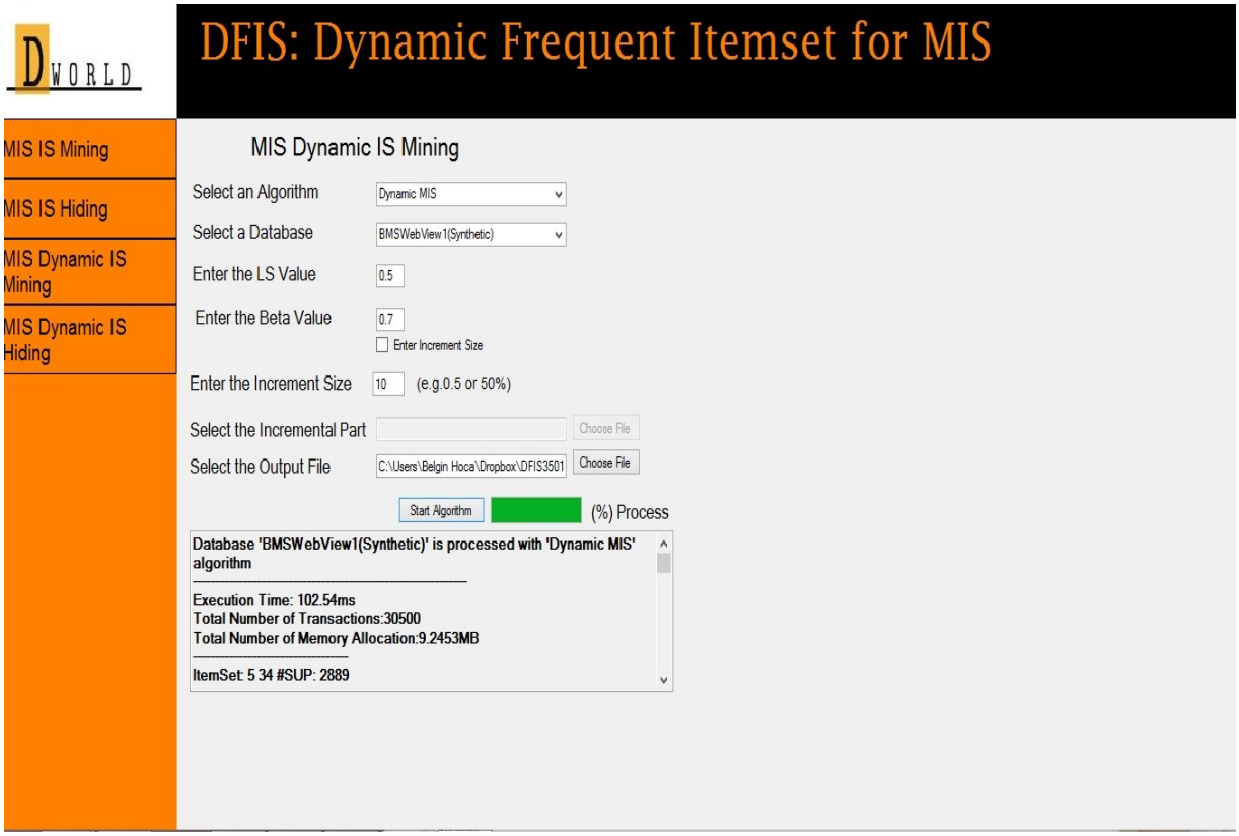
Şekil 5. Çoklu destek eşiklerinde sık kümeler madenciliği bileşeni önyüzü

Kullanıcının algoritma seçimini takiben algoritmanın üzerinde çalışacağı veri setini belirlemesi istenmektedir. Kullanıcı platforma gömülü değişik özelliklerdeki (Tablo 6) veri setlerinden ya da daha önce kendisinin ürettiği veri setini seçerek devam eder. Daha sonra kullanıcının LS (en düşük destek eşiği) ve Beta (kontrol parametresi) değerlerini girmesi beklenmektedir. Bu iki parametre çoklu destek eşiklerini kümelerin gerçek destek değerlerine bağlı olarak oluşturmakta kullanılır. Ardından kullanıcı sonuçların yazılması istenen dosyanın erişim patikasını ve ismini girer. “Start Algorithm” tuşuna basıldığında algoritma çalıştırılır. Çalışma tamamlandığında programın çalışma süresi, toplam kayıt adedi, toplam bellek kullanımı ve üretilen sık küme adedi çıktı dosyasının başında yayınlanır. Aynı formun devamında programın ürettiği sık kümeler gösterilir. Çıktı formu aynı zamanda üzerinde ek çalışmaların yapılabileceği kullanıcının belirlediği çıktı dosyasına da kaydedilir.



