T.C.
ERCİYES ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

HİDROLOJİK VE METEOROLOJİK VERİTABANLARINDA HİDRO-METEOROLOJİK BİRLİKTELİK ÖRÜNTÜLERİNİN KEŞFİ

Hazırlayan
Ahmet Şakir DOKUZ

Danışman
Yrd. Doç. Dr. Mete ÇELİK
Yrd. Doç. Dr. Filiz DADAŞER ÇELİK

Yüksek Lisans Tezi

Temmuz 2013
KAYSERİ

T.C.
ERCİYES ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

HİDROLOJİK VE METEOROLOJİK VERİTABANLARINDA HİDRO-METEOROLOJİK BİRLİKTELİK ÖRÜNTÜLERİNİN KEŞFİ

(Yüksek Lisans Tezi)

Hazırlayan
Ahmet Şakir DOKUZ

Danışman
Yrd. Doç. Dr. Mete ÇELİK
Yrd. Doç. Dr. Filiz DADAŞER ÇELİK

Bu çalışma, TÜBİTAK (Proje Kodu: ÇAYDAG 110Y110) tarafından desteklenmiştir.

Temmuz 2013
KAYSERİ

BİLİMSEL ETİĞE UYGUNLUK

Bu çalışmadaki tüm bilgilerin, akademik ve etik kurallara uygun bir şekilde elde edildiğini beyan ederim. Aynı zamanda bu kural ve davranışların gerektirdiği gibi, bu çalışmanın özünde olmayan tüm materyal ve sonuçları tam olarak aktardığımı ve referans gösterdiğimi belirtirim.



Ahmet Şakir DOKUZ

"Hidrolojik ve Meteorolojik Veritabanlarında Hidro-Meteorolojik Birliktelik Örüntülerinin Keşfi" adlı yüksek lisans tezi, Erciyes Üniversitesi Lisansüstü Tez Önerisi ve Tez Yazma Yönergesi'ne uygun olarak hazırlanmıştır.

Tezi Hazırlayan
Ahmet Şakir DOKUZ

Danışman
Yrd. Doç. Dr. Mete ÇELİK

Bilgisayar Mühendisliği ABD Başkanı

Prof. Dr. Derviş KARABOĞA

Yrd. Doç. Dr. Mete ÇELİK danışmanlığında **Ahmet Şakir DOKUZ** tarafından hazırlanan **"Hidrolojik ve Meteorolojik Veritabanlarında Hidro-Meteorolojik Birliktelik Örüntülerinin Keşfi"** adlı bu çalışma, jürimiz tarafından Erciyes Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalında **Yüksek Lisans** tezi olarak kabul edilmiştir.

03/07/2013

JÜRİ:

Danışman : Yrd. Doç. Dr. Mete ÇELİK

Üye :Prof. Dr. Derviş KARABOĞA

Üye :Yrd. Doç. Dr. Serkan ÖZTÜRK

ONAY:

Bu tezin kabulü Enstitü Yönetim Kurulunun 09/07/2013.. tarih ve 2013/30-07 sayılı kararı ile onaylanmıştır.

09 / 07 / 2013

Doç. Dr. Hasan YALÇIN

Enstitü Müdürü

TEŞEKKÜR

Çalışmalarım boyunca bilimsel yaklaşımları ve yol göstericilikleriyle beni aydınlatan, tez çalışmalarım boyunca yakın ilgi ve yardımlarını esirgemeyen ve akademik çalışmayı bana öğreten çok değerli danışmanlarım Yrd. Doç. Dr. Mete ÇELİK ve Yrd. Doç. Dr. Filiz DADAŞER ÇELİK'e teşekkürü bir borç bilirim.

Çalışmalarım süresince sabır göstererek beni daima destekleyen sevgili eşim Yeşim DOKUZ'a en içten teşekkürlerimi sunarım.

Hayatım boyunca beni destekleyen, okutan, üzerimde emeği olan anne ve babama en içten teşekkürlerimi sunarım.

Ayrıca lisans ve yüksek lisans eğitimim süresince emeği geçen, başta Prof. Dr. Derviş KARABOĞA olmak üzere tüm hocalarıma teşekkürlerimi sunarım.

Bu tez çalışmasına maddi destek veren Türkiye Bilimsel ve Teknik Araştırma Kurumu'na (Proje No: ÇAYDAG 110Y110) teşekkür ederim. Çalışmada kullanılan verileri sağlayan Meteoroloji Genel Müdürlüğü ve Elektrik İşleri Etüd İdaresi'ne teşekkür ederim.

Çevre Yüksek Mühendisli Eda CENGİZ'e veri kümesi üzerindeki istatistiksel testler ve değerlendirmelerdeki yardımlarından dolayı teşekkür ederim.

Ayrıca Erciyes Üniversitesi Veri Madenciliği Araştırma Grubu üyelerine yorum ve eleştirilerinden dolayı teşekkür ederim.

Ahmet Şakir DOKUZ

Kayseri, Temmuz 2013

HİDROLOJİK VE İKLİMSEL VERİTABANLARINDA HİDRO-METEOROLOJİK BİRLİKTELİK ÖRÜNTÜLERİNİN KEŞFİ

Ahmet Şakir DOKUZ

Erciyes Üniversitesi, Fen Bilimleri Enstitüsü

Yüksek Lisans Tezi, Temmuz 2013

Danışman: Yrd. Doç. Dr. Mete ÇELİK

KISA ÖZET

Hidro-meteorolojik birliktelik örüntü madenciliği, her bir kayda ait hidrolojik ve meteorolojik verilerin bulunduğu veritabanları üzerinde, veritabanları içerisindeki parametreler arasındaki anlamlı birliktelik örüntülerini ortaya çıkarmada kullanılan bir veri madenciliği yöntemidir. Mekân-zamansal birliktelik örüntü madenciliği, mekân ve zaman bilgisi içeren her türlü veri üzerinde uygulanabilir olmakla birlikte, sıklıkla, hidrolojik ve meteorolojik parametreler arasındaki mekân-zamansal birliktelikleri ortaya çıkarmada, telekomünikasyon hatalarının tespitinde, salgın hastalıkların yayılımlarının tespitinde, DNA gen dizilimlerinin belirlenmesinde kullanılır.

Veritabanlarının tekrar tekrar taranması ve yüksek hesaplama maliyetleri sebebiyle, hidro-meteorolojik birliktelik analizi algoritmaları yavaş çalışmaktadır. Ayrıca örüntülerin keşfedilmesinde değerlendirme ölçütlerinin çalışmaya uygunluğu da algoritma sonucunu etkileyen bir unsur olmaktadır. Bu tez çalışmasında, hidro-meteorolojik birliktelik örüntülerini keşfetmek için kullanılan algoritmalar incelenmiş, yeni bir algoritma önerilmiş ve yeni bir değerlendirme ölçütü tanımlanmıştır. Geliştirilen algoritma Türkiye'ye ait hidrolojik ve iklimsel veritabanlarında uygulanmış ve hidro-meteorolojik birliktelik örüntüleri tespit edilmiştir.

Anahtar Kelimeler: Mekân-zamansal birliktelik örüntü madenciliği; veri madenciliği; birliktelik analizi; mekân-zamansal veritabanları; hidro-meteorolojik örüntü madenciliği; değerlendirme ölçütleri.