Şekil 6. Çoklu destek eşiklerinde sık kümeler gizleme bileşeni önyüzü

Şekil 6 “MIS IS Hiding” sekmesi ile çalışan ara-yüz formunu göstermektedir. Formda görüldüğü gibi ilkin çalıştırılmak istenen algoritmanın seçilmesi gerekmektedir. Bir önceki bölümde anlatıldığı gibi bu grup altında 6 algoritma bulunmaktadır. Bunlardan PGBS (Öztürk ve Ergenç, 2017) ve TPGBS (Öztürk ve Ergenç, 2018) projenin Gizleme paketi içinde geliştirilen bir konferans kitapçığında ve bir bilimsel dergide yayınlanmış algoritmalardır. Üçüncü algoritma olan IPGBS algoritmasının ilişkin yayının hazırlığı projenin Entegrasyon paketinde yapılmış ve bir bilimsel dergiye gönderilmiştir. Bu sekmede ayrıca 3 rakip algoritma da bulunmaktadır; TTBS (Kuo vd., 2008), SWA (Oliveira ve Zaiane, 2003) ve RS (Cheng vd., 2016). Formda gösterilen örnek senaryoda istenen algoritma olarak PGBS algoritması seçilmiştir.



Şekil 7. Çoklu destek eşiklerinde dinamik sık kümeler madenciliği bileşeni önyüzü

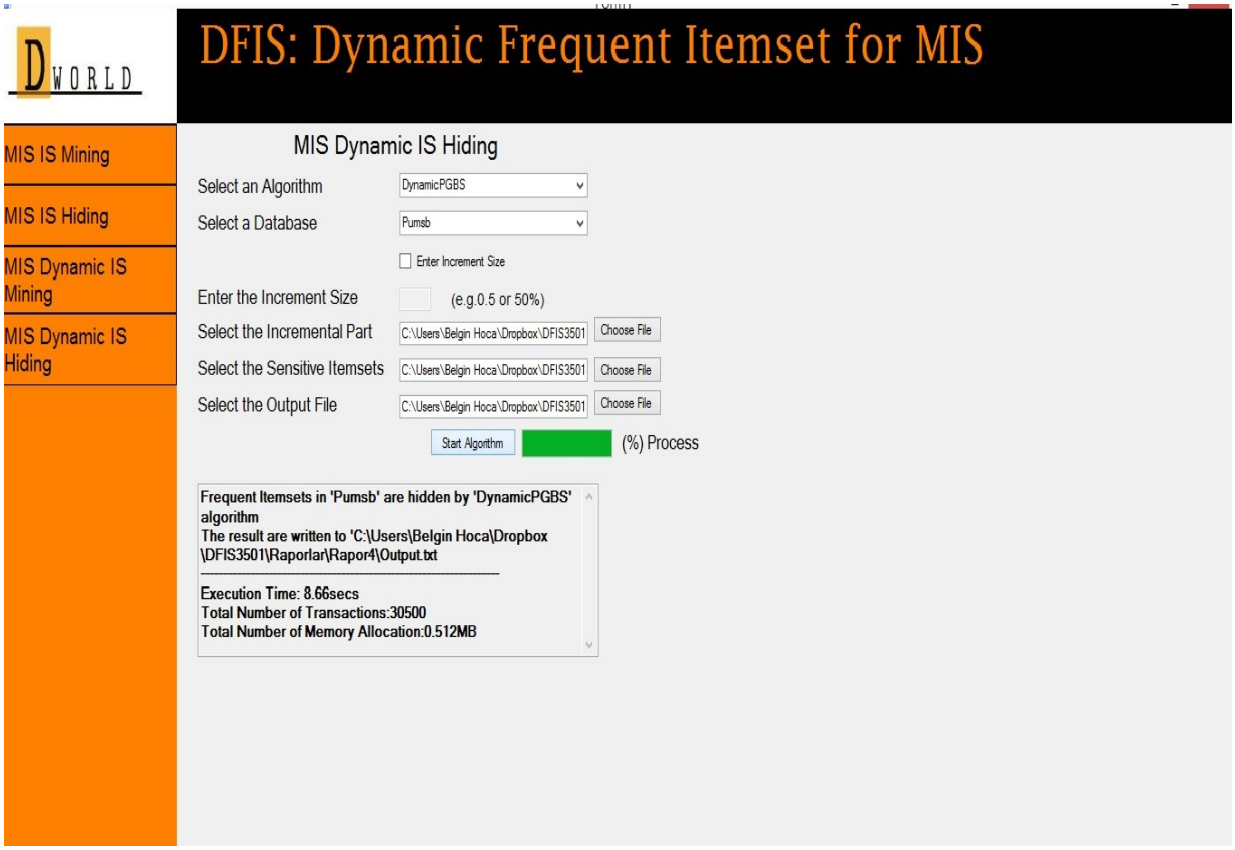
Kullanıcının algoritma seçimini takiben algoritmanın üzerinde çalışacağı veri setini belirlemesi istenmektedir. Kullanıcı platforma gömülü değişik özelliklerdeki (Tablo 6) veri setlerinden ya da daha önce kendisinin ürettiği veri setini seçerek devam edebilir. Daha sonra kullanıcının tüm duyarlı ("sensitive") kümeleri ve belirlediği duyarlı destek eşiklerini ("sensitive thresholds") durduğu dosyayı seçmesi istenir. Ardından kullanıcı sonuçların yazılması istenen dosyanın erişim patikasını ve ismini girer. "Start Algorithm" tuşuna basıldığında algoritma çalıştırılır. Çalışma tamamlandığında gizlemenin yapıldığı veri setinin hangisi dosya olduğu alttaki pencerede görülür. Aynı pencerenin devamında programın çalışma süresi, toplam kayıt adedi, toplam bellek kullanımı bilgileri de bulunmaktadır.

Şekil 7 "MIS Dynamic IS Mining" sekmesi ile çalışan ara-yüz formunu göstermektedir. Formda görüldüğü gibi ilkin çalıştırılmak istenen algoritmanın seçilmesi gerekmektedir. Bir önceki bölümde anlatıldığı gibi bu grup altında 3 algoritma bulunmaktadır. Bunlardan Dynamic MIS (Abuzayed ve Ergenç, 2016) ve Dynamic CFP-Growth (Abuzayed ve Ergenç, 2017) algoritmaları projenin Dinamik paketinde hazırlanmış ve iki ayrı uluslararası konferansta sunulmuşlardır. Bu bölümde çalıştırılabilecek üçüncü algoritma CFP-Growth++ (Kiran vd., 2011) algoritmasıdır. Formda gösterilen örnek senaryoda istenen algoritma olarak Dynamic MIS algoritması seçilmiştir.

Kullanıcının algoritma seçimini takiben algoritmanın üzerinde çalışacağı veri setini belirlemesi istenmektedir. Kullanıcı platforma gömülü değişik özelliklerdeki (Tablo 6) veri setlerinden ya da daha önce kendisinin ürettiği veri setini seçerek devam edebilir. Daha sonra kullanıcının, çoklu destek eşiklerinin belirlenmesi için kullanılan LS ve Beta parametrelerini girmesi gerekir. Bu dinamik özellikli algoritmalar için kullanıcının "increment" (ilave) büyüklüğü ya da ilavenin bulunduğu dosyayı girmesi istenir. Ardından kullanıcı sonuçların yazılması istenen dosyanın erişim patikasını ve ismini girer. "Start Algorithm" tuşuna basıldığında algoritma çalıştırılır. Çalışma tamamlandığında alttaki pencerede programın çalışma süresi, toplam kayıt adedi ve toplam bellek kullanımı bilgileri ve programın ürettiği sık kümeler görülebilmektedir. Çıktı formu aynı zamanda üzerinde ek çalışmaların yapılabileceği kullanıcının belirlediği çıktı dosyasına da kaydedilir.

Şekil 8 "MIS Dynamic IS Hiding" sekmesi ile çalışan ara-yüz formunu göstermektedir. Formda görüldüğü gibi ilkin çalıştırılmak istenen algoritmanın seçilmesi gerekmektedir. Bir önceki bölümde anlatıldığı gibi bu grup altında 3 algoritma bulunmaktadır. Bunlardan Dynamic PGBS (Öztürk ve Ergenç, 2018) algoritması projenin Dinamik Gizleme paketinde hazırlanmış ve bir bilimsel dergide yayınlanmıştır. Bu bölümde çalıştırılabilecek diğer iki algoritma rakip algoritmalardır. Bunlar SPITF (Dai ve Chiang, 2010) ve RHID (Jadav vd., 2014) algoritmalarıdır. Formda gösterilen örnek senaryoda istenen algoritma olarak DynamicPGBS seçilmiştir.

Kullanıcının algoritma seçimini takiben algoritmanın üzerinde çalışacağı veri setini belirlemesi istenmektedir. Kullanıcı platforma gömülü değişik özelliklerdeki (Tablo 6) veri setlerinden ya da daha önce kendisinin ürettiği veri setini seçerek devam edebilir. . Daha sonra kullanıcının tüm duyarlı ("sensitive") kümeleri ve belirlediği duyarlı destek eşiklerini ("sensitive thresholds") durduğu dosyayı seçmesi istenir. Bu dinamik özellikli algoritmalar için kullanıcının "increment" (ilave) büyüklüğü ya da ilavenin bulunduğu dosyayı girmesi istenir. Ardından kullanıcı sonuçların yazılması istenen dosyanın erişim patikasını ve ismini girer. "Start Algorithm" tuşuna basıldığında algoritma çalıştırılır. Çalışma tamamlandığında gizlemenin yapıldığı veri setinin hangisi dosya olduğu alttaki pencerede görülür. Aynı pencerenin devamında programın çalışma süresi, toplam kayıt adedi ve toplam bellek kullanımı bilgileri de bulunmaktadır.