HYDRO-METEOROLOGICAL ASSOCIATION PATTERN DISCOVERY ON HYDROLOGICAL AND CLIMATOLOGICAL DATABASES

Ahmet Şakir DOKUZ

Erciyes University, Graduate School of Natural and Applied Sciences

M.Sc. Thesis, July 2013

Supervisor: Asist. Prof. Dr. Mete ÇELİK

ABSTRACT

Hydro-meteorological association pattern mining is a data mining approach to extract meaningfull association patterns over parameters in a database which stores hydrological and meteorological information for every record in the database. Spatio-temporal association pattern mining can be applied on every data which have spatial and temporal aspects. It is mostly used on discovery of spatio-temporal relationships on climatic parameters, telecommunication fault detection, the spread of epidemic diseases over spatial and temporal aspects.

Because of the repeatitive database scan and high calculation costs, hydro-meteorolgical association analysis algorithms are computationally complex. Also, interest measures, which are used to evaluate patterns, is an important factor on spatio-temporal association pattern mining. In this thesis, hydro-meteorological association pattern discovery algorithms were analyzed, a new algorithm and a new interest measure were proposed to evaluate results. Proposed algorithm was applied on hydrological and meteorological database in Turkey and hydro-meteorological association patterns were discovered.

Keywords: Spatio-temporal association pattern mining; data mining; association analysis; spatio-temporal databases; hydro-meteorological pattern mining; interest measures.

İÇİNDEKİLER

HİDROLOJİK VE İKLİMSEL VERİTABANLARINDA HİDRO-METEOROLOJİK BİRLİKTELİK ÖRÜNTÜLERİNİN KEŞFİ

Sayfa

BİLİMSEL ETİĞE UYGUNLUK SAYFASI ..... **ii**
YÖNERGEYE UYGUNLUK SAYFASI ..... **iii**
KABUL VE ONAY SAYFASI ..... **iv**
TEŞEKKÜR ..... **v**
KISA ÖZET ..... **vi**
ABSTRACT ..... **vii**
İÇİNDEKİLER ..... **viii**
TABLOLAR LİSTESİ ..... **xi**
ŞEKİLLER LİSTESİ ..... **xii**

GİRİŞ **1**

1. BÖLÜM

GENEL BİLGİLER

1.1.Veri Madenciliği **5**
1.1.1. Veri Madenciliğinin Önemi **6**
1.2.Veri Madenciliğinin Adımları **7**
1.3.Veri Madenciliğinin Avantaj ve Dezavantajları **8**
1.4.Veri Madenciliğinin Temel Zorlukları **9**
1.4.1. Metotları Açısından **9**
1.4.2. Kullanıcı Etkileşimi Açısından **10**
1.4.3. Etkinlik ve Ölçeklenebilirlik Açısından **10**
1.4.4. Veritabanı Türlerinin Değişkenliği Açısından **11**
1.5.Veri Madenciliğinin İlgilendiği Konular **11**
1.5.1. Birliktelik Analizi **11**

1.5.2. Sınıflandırma **13**
1.5.3. Kümeleme **15**
1.5.4. Anormallik Tespiti **17**
1.6.Mekân-Zamansal Birliktelik Analizi **17**
1.7.Mekân-Zamansal Birliktelik Analizi Değerlendirme Ölçütleri **19**
1.8.Literatür Taraması **21**
1.8.1. Mekân-Zamansal Birliktelik Analizi **21**
1.8.2. Hidrolojik ve Meteorolojik Veritabanlarında Veri Madenciliği Uygulamaları **23**

2. BÖLÜM

HİDRO-METEOROLOJİK BİRLİKTELİK ÖRÜNTÜ MADENCİLİĞİ

2.1. Temel Tanımlar **26**
2.1.1. Birliktelik Kuralı **26**
2.1.2. Destek Ölçütü **27**
2.1.3. İstasyon içi Örüntü **27**
2.1.4. İstasyon içi Yaygın Örüntü **27**
2.1.5. İstasyonlar arası Örüntü **27**
2.1.6. İstasyon Yaygınlık Ölçütü **27**
2.1.7. İstasyonlar arası Yaygın Örüntü **28**
2.1.8. Hidro-meteorolojik Birliktelik Örüntüsü **28**
2.1.9. Hidro-meteorolojik Yaygınlık İndeksi **28**
2.1.10. Hidro-meteorolojik Yaygın Örüntü **29**
2.1.11. Güven Ölçütü **29**
2.2.Hidro-meteorolojik Birliktelik Örüntü Madenciliği Problem Tanımlaması **29**
2.3.Hidro-meteorolojik Birliktelik Örüntü Madenciliği **30**
2.3.1. Toy (Naive) Yaklaşım **30**
2.3.2. HMP-Miner Algoritması **34**
2.4.HMP-Miner Algoritmasına Ait Örnek Çalışma **37**

3. BÖLÜM

DENEYSEL ÇALIŞMA

3.1. Deney Düzeneği **44**
3.2. Kullanılan Veri Kümesi **45**
3.3. Veri Üzerinde Yapılan İşlemler **46**
3.3.1. Veri Seçme **46**
3.3.2. Verilerin Aylık Forma Dönüştürülmesi **47**
3.3.3. Ayrıklaştırma **47**
3.4. Deneyler **48**
3.4.1. Minimum Destek Eşik Değerinin Etkisi **48**
3.4.2. Minimum İstasyon Eşik Değerinin Etkisi **49**
3.4.3. İstasyon Sayısının Etkisi **51**
3.5. Deney Sonuçları **51**

4. BÖLÜM

SONUÇ, TARTIŞMA VE ÖNERİLER

4.1. Sonuç **55**
4.2. Tartışma **56**
4.3. Gelecekteki Çalışmalar **57**

KAYNAKÇA **58**
ÖZGEÇMİŞ **63**

TABLOLAR LİSTESİ

Tablo 2.1. Toy yaklaşıma ait kaba kod 32
Tablo 2.2. HMP-Miner algoritmasına ait kaba kod 36
Tablo 2.3. Örnek çalışma için kullanılacak olan veri kümesi 37
Tablo 2.4. 1 boyutlu örüntülerin üretimi 39
Tablo 2.5. 2 boyutlu aday örüntüler 39
Tablo 2.6. 2 boyutlu yaygın örüntüler 41
Tablo 2.7. 3 boyutlu aday örüntüler 42
Tablo 2.8. 4 boyutlu aday örüntüler 42
Tablo 2.9. Güven-tabanlı eleme ve HMP-Miner algoritmasının çıktısı 43
Tablo 3.1. HMP-Miner algoritmasının bulduğu kurallar 53

ŞEKİLLER LİSTESİ

Şekil 1.1. Veri madenciliği adımları 7
Şekil 1.2. Apriori metoduna ait kafes yapısı ve elenen elemanlar 12
Şekil 1.3. FP-Tree metoduna ait çalışma şeması 13
Şekil 1.4. Literatür taramasında tez çalışmasının yeri 21
Şekil 2.1. Toy (naive) yaklaşımın çalışma mantığı 31
Şekil 2.2. HMP-Miner algoritmasına ait çalışma mantığı 35
Şekil 3.1. Deney düzeneği 45
Şekil 3.2. Akım ölçüm ve meteoroloji istasyonların konumu 46
Şekil 3.3. Akım parametresine ait örnek bir ayrıklaştırma 48
Şekil 3.4. Minimum destek eşik değerinin etkisi 49
Şekil 3.5. Minimum istasyon eşik değerinin etkisi 50
Şekil 3.6. İstasyon sayısının etkisi 51
Şekil 3.7. HMP-Miner algoritmasının bulduğu kurallar 54