Şekil 8. Çoklu destek eşiklerinde dinamik sık kümeler gizleme bileşeni önyüzü

5. SONUÇ

Veri madenciliği, büyük veri tabanlarında klasik veritabanı sorgularıyla çıkarılamayacak gizli örüntüleri bulmayı hedefler. İlişki kuralları madenciliği ya da sık kümeler madenciliği bu alanın en çok kullanılan dolayısıyla en çok araştırılan görevlerindendir. Burada, veri tabanı nesneleri arasındaki ilginç ilişkiler bulunmaya çalışılır. İlginç ilişkileri oluşturan kümeler tanımlanan destek eşiğini geçen sıklıkta olan kümelerdir. Bu alanın araştırmalarının önemli bir bölümü sık kümeleri bulurken tek bir eşik değeri olduğunu ve veritabanlarının değişmediğini varsayarlar. Oysaki tek eşik değeri farklı önemdeki sık kümeleri bulmayı zorlaştırır. Aynı şekilde sürekli güncellenen veri tabanlarında, sık kümelerin her defasında madencilik işlemini baştan yapmadan güncellenebildiği madencilik yaklaşımlarına ihtiyaç vardır. Sık kümeler madenciliğinin paylaşılan veri üzerinde yapıldığı durumlarda ise veri sahiplerinin kendileri için stratejik olan sık kümeleri gizlemeleri gerekmektedir. Dolayısıyla sık kümeler madenciliği alanında madencilik çözümleri kadar gizleme çözümlerine de ihtiyaç vardır. Bu proje kapsamında sözü edilen sık kümeler madenciliği, dinamik sık kümeler madenciliği, sık küme gizlemesi ve dinamik sık küme gizlemesi araştırmaları yapılmıştır. Tüm çalışmalarda çoklu destek eşikleri dikkate alınmıştır. Aynı şeklide tüm çözümlerin veri büyüklüğüyle baş edebilecek etkin veri yapıları kullanmaları sağlanmıştır.

Proje bir sınama platformunu parça parça oluşturmayı ve son bölümde tüm parçaları bir araya getirmeyi hedeflemiştir. Parçalar farklı işlevleri yapan, proje kapsamında önerilerek geliştirilen ya da rakip algoritma gruplarıdır. Her bir grup elemanının kendi grubu içindeki diğer algoritmalarla karşılaştırması yapılmıştır. Önerilen tüm algoritmalar proje süresince konferans bildirisi olarak ya da bilimsel dergi makalesi olarak dökümante edilmiş ve araştırma dünyası ile paylaşılmıştır. Önerilen algoritmalar aynı zamanda projede çalışan bursiyerlerin yüksek lisans ya da doktora tezlerinde yer almıştır. Tüm yayınlar bu rapor eki olarak da verilmiştir.

Oluşturulan sınama platfoormu 4 grup algoritmayı içermektedir. İlk grupda çoklu destek eşiklerinde sık kümeler madenciliği yapabilen algoritmalar vardır. Bunlar MISFP-Growth (Darrab ve Ergenç, 2016) ve MIS-eclat (Darrab ve Ergenç, 2017) algoritmalarıdır. İkinci grup algoritma dinamik veri tabanlarında ve yine çoklu destek eşklerini dikkate alarak sık kümeleri bulmayı hedefler. Bu grup içinde 2 yeni algoritma önerilerek geliştirilmiştir; Dynamic MIS (Abuzayed ve Ergenç, 2016) ve Dynamic CFP-Growth (Abuzayed ve Ergenç, 2017). Üçüncü grup madalyonun öbür yüzündeki gereksinim olan gizleme algoritmalarından oluşur. Bu grubun ilk algoritması PGBS (Öztürk ve Ergenç, 2017) dir, sonra TPGBS (Öztürk ve Egenç, 2018a) ve IPGBS (Öztürk ve Ergenç, 2018b) algoritmaları gelmiştir. Dördüncü grup dinamik veri tabanları için hazırlanan sık kümeler madenciliği algoritmalarından oluşmuştur. Proje

önerisi algoritma Dinamik PGBS algoritmasıdır (Öztürk ve Ergenç, 2018a). Kısaca söylemek gerekirse projenin yaygın etkisi olarak, platform bileşenleri 5 konferans kitapçığında ve 1 bilimsel dergide yer almışlardır, ikinci dergi makalesi için de hakem görüşü beklenmektedir. Proje çıktıları arasında tamamlanmış olan iki yüksek lisans tezi ve bir de yazım aşamasında olan doktora tezi de vardır.

Araştırmacı yetiştirme ya da yürütücünün kariyerine verilen katkı olarak proje değerlendirildiğinde sonuçlar olumludur. Projede çalışan üç bursiyer alan tarama, yeni fikir oluşturma, fikri algoritmaya evirebilme, rakip çalışmaları anlama-kodlama, önerilen algoritmayı rakiplerle karşılaştırabilme, tüm çalışmayı yayın olarak dökümante edebilme gibi araştırma dünyasının temel becerilerini kazanmışlardır. Yürütücü açısından olaya bakıldığında ise çalışma grubunda alana dair bilgi birikiminin arttığını, projenin yeni proje fikirleri doğurduğunu, proje oluşturma-yönetme-sonuçları dökümante etme-yayın oluşturma becerilerini pekiştirdiğini söyleyebiliriz. Yürütücü bu kariyer projesini yürütürken aynı zamanda kariyerinde bir basamak yükselerek doçent ünvanı almıştır.