GİRİŞ

Bilişim teknolojilerindeki gelişmelerle birlikte, günümüzde, bilgisayarlar çok karmaşık işlemleri son derece hızlı ve rahatlıkla yapabilecek seviyeye gelmiştir. Bu gelişmelere paralel olarak, bilgisayarlar artık veri depolama sistemleri de sağlamaya başlamıştır. Bilgisayarlardan önce veriler dosyalar ve klasörlerde fiziksel olarak tutuluyordu. Bilgisayarların hızlı işlem yapabilme yeteneğiyle birlikte bu dosyalama işlemleri de dijital ortama geçirilmeye başlanmıştır. Günümüzde artık çok gizli bilgiler dahi bilgisayarlarda tutulur hale getirilmiştir.

Veri depolama sistemlerinde, başlangıçta veriler dosyalarda tutulmaktaydı. İlk sistemlerde, birbiriyle ilişkili dosyaların olmaması, veri tekrarının çok büyük boyutlarda olmaması ve veri boyutlarının küçük olması sebebiyle dosyalama işlemi yeterli olmaktaydı. Ancak zamanla veri boyutlarının ve birbiriyle bağlantılı verilerin artmasıyla, birbiriyle bağlantılı dosyaların ayrıştırılması, bu dosyalardaki verilerin birbirleriyle ilişkilerinin incelenmesinde gereksiz pek çok ara işlemlere ihtiyaç duyulması gibi sorunları ortaya çıkarmıştır. Bu durumda, artık klasik dosyalama sistemleri yeterli olmaktan uzaklaşmış ve etkin bir şekilde veri işleyebilecek bir sistem ihtiyacı ortaya çıkmıştır.

Veritabanı sistemleri, bu ihtiyaçlar çerçevesinde şekillenmiştir. Veritabanı sistemleri, her ne kadar altyapısal olarak dosyalama sistemlerini kullansa da, verileri ağaçlar gibi veri yapısı sistemlerinde tutmaları nedeniyle, çok daha hızlı ve verimli bir şekilde saklayabilmekte, erişebilmekte ve işleyebilmektedir. Veritabanına kaydedilen bir veri grubu, herhangi bir ara işleme gereksinim duyulmadan, her türlü erişim ve işleme hazır hale gelmektedir. Bu veriler üzerinde, birbirinden farklı veritabanı işlemleri yapılabileceği gibi, çok çeşitli algoritmalar çalıştırılarak da veriler işlenebilmektedir.

Veritabanı sistemleri klasik sorguları ile belirli sorulara beklenen cevapları verebilmektedir. Verilerin doğası, tek başına tek bir bilgiyi içermek olsa da, verilerin grup halinde iken görülmeyen, tahmin edilemeyen, grupta aykırı olan bilgiler de içermektedir. Veri madenciliği, bir veritabanı üzerinde, veritabanındaki veriler arasındaki ilişkileri inceleyen bir bilgisayar mühendisliği alt dalıdır. Verilerin veri grupları içerisindeki davranışlarını incelemek için veri madenciliği çeşitli metotlar ve algoritmalar sunmaktadır. Klasik veritabanı sorguları yerine, sıradan işlemlerle ortaya çıkarılamayacak, ilk bakışta görülemeyen, ilginç, anlamlı örüntüler elde etmeyi amaçlayan veri madenciliği; istatistik, makineli öğrenme ve veritabanı sistemlerinin bir araya getirilmesiyle ortaya çıkmıştır.

Tezin Kapsamı ve Önemi

Veri madenciliği, en genel tabiriyle, büyük ölçekli veriler arasından bilgi çıkarımı işlemidir. Bir diğer deyişle veri madenciliği, çok büyük boyutlu ve karmaşık veritabanları üzerinde, manuel olarak çok uzun sürecek veya klasik sorgular yardımıyla elde edilemeyecek olan gözle görülemeyen, ilginç ve anlamlı örüntülerin keşfi işlemidir.

Veri madenciliği, son yıllarda artan boyuttaki veritabanları için çok etkin ve hızlı çalışan algoritmalar önermektedir. Veri üzerindeki dolanım sayısı, yapılan işlem sayısı ve gerektirdiği hafıza alanları ile veri madenciliği algoritmaları diğer yöntemlerden üstünlüğünü ortaya koymaktadır. Bu açılardan yapay zeka, makineli öğrenme ve özellikle istatistiksel yöntemler yerine veri madenciliği algoritmaları daha yaygın olarak kullanılmaya başlanmıştır.

Klasik veritabanı verilerinden ayrı olarak verilerin ait olduğu mekana ve zamana dayalı bilgilerin tutulduğu veritabanları üzerinde mekan ve zaman bilgisine bağlı kalarak bazı sonuçlar çıkarmak gerekmektedir. Buna en iyi örnek Akdeniz Bölgesi için yaz mevsiminde 30 °C sıcaklık normal iken kış mevsiminde bu sıcaklık değeri aşırı yüksek olmaktadır. Veya Karadeniz Bölgesi için yaz döneminde 30 °C sıcaklık normal iken Güneydoğu Anadolu Bölgesi için bu sıcaklık düşük olarak değerlendirilmektedir. Doğru bir sonuç çıkarımında bulunmak için bu veriyi ait olduğu mekana ve içerisinde bulunduğu zamana göre analiz etmek gerekmektedir.

Tezin Amacı

Son yıllarda artan bilişim teknolojileri (gözlem, sensör teknolojileri gibi) yardımıyla, zamana dayalı verilerin de toplanması kolaylaşmış ve otomatik hale getirilmiştir. Bu teknolojiler kullanılarak oluşturulan veritabanları ise hızlı bir şekilde büyüdüğü için, bu verilerin analizi için otomatik yöntemler geliştirmek ihtiyacı ortaya çıkmıştır. Ekstrem iklim olaylarının tespitinde, uzay gözlem verilerinin analizinde, telekomünikasyon hatalarının tespitinde zamana dayalı analizler yapmak gerekmektedir. Veri madenciliği de bu problemleri çözmek için kullanılan yöntemlerden bir tanesidir. İçerdiği mekan-zamansal veritabanı analizi algoritmaları yardımıyla etkili ve hızlı bir şekilde sonuçlar elde edebilmektedir.

Bu tez çalışmasının hedefi, bir hidrolojik ve meteorolojik veritabanları üzerinde, veri madenciliği alt dalı olan birliktelik analizi algoritmaları kullanılarak hidro-meteorolojik birliktelik örüntülerinin tespitini yapmaktır. Diğer bir deyişle, pek çok istasyondan görece uzun bir zaman dilimi için oluşturulmuş olan hidrolojik ve meteorolojik veritabanları üzerinde çeşitli parametrelerin birbiriyle olan etkileşimlerini incelemek için algoritma geliştirmektir.