Projenin önemli çıktısı olan ve tüm çalışmalara tek bir yazılım önyüzü ile erişilebilmesini sağlayan DFIS önyüzünün yeteneklerinin geliştirilerek internetde bir servis olarak açılması proje sonrası planlanan çalışmalar arasındadır. Bu önyüze ilave edilmesi düşünülen seçenekler şunlardır; girdi olarak kullanılacak veri setlerinin hazırlanabildiği bölüm, gizleme işlevleri çalıştırılırken girdi olarak verilen duyarlı küme dosyalarının hazırlanabildiği bölüm, üretilen sonuçların görselleştirilmesi için bir bölüm, farklı algoritmaların sonuçlarının otomatik olarak karşılaştırılabilmesi için bir bölüm.

EKLER

Ek-1. MISFP-Growth (Darrab ve Ergenç, 2016)

Ek-2. Sadeq Darrab'ın Yüksek Lisans Tezi

Ek-3. MIS-eclat (Darrab ve Ergenç, 2017)

Ek-4. Dynamic MIS (Abuzayed ve Ergenç, 2016)

Ek-5. Nourhan Abuzayed'in Yüksek Lisans Tezi

Ek-6. Dynamic CFP-Growth (Abuzayed ve Ergenç, 2017)

Ek-7. PGBS (Öztürk ve Ergenç, 2017)

Ek-8. Dynamic PGBS (Öztürk ve Ergenç, 2018a)

Ek-9. IPGBS (Öztürk ve Ergenç, 2018b)

KAYNAKLAR

Abul O. 2009. "Hiding Co-Occuring Frequent Itemsets", In: 2nd International Workshop on Privacy and Anonymity in the Information Society (PAIS), St. Petersburg, Russia.

Abuzayed N, Ergenç B. 2016. "Dynamic Itemset Mining under Multiple Support Thresholds". Fuzzy Systems and Data Mining II, IOS Press, pp.141-148.

Abuzayed N, Ergenç B. 2017. "Comparison of Dynamic Itemset Mining Algorithms for Multiple Support Thresholds". 21st International Database Engineering and Applications Symposium IDEAS 2017, Byte Press, 309-316, pp. 12-14 July, Bristol, England.

Agrawal R, Imielinski T, Swami A. 1993. "Mining Association Rules Between Sets of Items in Large Databases". ACM SIGMOD Record; 22(2): 207-216.

Agrawal R, Srikant R. 1994. "Fast Algorithms For Mining Association Rules". 20th VLDB 1994; 1215: pp. 487-499.

Agrawal C, Han J. 2014. "Frequent Pattern Mining". Springer.

Alhajj M.R, Barker K. 2008. "Alternative Method for Incrementally Constructing the FP-Tree". Studies in Computational Intelligence 109, 361–377.

Amiri A. 2007. "Dare to Share: Protecting Sensitive Knowledge with Data Sanitization". Decision Support Systems, 43(1), pp.181–191. doi:10.1016/j.dss.2006.08.007

Amornchewin R, Kreesuradej W. 2007. "Incremental Association Rule Mining Using Promising Frequent Itemset Algorithm". In Proceedings of the 6th International Conference on Information, Communications and Signal Processing, Singapore.

Atallah M, Bertino E, Elmagarmid A, Ibrahim M, Verykios V.S. 1999. "Disclosure Limitation Of Sensitive Rules". In Workshop on Knowledge and Data Engineering Exchange, pp. 45-52.

Ayan N, Tansel A, Arkun M. 1999. "An Efficient Algorithm to Update Large Itemsets with Early Pruning". In Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, 287–291.

Ayav T, Ergenç B. 2015. "Full Exact Approach For Itemset Hiding". International Journal of Data Warehousing and Mining, 11(4), pp. 49-63.

Blake CL, Merz, CJ. 1998. "UCI Repository of Machine Learning Databases". University of California, Irvine, Department of Information and Computer Science.

Boora R.K, Shukla R, Misra A. 2009. "An İmproved Approach To High Level Privacy Preserving Itemset Mining", International Journal of Computer Science and Information Security. 6(3), pp. 216-223.

Brijs T, Swinnen G, Vanhoof K, Wetz G. 1998. "Using Association Rules For Product Assortment Decisions: A Case Study". In Knowledge Discovery and Mining, pp. 254-260.

Chaudhary V. 2014. "Multiple Minimum Support Implementations with Dynamic Matrix Apriori Algorithm For Efficient Mining of Association Rules". International journal for Scientific Research and Development 2(7), 489 –500.

Cheng P, Roddick, J.F, Chu, S.C. 2016. "Privacy Preservation Through A Greedy, Distortion-Based Rule Hiding Method". Applied Intelligence, vol.44, pp. 295-306.

Cheungt D. W, Han J, Ng V. T, Wong C. Y. 1996. "Maintenance Of Discovered Association Rules İn Large Databases, An İncremental Updating Technique", In: The 12th International Conference on Data Engineering. pp. 106–114.

Cheung W, Zaïane O. R. 2003. "Incremental Mining of Frequent Patterns Without Candidate Generation or Support Constraint", In: The 7th International Database Engineering and Applications Symposium. pp. 111–116.

Dai B. R, Chiang L. H. 2010. "Hiding Frequent Patterns In The Updated Database". International Conference on Information Science and Applications (ICISA). doi:10.1109/ICISA.2010.5480385

Darrab S, Ergenç B. 2016. "Frequent Pattern Mining under Multiple Support Thresholds". Wseas Transactions on Computer Research, 4

Darrab S, Ergenç B. 2017. "Vertical Pattern Mining Algorithm for Multiple Support Thresholds". 21st International Conference on Knowledge-Based and Intelligent Information & Engineering Systems, KES2017, 6-8 September, Marseille, France

Dehkordi M.S, Dehkordi M.N. 2016. "Introducing an Algorithm for Use to Hide Sensitive Association Rules Through Perturbation Technique". Journal of AI and Data Mining, doi: 10.5829/- IDOSI.JAIDM.2016.04.02.10

Ezeife C, Su Y. 2002. "Mining Incremental Association Rules with Generalized FP-Tree". Proceedings of the 15th Conference of the Canadian Society for Computational Studies of Intelligence on Advances in Artificial Intelligence, Calgary, Canada, pp. 147–160.

Gan W, Lin J, Viger P, Chao H. 2016. "More Efficient Algorithm for Mining Frequent Patterns with Multiple Minimum Supports". International Conference on Web-Age Information Management, pp. 3-16.

Garg V, Singh A, Singh D. 2014. "A Hybrid Algorithm for Association Rule Hiding Using Representative Rule". International Journal of Computer Applications, vol.97,. doi: 10.1007/978-3-319-07455-9_9

Ghanem A, Sallam H. 2011. "Hybrid Search Based Association Rule Mining". IEEE Pacific Rim Conference on Communications, Computers and Signal Processing.

Gkoulalas-Divanis A, Verykios V.S. 2006. "An İnteger Programming Approach For Frequent İtemset Hiding". In 15th ACM International Conference on Information and Knowledge Management. doi:10.1145/1183614.1183721

Gkoulalas-Divanis A, Verykios V.S. 2008. "A Parallelization Framework for Exact Knowledge Hiding İn Transactional Databases". In IFIP International Federation for Information Processing, vol. 278, (pp. 349-363). doi:10.1007/978-0-387-09699-5_23

Gkoulalas-Divanis A, Verykios V.S. 2009. "Hiding Sensitive Knowledge Without Side Effects". Knowledge and Information Systems, 20(3), pp. 263-299. doi:10.1007/s10115-008-0178-7

Grahne G, Zhu J. 2005. "Fast Algorithms for Frequent Itemset Mining Using FP-Trees". Knowledge and Data Engineering; 17(10): pp. 1347-1362.

Guo Y. 2007. "Reconstruction Based Association Rule Hiding". In SIGMOD Ph.D. Workshop on Innovative Database Research. http://www.borgelt.net/apriori.html

Han J, Pei J, Yin Y. 2000. "Mining Frequent Patterns Without Candidate Generation". ACM Sigmod Record, 29(2).

Hoque F.A, M. Debnath, M, Easmin, N, Rashad, k. 2011. "Frequent Pattern Mining For Multiple Minimum Supports With Support Tuning And Tree Maintenance On İncremental Database". Research Journal of Information Technology, 3(2), 79–90.

Hong T-P, Lin C-W, Yang K-T, Wang S-L. 2013. "Using Tf-Idf to Hide Sensitive Itemsets". Applied Intelligence, 38(4), pp. 502–510. doi:10.1007/s10489-012-0377-5

Hu Y, Chen Y. 2006. "Mining Association Rules with Multiple Minimum Supports: A New Mining Algorithm And A Support Tuning Mechanism". Decision Support Systems; 42(1):1-24.

Jadav K.B, Vania J, Patel D.R. 2014. "Efficient Hiding of Sensitive Association Rules For Incremental Datasets". International Journal of Innovations & Advancement in Computer Science IJIACS, ISSN 2347 – 8616, 3(4).

Jalan S, Srivastava A, Sharma G. 2009. "A Non-Recursive Approach For FP-Tree Based Frequent Pattern Generation". Research and Development (SCOReD: pp. 160-163.

Jayasudha V. 2013. "EUP-Growth+ - Efficient Algorithm for Mining High Utility Itemset". International Journal of Engineering Research and Development, 9(5), pp. 38-50.

Kanimozh C.S, Tamilarasi A. 2009. "Mining Association Rules with Dynamic And Collective Support Thresholds". International Journal of Engineering and Technology, 1(3).

Keer S, Singh A., (2012). "Hiding Sensitive Association Rule Using Clusters Of Sensitive Association Rule". International Journal of Computer Science and Network, 1(3).

Kiran R, Uday P, Krishna R. 2011. "Novel Techniques to Reduce Search Space in Multiple Minimum Supports-Based Frequent Pattern Mining Algorithms". Proceedings of the 14th International Conference on Extending Database Technology. ACM.