Tez çalışması kapsamında, gerçek veriler üzerinde mekan-zamansal birliktelik analizi algoritmalarının çalışması incelenmiş ve bir mekan-zamansal birliktelik analizi algoritması kullanılarak sonuçlar elde edilmiştir. Ayrıca yeni bir değerlendirme ölçütü önerilerek elde edilen sonuçların daha etkin değerlendirilmesi sağlanmıştır.

Tezin İçeriği

Bu tez kapsamında yapılan çalışmalar bölümlere ayrılarak anlatılmıştır. Bölüm 1'de veri madenciliği hakkında genel bilgiler verilmiş, avantaj ve dezavantajlarına değinilmiş, veri madenciliği görevleri anlatılmış, veri madenciliğine ait değerlendirme ölçütleri hakkında genel bilgiler verilmiş ve mekan-zamansal veri madenciliği ile ilgili olarak yapılmış olan çalışmalara ait literatür çalışması verilmiştir.

Bölüm 2'de hidro-meteorolojik birliktelik analizi problemi ve bu problemin önemi tartışılmıştır. Bu tez çalışmasında geliştirilmiş olan HMP-Miner algoritması detaylı bir şekilde anlatılmıştır. Ayrıca algoritmanın örnek bir veri kümesindeki çalışması da gösterilerek, algoritmanın daha iyi anlaşılması sağlanmaya çalışılmıştır.

Bölüm 3’te HMP-Miner algoritmasının eldeki veri kümesi üzerindeki performansı değerlendirilmiş ve toy yaklaşımla olan farkları incelenmiştir.

Bölüm 4’te çalışma sonuçlar değerlendirilmiş ve gelecek çalışmalar için yapılabilecek konular üzerinde öneriler verilmiştir.

1. BÖLÜM

GENEL BİLGİLER

Artan veritabanı içeriği, bu veritabanlarındaki verilerin analiz ihtiyacını ortaya çıkarmıştır. Klasik veritabanı sorgularının yetersizliği, istatistiksel metotların tüm veri üzerinde çalışmaması, yapay zeka ve makineli öğrenme tekniklerinin ise veriyi analiz etmede geçirdiği sürenin fazla olması, veritabanlarındaki verilerin karakteristiğini ortaya çıkaran yeni ve hızlı yöntemlerin geliştirilmesini gerektirmiştir.

Veri madenciliği, diğer bilgisayar mühendisliği alanlarının yetersiz kaldığı durumlarda diğer alanların eksikliklerini ortadan kaldırmak için önerilmiş bir bilgisayar mühendisliği alanıdır.

1.1. Veri Madenciliği

Literatürde veri madenciliğine ait pek çok tanımlama yapılmıştır. Bunlardan öne çıkan birkaçı aşağıda verilmiştir.

1. Veri madenciliği, büyük boyutlardaki veri içerisinden ilginç örüntüler veya bilgi keşfetme işlemidir.[1]
2. Veri madenciliği, büyük veri havuzları içerisinden otomatik olarak anlamlı bilgiler keşfetme işlemidir.[2]
3. Veri madenciliği, şüphelenilmeyecek ilişkileri bulmak için gözlemsel veri kümelerinin analizi ve veriyi, veri sahibi tarafından anlaşılır ve kullanılabilir hale getirecek özgün yöntemlerle özetleme işlemidir.[6]

Bu tanımlardan yola çıkarak, veri madenciliği, eldeki veri kümeleri veya veritabanları

üzerinde, bu veri kümeleri veya veritabanlarındaki veriler arasındaki gözle görülemeyen, ilginç ve veriyi kullanacak kişiler için anlamlı ilişkileri ortaya çıkarma işlemidir denilebilir [1, 2].

Veri madenciliği istatistik, makineli öğrenme, yapay zeka ve veritabanı sistemlerinin bir araya gelmesiyle ortaya çıkmış bir bilgisayar mühendisliği alanıdır. Bu alanların her biri ayrı ayrı veri işleme yöntemleri olarak görülebilmesine rağmen veri madenciliği algoritmaları bu alanların güçlü özelliklerini bir araya getirmesi yönüyle diğer alanlardan ayrılır. İstatistiksel yöntemler, verinin tamamı yerine örnekleme yöntemiyle verinin bir kısmı üzerinden sonuçlar çıkarmaktadır. Ancak örneklem uzayının verinin tamamını temsil edip etmediği tartışmalıdır ve iyi belirlenmelidir. Yapay zeka ve makineli öğrenme yöntemleri verinin boyutunun büyümesi durumunda yavaşlayabilmekte ve heterojen veritabanlarında istenen sonuçları verememektedir. Ayrıca yapay zeka yöntemlerindeki sezgisellik sebebiyle her zaman iyi sonuca gidememe durumu da vardır. Veritabanı sistemleri ise tek başlarına bir veri işleme görevini tam anlamıyla yerine getirememektedir. Veri madenciliği, tüm bu alanların yetersizliklerini ortadan kaldıracak etkili ve hızlı çalışan algoritmalar geliştirmeyi amaçlamaktadır.

1.1.1. Veri Madenciliğinin Önemi

Veri depolama sistemlerindeki gelişmeler sonucunda, günümüzde artık her türden veri depolanabilir hale gelmiştir. Artık veritabanları terabyte hatta petabyte boyutlarındaki verileri saklayabilmektedir. Depolanan bu verileri analiz etmek ve verilerden bilgi elde etmek ihtiyacı ortaya çıkmıştır. Olağanüstü boyutlardaki veriden otomatik olarak değerli bilgilerin çıkarılması ve bu verilerin organize edilmiş bilgiye dönüştürülmesi veri madenciliğinin ortaya çıkmasına sebep olmuştur.

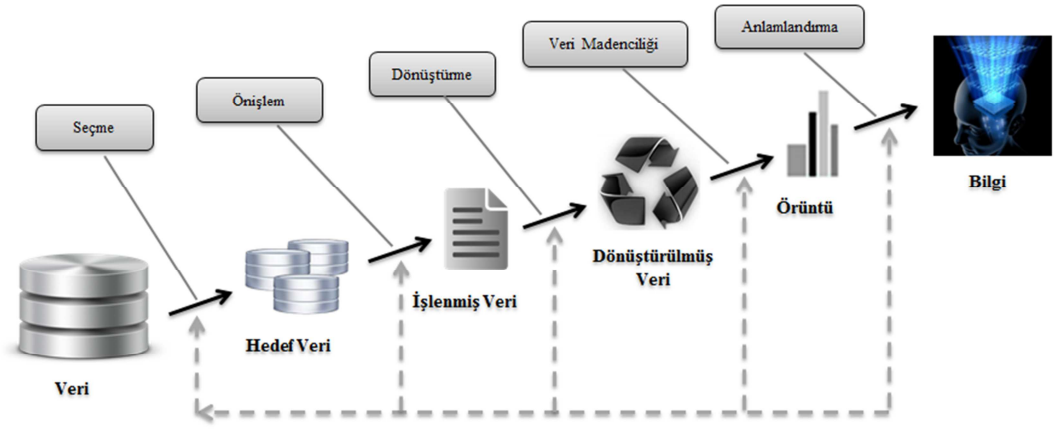
Yapılacak olan işlemin kısmen de olsa otomatik olması ve elde edilen bilginin anlamlı bir şekilde kullanıcıya sunulması önem arz etmektedir. Veri madenciliği, bu şartları sağlaması yönüyle ve verinin anlamlı bilgiye dönüştürülmesi aşamalarını başarılı bir şekilde gerçekleştirmesi ile diğer veri üzerinden bilgi keşfi yöntemlerinden ayrılmaktadır.

Örnek olarak bir süpermarket veritabanına her gün binlerce kayıt aktarılmaktadır. Bu kayıtlar arasında birbiriyle ilişkili, birbirini takip eden veya birlikte oluşan kayıtların

ortaya çıkarılması market yöneticilerinin çeşitli stratejik kararlar almalarını ve elde edilecek kârı arttırmalarını sağlayabilecektir.

1.2. Veri Madenciliğinin Adımları

Her problem çözme yönteminde olduğu gibi veri madenciliğinde de veriden bilgi keşfi işlemi alt parçacıklara bölünerek, sırasıyla bu parçacıklar tamamlanarak yapılmaktadır. Veri madenciliği adımları olarak da ifade edilebilecek bu adımlar klasik problem çözme adımlarının veri madenciliği için özelleştirilmiş bir türü olarak görülebilir.



Şekil 1.1: Veri madenciliği adımları[1, 2]

Şekil 1.1'de verilen adımlar aşağıda sırasıyla açıklanmaktadır[1, 2]:

a. **Veri Seçme:** İlk adım olarak öncelikle veritabanındaki veriler arasından, üzerinde çalışma yapılacak veri kümesi seçilerek gereksiz veri erişimleri engellenmiş ve algoritma için daha sade bir veri kümesi oluşturulmuş olmaktadır.
b. **Veri Önişlemi:** Önişlem aşamasında seçilmiş olan veri kümesi üzerinde algoritma çalıştırılmadan önce seçilen verinin, algoritmanın kabul edeceği formata uygun hale getirilmesi işlemi yapılmaktadır. Önişlem adımında veri üzerinde yapılabilecek işlemler; i) ayrıklaştırma – veriyi sürekli formdan ayrık forma çevirmek, ii) birleştirme – veriyi günlük formattan aylık formata çevirme, iii) örnekleme – verinin sadece belli bir kısmını alma ve iv) boyut indirgeme – verideki fazla parametreleri temizleme sayılabilir. Bu aşamada yapılan işlemler halen algoritmaya ait işlemler olmamakta sadece veriyi algoritmanın çalışmasına uygun hale getirmeyi amaçlamaktadır.

c. **Veri Dönüştürme:** Elde edilen veri kümesindeki veriler bir formdan başka bir forma dönüştürülecekse bu adımda dönüştürme işlemi yapılır. Veri madenciliği algoritmasından önceki son aşama verinin dönüştürülmesi aşamasıdır. Bu adım veri madenciliği işlemleri için zorunlu bir adım değildir. Bazı durumlarda verinin durumuna bağlı olarak bu adımın veri madenciliği işlemi içerisinde gerçekleştirilmesi gerekebilir.
d. **Veri Madenciliği:** Bu adımda önişlem ve dönüştürme adımlarında hazırlanmış olan veri kümesi üzerinde veri madenciliği algoritmaları çalıştırılmaktadır. Algoritmalar verilerin özelliğine göre çeşitli sonuçlar ortaya çıkarırlar. Bu sonuçlar veri madenciliği algoritmasının sonuçlarıdır ve kullanıcıların anlayabilmeleri için algoritmaların işleyişini bilmeleri gerekmektedir. Veri madenciliği adımı sonucu çıkarılan örüntülere ham örüntüler de denilebilir.
e. **Sonuç Anlamlandırma:** Bu adımda veri madenciliği algoritmalarından elde edilen sonuçlar anlamlandırılarak algoritmanın işleyişini bilmeyen son kullanıcıların anlayabileceği hale getirilmektedir. Bu adım sonucunda artık veri madenciliği işlemi tamamlanmış ve sonuçları toplanmıştır denilebilir.

Yukarıdaki adımlar sırasıyla takip edilerek veri madenciliği yöntemiyle veritabanlarından anlamlı sonuçlar ve örüntüler çıkarılmaktadır. Belirtilen adımların bazıları mecburi olmakla birlikte bazıları atlanabilirdir. Örneğin eğer veritabanında bulunan verilerin tamamı kullanılacaksa veri seçme adımı atlanabilir. Veri üzerinde bir dönüştürme işlemi yapılmayacaksa bu adım atlanabilir. Mecburi olan adımlar ise veri madenciliği ve sonuç anlamlandırma adımlarıdır. Bu adımlar olmadan veri madenciliği yapılmıştır denilemez.

1.3.Veri Madenciliğinin Avantaj ve Dezavantajları

Veri madenciliği, örüntüleri ve birliktelikleri tespit etmede var olan sistemlerden daha farklı yaklaşımlar önermektedir. Önerilen bu yaklaşımların ise hem çeşitli avantajlarını hem de dezavantajları ortaya çıkmaktadır. Aşağıda veri madenciliğine ait avantaj ve dezavantajlar incelenmektedir.

Veri madenciliğinin avantajları[1, 2]:

1. Veri madenciliği algoritmaların hızlı çalışabilmesi,
2. Veri madenciliği algoritmalarının esnek olması,
3. Her türlü veri üzerinde uygulanabilir olması,
4. Üzerinde çalışılan veri içerisindeki ani değişimleri tespit edebilmesi,
5. Çalışılan verinin boyutuna bağlı olarak ölçeklenebilir olması,
6. Artan veritabanları üzerinde etkin bir şekilde çalışabilmesi,

sayılabilir.

Veri madenciliğinin dezavantajları[1, 2]:

1. Kullanılacak verinin mahremiyeti,
2. Verinin güvenliği,
3. Veri için en uygun yöntemin belirlenmesinin zorluğu

sayılabilir.

1.4. Veri Madenciliğinin Temel Zorlukları

Veri madenciliği, sağladığı avantajları ile değerlendirilmesiyle birlikte bu avantajları elde edebilmek için çeşitli zorlukları da bulunmaktadır. Bu zorluklar metotlar açısından, kullanıcı etkileşimi açısından, etkinlik ve ölçeklenebilirlik açısından ve veritabanı türlerinin değişkenliği açısından olmak üzere dört başlıkta incelenebilir [1]:

1.4.1. Metotları Açısından

1. **Çeşitli ve yeni tür bilgilerin madenciliği:** Aynı veri kümesi, birbirinden farklı türde veri madenciliği yöntemleri açısından incelenerek farklı boyutlarda bilgiler elde edilmeye çalışılabilir ve bu işlem çok büyük çaba gerektirmektedir.
2. **Çok boyutlu verilerin madenciliği:** Üzerinde çalışılan veri kümesi çok fazla sayıda parametre içeriyorsa bu parametreler arasındaki ilginç örüntülerin keşfi zorlaşmaktadır.
3. **Disiplinler-arası çalışma:** Veri madenciliğinde kullanılan veri kümelerinin ait oldukları alanla ilgili sorunlar ve veri üzerinde yapılması gereken ek işlemler veri madenciliği işlemine olumsuz yönde etki etmektedir.