Kuo Y, Lin P.Y, Dai B.R. 2008. "Hiding Frequent Patterns Under Multiple Sensitive Thresholds". Database and Expert Systems Applications (DEXA), (pp. 5-18). doi: 10.1007/978-3-540-85654-2_2

Lee S.D, Cheung D.W, Kao B. 1998. "Is Sampling Useful in Data Mining? A Case in the Maintenance of Discovered Association Rules". Data mining and Knowledge Discovery 2(3),pp. 233– 262.

Li Y.C, Yeh J.S, Chang C.C. 2007. "MICF: An Effective Sanitization Algorithm for Hiding Sensitive Patterns on Data Mining". Advanced Engineering Informatics, vol 21, pp. 269–280

Lin J, Liu J.Y.C. 2007. "Privacy Preserving Itemset Mining Through Fake Transactions". In 22nd ACM Symposium on Applied Computing, SAC, (pp.375–379), Seoul, Korea. doi:10.1145-/10.1145/1244002.1244092

Liu B, Hsu W, Ma Y. 1999. "Mining Association Rules with Multiple Minimum Supports". ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 337-341.

Menon S, Sarkar,S, Mukherjee S. 2005. "Maximizing Accuracy of Shared Databases When Concealing Sensitive Patterns". Information Systems Research 16(3), pp. 256-270.

Moustakides G.V, Verykios V.S. 2008. "A Maxmin Approach For Hiding Frequent Itemsets". Data and Knowledge Engineering, 65(1), pp. 75-89.

Nhan L, Thuy N, Chung T. 2010. "Bitapriori: An Apriori-Based Frequent Itemsets Mining Using Bit Streams". International Conference on Information Science and Applications (ICISA); pp. 1-6.

Oğuz D, Ergenç B. 2012. "Incremental Itemset Mining Based on Matrix Apriori", DEXA-DaWaK, Vienna, Austria, September 3-6, LNCS Volume 7448, pp.192-204.

Oliveira S.R.M, Zaiane O.R. 2002. "Privacy Preserving Frequent Itemset Mining". In International Conference on Data Mining (ICDM), pp. 43-54, Maebashi City, Japan.

Oliveira S.R.M, Zaiane O.R. 2003. "Algorithms For Balancing Privacy And Knowledge Discovery İn Association Rule Mining". In 7th International Database Engineering & Applications Symposium, (pp. 54–63). doi:10.1109/IDEAS.2003.1214911

Öztürk A.C, Ergenç B. 2017. "Itemset Hiding Under Multiple Sensitive Support Thresholds". 9th International Conference on Knowledge Management and Information Sharing, KMIS2017, 1-3 November, Funchal, Madeira, Portugal

Öztürk A.C, Ergenç B. 2018a. "Dynamic Itemset Hiding Algorithm for Multiple Sensitive Support Thresholds". International Journal of Data Warehousing and Mining. 11(2).

Öztürk A.C, Ergenç B. 2018b. "Minimizing Information Loss in Shared Data: Hiding Frequent Patterns with Multiple Sensitive Support Thresholds". Turkish Journal of Electrical Engineering and Computer Science. (waiting for decision).

Park J, Chen M, Yu P. 1997. "Using A Hash-Based Method With Transaction Trimming For Mining Association Rules". Knowledge and Data Engineering, 9(5): pp. 813-825.

Pavón J, Viana S, Gómez S. 2006. "Matrix Apriori: Speeding up the Search For Frequent Patterns". Databases and Applications, pp. 75-82.

Pontikakis E.D, Tsitsoni, A.A, Verykios V.S. 2004. "An Experimental Study Of Distortion-Based Techniques For Association Rule Hiding". In 8th Annual Conference on Data and Applications Security, vol 144. Catalonia, Spain. doi: 10.1007/1-4020-8128-6_22

Sinthuja M, Rachel S, Janani G. 2011. "MIS-Tree Algorithm for Mining Association Rules with Multiple Minimum Supports". Bonfring International Journal of Data Mining, 1: 1-5.

Stavropoulos E.C, Verykios V.S, Kagklis V. 2016. "A Transversal Hypergraph Approach for the Frequent Itemset Hiding Problem". Knowledge and Information Systems, pp. 625-645.

Sun X, Yu P.S. 2005. "A Border-Based Approach for Hiding Sensitive Frequent Itemsets", In: 5th IEEE International Conference on Data Mining (ICDM), pp. 426-433.

Sun X, Yu P.S. 2007. "Hiding Sensitive Frequent Itemsets By A Border-Based Approach". Computing Science and Engineering 1(1), pp.74-94.

Taha M, Gharib T, Nassar H. 2011. "DARM: Decremental Association Rules Mining" Journal of Intelligent Learning Systems and Applications, 3, pp. 181–189.

Thieme L. 2004. "Algorithmic Features Of Eclat". FIMI.

Vaidya J. 2001. "Privacy Preserving Association Rule Mining in Vertically Partitioned Data", In: The Eight ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 639-644.