4. **Veri kümesinin ağ üzerinde bulunması:** Veri kümesi birbirine bağlantılı çevrelerde dağılmış bir şekilde bulunabilir (Web, dosyalar, dokümanlar vb.). Bu verilerin bir araya getirilerek algoritmaların çalıştırılması daha zor ve daha fazla kaynak gerektiren bir işlem olmaktadır.
5. **Kesin ve tam olmayan, gürültü içeren veri:** Veriler genellikle gürültü içerir ve kesin ve tam değildir. Eksik ve tam olmayan veriler üzerinde çalıştırılan veri madenciliği algoritmalarının sonuçları beklenen şekilde olmayacaktır. Eksik ve gürültülü veriler üzerinde yapılacak olan işlemler de veri madenciliği sürecini etkileyen bir etmen olacaktır.
6. **Örüntülerin değerlendirilmesi:** Veri madenciliği algoritmaları tarafından ortaya konulan kurallardan hangilerinin ilginç ve alakalı olduğu, hangilerinin alakasız olduğuna karar verecek bir yapı oluşturmak çok zordur ve tam etkin çalışacak bir yapı kurulamamaktadır.

1.4.2. Kullanıcı Etkileşimi Açısından

1. **İnteraktif veri madenciliği:** Veri madenciliği işlemleri interaktif olmak durumundadır. Esnek kullanıcı arayüzleri tasarlanmalı ve veri madenciliği aşamaları kullanıcı kontrolünde olmalıdır.
2. **Altyapısal bilginin kullanılması:** Çalışılan alanla ilgili sabitler, kısıtlamalar ve kurallar sürece dahil edilebilmelidir.
3. **Veri madenciliği sonuçlarının sunumu ve görselleştirilmesi:** Veri madenciliği işlemi sonucu elde edilen sonuçlar etkin ve kolay anlaşılabilir bir şekilde sunulabilmelidir.

1.4.3. Etkinlik ve Ölçeklenebilirlik Açısından

1. **Veri madenciliği algoritmalarının etkinliği ve ölçeklenebilirliği:** Büyük boyutlardaki veriler içerisinden efektif bir şekilde bilgiler çıkarabilmek için veri madenciliği algoritmaları etkin ve ölçeklenebilir olmalıdır. Etkinlik, ölçeklenebilirlik, performans, optimizasyon ve gerçek zamanlı çalışabilme anahtar kriterlerdir.
2. **Paralel, dağıtık ve artırımlı madencilik algoritmaları:** Çok büyük boyutlu veri kümeleri, verinin dağıtılmış olması ve bazı algoritmaların hesaplama karmaşıklığı paralel ve dağıtık algoritmalar geliştirilmesine sebep olmuştur. Ancak bu yapılar da veri madenciliği aşamalarını karmaşıklaştırarak algoritmaların kullanılmasını zorlaştırmaktadır. Ayrıca artırımlı veriler üzerinde

çalışmak için gerçek zamanlı erişim ve karmaşık algoritmalar gerekmektedir. Bu işlemler de veri madenciliğini zorlaştıran etmenler arasındadır.

1.4.4. Veritabanı Türlerinin Değişkenliği Açısından

1. **Kompleks veri türleri:** Veri kümeleri; kararlı veri ambarlarından dinamik veri akışlarına, temel veri türlerinden zamansal veri türlerine, biyolojik diziler, sensor verileri, mekânsal veriler, multimedya verileri, Web verisi ve sosyal medya verilerine kadar pek çok türde bulunabilmektedir. Bu türleri efektif ve etkili bir şekilde bir araya getirebilmek çaba gerektiren bir işlemdir.
2. **Dinamik, ağ bağlantılı ve küresel veri ambarlarında veri madenciliği:** Bir önceki maddede belirtilen türlerdeki verilerin tümü ile ve ayrıca bu verilerin birlikte kullanımları ile etkili bir şekilde çalışabilecek veri madenciliği algoritmaları geliştirmek son derece büyük çabalar gerektiren bir işlemdir. Bu tarz algoritmaları geliştirebilmek ise veri madenciliğinin zorluklarından birisidir.

1.5. Veri Madenciliğinin İlgilendiği Konular

Veri madenciliği, hedeflediği sonuçları ve veriyi işleme şekillerine göre birbirinden farklılık gösteren 4 ana konudan oluşmaktadır. Bu konular, birliktelik analizi, anormallik tespiti, kümeleme ve sınıflandırmadır. Her bir konu ayrı ayrı aşağıda anlatılmaktadır.

1.5.1. Birliktelik Analizi

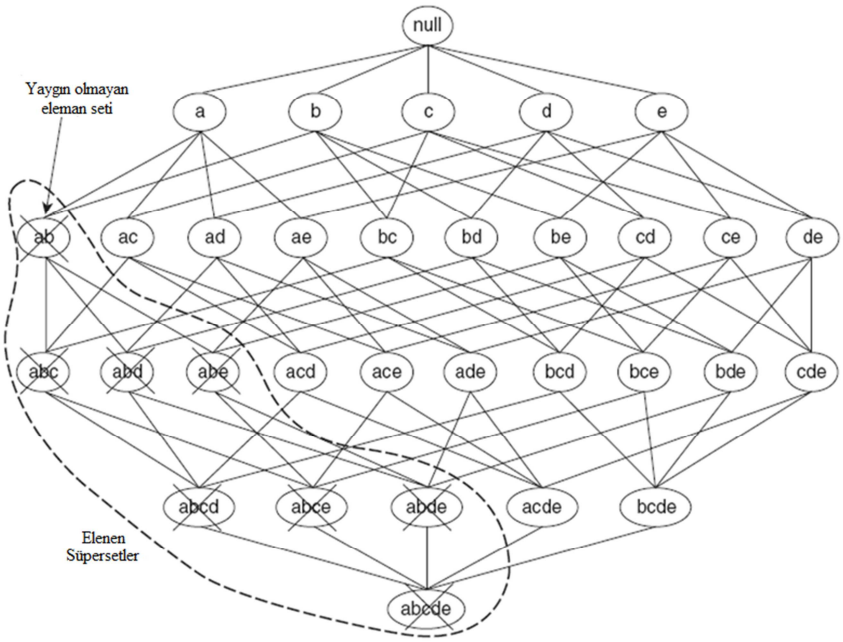
Birliktelik analizi, bir veri kümesindeki parametrelerin bir arada oluşumlarını inceleyen ve sonucunda birliktelik kuraları ortaya çıkaran bir veri madenciliği tekniğidir. Parametrelerin birlikte oluşturdukları örüntüler belirlenmiş eşik değerlerinden yüksek ise bu örüntülerin veri kümelerinde yaygın olarak bulundukları ifade edilir ve bu örüntüler birliktelik kuralları olarak kabul edilir. Birliktelik analizi sonucunda, “Eğer A parametresi var ise, B parametresi de verilen eşik değeri olasılıkla olacaktır.” şeklinde örüntüler elde edilir.

Birliktelik analizi için en uygun örnek market sepeti analizidir. Amaç, müşterilerin satın aldığı ürünler arasındaki birlikteliklerin bulunması ile müşterilerin satın alma alışkanlıklarını analiz etmektir. Bu birlikteliklerin keşfedilmesi ile satıcılar, müşterilerin hangi ürünleri sıklıkla birlikte satın aldıklarını bilecek ve satış stratejilerini bu bilgilerden yola çıkarak geliştirebileceklerdir. Örneğin; eğer müşteriler cips ve kola ikilisini sıklıkla birlikte satın alıyor iseler bu iki ürün için bir kampanya düzenlemek

veya bu iki ürünü müşteriler için daha görülebilir yerlere koymak bu ürünlerin satışını artıracaktır.

Birliktelik analizi, çok geniş bir uygulama aralığına sahip ise de yoğunlukla finans servislerinde, bankacılıkta, reklamcılık ve işletme sektörlerinde ve elektronik ticaret alanlarında kullanılmaktadır. Ayrıca veri madenciliği e-ticaret sektöründe, metin madenciliği alanında, tıbbi verilerin işlenmesinde, kısacası verinin bulunduğu her alanda yoğun bir şekilde kullanım potansiyeline sahiptir.

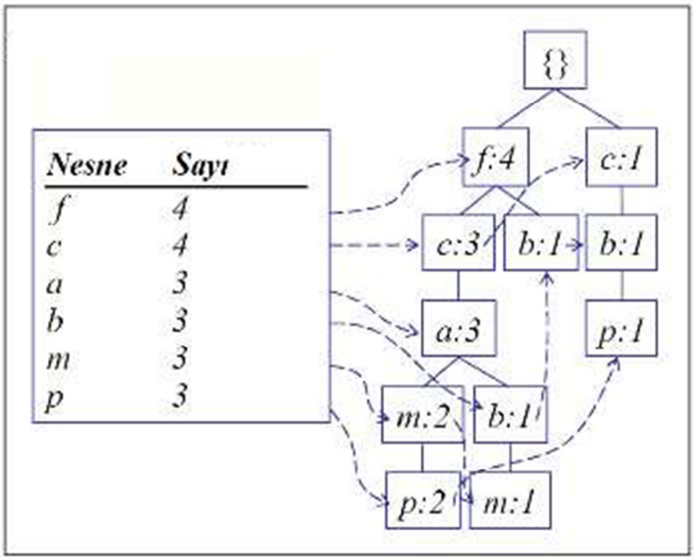
Birliktelik analizi algoritmaları a-priori prensibi çerçevesinde şekillenmektedir [4]. A-priori prensibine göre, eğer bir örüntü veri kümesinde yaygın ise, bu örüntünün tüm alt kümeleri de yaygın olmak zorundadır. Diğer bir deyişle, bir örüntünün yaygın olduğuna karar verebilmek için bu örüntüye ait tüm alt örüntülerin veri kümesinde yaygın olması gerekmektedir. Bu prensip yardımıyla gereksiz örüntülerin üretimi engellenebilmekte ve algoritmalar daha hızlı çalışabilmektedir [1].



Şekil 1.2. Apriori metoduna ait kafes yapısı ve elenen elemanlar

Bir diğer yaklaşım ise FP-Tree gelişim ağacı yaklaşımıdır [1]. Bu yaklaşım da A-priori yaklaşımına ait temelleri paylaşmakla birlikte, veri kümesindeki dolanım sayısını azaltması, daha az yer kaplaması ve ağaç yapısının verdiği avantajları ile A-priori

yaklaşımından ayrılmaktadır. Her iki yaklaşım da etkili ve güçlü birliktelik analizi algoritmaları içermektedir [2].



Şekil 1.3. FP-Tree metoduna ait çalışma şeması ([1]'den modifiye edilmiştir.)

Birliktelik analizinde önemli bir unsur da örüntülerin kural olarak kabul edilme eşik değerleridir. Bu eşik değerleri değerlendirme ölçütleri (interest measures) olarak ifade edilmektedir. Çeşitli problemler için birbirinden farklı değerlendirme ölçütleri önerilmiştir. İlgilenilen problemin doğası hangi ölçüte uygun ise ve istenilen sonuçları hangi ölçütün daha iyi vereceği düşünülüyor ise o ölçüt tercih edilmektedir. En yaygın kullanılan değerlendirme ölçütleri destek (support) - güven (confidence) ölçütleridir. Destek değeri parametrelerin veri kümesi içerisinde bulunma olasılığı, güven değeri de parametrelerin birbirlerine bağlı olarak bulunma olasılığını ifade etmektedir. Örüntüler, bu ölçütler kullanılarak değerlendirilir ve kural olarak seçilip seçilmeyeceği belirlenir. Değerlendirme aşamasında eğer eldeki örüntünün destek değeri kullanıcı tanımlı minimum destek değerinden yüksekse ve örüntünün güven değeri de kullanıcı tanımlı minimum güven değerinden yüksekse bu örüntü kuraldır denir ve kural setine dahil edilir.

1.5.2. Sınıflandırma

Sınıflandırma, veri kümeleri üzerinde önemli veri sınıflarını tanımlayacak modeller ortaya koyan bir veri madenciliği yöntemidir. Sınıflandırmada, eldeki veri kümesindeki parametreler kullanılarak verinin ait olduğu sınıf belirlenmeye çalışılmaktadır.

Sınıflandırma işleminin ne kadar güvenilir olduğu bu sınıfların seçiminin ne kadar doğru olduğuyla da ilgilidir. Çünkü ortaya konulan model bu sınıflar yardımıyla etiketlenmemiş verileri sınıflandırmaya çalışmaktadır. Bir şirket yönetiminin eleman alımında sınıflandırma yöntemini kullanmak istediğini düşünelim. Bu şirket için önemli olan parametreler, çalışanın hangi sosyal statüde olduğundan daha çok hangi bilgilere sahip olduğu, etnik kökeninden çok hangi işlerde çalıştığının daha önemli olması gerekmektedir. Bu sebeple parametre seçimi sınıflandırma işlemi için neredeyse en önemli karar aşamasıdır.

Sınıflandırma için örnek olarak, banka kredi uygulamaları verilebilir. Başvuran şahıslar arasından kredi verilecek olanların seçilmesi bankalar için riskli bir karardır. Bankalar, sonucunda krediyi geri ödeyebileceğini garanti edecekleri adaylara kredi vermek isterler. Bu durumda, eldeki geçmiş başarılı ve başarısız kredi bilgilerini kullanarak bir karar mekanizması oluşturmak istediklerinde, bu işlemi en iyi yapacak olan sistem sınıflandırma sistemidir. Bu örnekte iki sınıf bulunmaktadır; başarılı ve başarısız. Seçilebilecek parametreler olarak da, başvuru sahibinin aylık geliri, evlilik durumu, ev sahibi olup olmaması, araba sahibi olup olmaması, mesleği vb. parametreler kullanılabilir. Bu parametreler kullanılarak başvuru sahibinin verilen krediyi geri ödeyip ödeyemeyeceği belirlenebilir. Sınıflandırma yöntemi ile de bu parametreler kullanılarak yeni başvuran adaylar için karar aşaması sağlanabilmektedir.