Verykios V. S, Emagarmid A. K, Bertino E, Saygin Y, Dassen, E. 2004. "Association Rule Hiding". IEEE Transactions on Knowledge and Data Engineering, 16(4), pp. 434–447. doi:10.11-.09/TKDE.2004.1269668

Weng C, Chen S, Che Lo H. 2008. "A Novel Algorithm for Completely Hiding Sensitive Association Rules". In 8th International Conference on Intelligent Systems Design and Applications. doi: 10.1109/ISDA.2008.180

Wu YH, Chiang CM, Chen A. 2007. "Hiding Sensitive Association Rules With Limited Side Effects". IEEE Transactions on Knowledge and Data Engineering, 19(1), pp. 29-42.

Xu T, Dong X. 2013. "Mining Frequent Patterns with Multiple Minimum Supports Using Basic Apriori". In Ninth International Conference on Natural Computation (ICNC), 957-961.

Yıldız B, Ergenç B. 2010. "Comparison of Two Association Rule Mining Algorithms Without Candidate Generation". 10th IASTED: 450–457.

Zaki M. 2000. "Scalable Algorithms for Association Mining". IEEE Transactions on Knowledge and Data Engineering, 12(3): 372-390.

TÜBİTAK
PROJE ÖZET BİLGİ FORMU

Proje Yürütücüsü:	Doç. Dr. BELGİN ERGENÇ
Proje No:	114E779
Proje Başlığı:	Dfis-Çoklu Destek Eşiklerinde Dinamik Sık Kümeler Madenciliği Ve Gizleme Platformu
Proje Türü:	3501 - Kariyer
Proje Süresi:	36
Araştırmacılar:	
Danışmanlar:	
Projenin Yürütüldüğü Kuruluş ve Adresi:	İZMİR YÜKSEK TEKNOLOJİ ENSTİTÜSÜ
Projenin Başlangıç ve Bitiş Tarihleri:	15/04/2015 - 15/04/2018
Onaylanan Bütçe:	216888.0
Harcanan Bütçe:	143859.47
Öz:	Bu proje kapsamında, veri madenciliği alanının en çok kullanılan yöntemi olan, ilişki kuralları (association rules) madenciliğinin başetmeye çalıştığı zorluklardan, veri büyüklüğü, veri dinamizmi, sık kümelerin (frequent itemsets) özel destek eşik (support threshold) değerlerinin dikkate alınması ve paylaşımında ortaya çıkabilecek duyarlı (sensitive) bilgilerin gizlenmesi (sensitive knowledge hiding) problemleri ile aynı anda uğraşan sınama platformunun geliştirilmesi hedeflenmektedir. Önerilecek olan platformdaki temel (baseline) ilişki kuralı madenciliği işlevi veri büyüklüğü ile başedebilmek için veritabanını çoklu taramayacak, kolay yönetilebilir veri tipleri kullanacak ve etkin bellek kullanımı yapacaktır. Söz konusu işlev, tüm platform için tek bir destek eşik değeri ile çalışmak yerine veri kümelerine özel destek eşik değerleri ile çalışabilir olacaktır. Platform parçalarından biri de temel ilişki kuralı madenciliği işlevinin dinamik sürümüdür; bu sürüm veri güncellemeleri geldiğinde tüm ilişki kuralı bulma sürecini baştan çalıştırmak yerine, güncellemeyi içeren veritabanı parçası ve önceki sonuçları dikkate alarak güncel sık kümeleri dinamik olarak bulur. Platform son olarak veritabanını, duyarlı bilgi çıkarımları yapılamayacak halde paylaşmaya hazırlayabilecek yani dinamik sık küme gizleme (itemset hiding) işlevi içermektedir.
Anahtar Kelimeler:	sık kümeler, dinamik veri, çoklu destek eşikleri, veri paylaşımı, bilgi gizleme, sık küme gizleme
Fikri Ürün Bildirim Formu Sunuldu Mu?:	Hayır
Projeden Yapılan Yayınlar:	1- Dynamic Itemset Mining under Multiple Support Thresholds (Bildiri - Uluslararası Bildiri - Sözlü Sunum), 2- Frequent Pattern Mining under Multiple Support Thresholds (Bildiri - Uluslararası Bildiri - Sözlü Sunum), 3- Dynamic Itemset Mining Under Multiple Supoort Thresholds (Tez (Araştırmacı Yetiştirilmesi) - Yüksek Lisans Tezi), 4- Development of a Framework for Frequent Itemset Mining under Multiple Support Thresholds (Tez (Araştırmacı Yetiştirilmesi) - Yüksek Lisans Tezi), 5- Dynamik Itemset Hiding Algorithm for Multiple Sensitive Support Thresholds (Makale - Diğer Hakemli Makale), 6- Vertical Pattern Mining under Multiple Support Thresholds (Bildiri - Uluslararası Bildiri - Sözlü Sunum), 7- Itemset Hiding under Multiple Sensitive Support Thresholds (Bildiri - Uluslararası Bildiri - Sözlü Sunum), 8- Comparison of Dynamic Itemset Mining Algorithms for Multiple Items Support Thresholds (Bildiri - Uluslararası Bildiri - Sözlü Sunum),