Sınıflandırma; banka kredi uygulamalarında, protein yapılarının sınıflandırılmasında, tıpta hastalıklara ve derecelerine karar vermede, şirketlerde çalışan alımlarında ve etiketli veri bulunabilecek her türlü alanda uygulanabilir olmaktadır.

Sınıflandırma işlemi genel olarak iki adımdan oluşmaktadır. Birincisi öğrenme adımı, ikincisi de sınıflandırma adımıdır. Öğrenme adımında, eldeki sınıflandırılmış bir veri kümesi kullanılarak, bilinen parametreler yardımıyla belirlenmiş sınıflar için modeller oluşturulmaktadır. Sınıflandırma adımında ise, öğrenme adımında ortaya konulmuş olan modele göre sınıflandırılmamış veriler bilinen sınıflardan birine dahil edilmektedir. Sınıflandırma işleminin sonucunda hangi sınıfa dahil olduğu bilinmeyen, etiketlenmemiş, verinin, belirlenmiş olan sınıflardan birine dahil edilmesi gösterilebilir.

En yaygın kullanılan sınıflandırma yöntemi karar ağaçlarıdır. Karar ağaçlarında her bir dal bir parametreyi ifade etmektedir ve bu dallara göre ulaşılmış olan her yaprak bir

kararı temsil etmektedir. Eldeki eğitim verisi kullanılarak oluşturulan ağaca sınıflandırılması istenen veriler girilerek, hangi sınıfa ait oldukları tespit edilebilir. Karar ağaçları, kural tabanlı sınıflandırma vb. yöntemler sınıflandırmada kullanılan yöntemlerden bazılarıdır.

1.5.3. Kümeleme

Kümeleme, bir veri kümesindeki nesnelerin, bir kümeye ait nesneler arası benzerliğin en yüksek ve diğer kümelere ait nesneler arasındaki benzerliğin de en düşük olacak şekilde birçok gruba veya kümeye gruplanmasıdır. Benzerlik ölçütleri nesnelerin özelliklerine göre belirlenmekle beraber genellikle uzaklık ölçütleri kullanılmaktadır. Kümelemenin sınıflandırmadan farkı, her bir nesnenin ait olduğu sınıf etiketi bilgisinin olmamasıdır. Bu durumda, grupları veya kümeleri belirleyebilmek için otomatik olarak veriler arasından çıkarım yapan yöntemler kullanılması gerekmektedir. Kümeleme işleminde kümeler, sınıflandırmadaki eğitim verisinin uzmanlar tarafından belirlemesinden farklı şekilde insanlar tarafından değil, kümeleme algoritmaları tarafından belirlenmektedir. Bu durumda da farklı kümeleme algoritmaları, aynı veri kümesi üzerinde farklı sayıda ve yapıda kümeler elde edebilirler. Bu durumda hangi kümeleme algoritmasının eldeki veri kümesine uygun olduğuna kullanıcı karar vermek durumundadır.

Kümelemeye örnek olarak bir işletmenin müşterileri arasında aynı özellikleri paylaşan ve aynı karakteristiğe sahip olanlarının bir araya getirilmesi verilebilir. Aynı özelliklere sahip olan müşteriler, aynı ürünleri almaya yatkın olmaktadırlar ve zevkleri de birbirine benzemektedir. Burada amaç, hedef kitleyi belirlemek ve bu hedef kitlelere uygun ürünler ortaya koymak olarak belirlenebilir. Ayrıca müdavim müşterilerin hangi kümeye dahil olduğu da tespit edilerek raf organizasyonunun, indirimlerin ve promosyonların buna uygun olarak düzenlenmesi de bu örneğe dahil edilebilir.

Kümeleme, sıklıkla ticari işletme stratejilerinde, görüntülerden örüntü tanımlamada, web arama motorlarında, biyolojide hayvan veya bitkileri verilen özelliklerine göre gruplandırmada ve güvenlik alanında kullanılmaktadır.

Bir nesnenin hangi kümeye ait olduğunun tespitini yapabilmek için, o nesnenin, o ana kadar ortaya çıkarılmış olan kümelere olan uzaklığı tespit edilir. Buna göre bir nesne, ya var olan kümelerden birine dahil olacak, ya yeni bir küme oluşturacak ya da gürültü

olarak belirlenecektir. Her bir nesne için bu karar mekanizması işletilir ve sonucunda tüm veri kümesi kümelenmiş olur.

Kümelemeden elde edilmek istenen sonuç göz önüne alındığında, birbirinden farklı yaklaşımlar ortaya konulmuştur. Bu yaklaşımlar; en genel şekliyle hiyerarşik ve hiyerarşik olmayan metotlar olarak ikiye ayrılabilir. Hiyerarşik metotlar yığıcı (agglomerative) ve bölücü (divisive) metotlar olarak ikiye ayrılabilir. Hiyerarşik olmayan metotlar ise ayırma (partitioning) metotları, yoğunluk-tabanlı (density-based) metotlar ve grid-tabanlı (grid-based) metotlar ve diğer yaklaşımlar olmak üzere dört kısımdan oluşmaktadır. Hiyerarşik metotlarda, veri kümesindeki nesneler bir ağaç yapısına sahip kümelere dahil edilirler. Bu şekilde bir kümenin altında yine kümeler veya nesneler bulunabilir. Hiyerarşik olmayan metotlarda ise veri kümesi belli sayıda kümeye ayrılır. Bir küme, kendi altında başka küme içeremez.

Kümeleme işleminin hedefine göre kümeleme metotlarından bir veya birkaçı kullanılabilir. Örneğin eğer veri kümesindeki nesnelerin sadece tek bir kümeye ait olması gerekiyor ise bu durumda bölümlendirme metotlarını kullanmak en mantıklı seçim olacaktır.

1.5.4. Anormallik Tespiti

Anormallik tespiti (veya aykırılık tespiti), bir veri kümesi içerisindeki nesneler arasında beklenen normal davranışı göstermeyen nesnelerin bulunması işlemidir. Anormallikle gürültü kavramları birbirinden farklıdır. Gürültü, ölçülen değişkenin değerindeki sapma veya hataya denmektedir. Veri analizinde gürültü genellikle ilginç değildir. Ancak anormallik, verinin kalanından farklı bir mekanizmayla üretildiğinden şüphe edildiği için ilginçtir. Bu sebeple, anormallik tespitinde, anormal nesnenin nasıl farklı bir mekanizmayla üretildiğini doğrulamak önemlidir. Doğrulama işlemi, verinin kalan kısmını ifade eden varsayımlarda bulunmak ve anormal nesnelerin bu varsayımlardan çok farklı olduğunu göstermekle yapılmaktadır.

Anormallik tespitine, bankaların müşterilerini kredi kartı kaçakçılığından korumak için kart işlemlerinden sonuçlar çıkarmaları örnek olarak verilebilir. Örneğin bir kart kullanıcısı, normal alışverişinin çok üzerinde bir alışveriş yaparsa veya normal alışveriş yaptığı saat dışında bir alışveriş işlemi yapılırsa veya yaşadığı şehir dışında bir şehirde

kredi kartı kullanılırsa bu kullanıcının kredi kartının çalınmış olma ihtimali vardır denilebilir. Çünkü yapılan işlem normal işlemlerden aykırı bir işlem olmaktadır.

Anormallik tespiti, hile tespitinde, tıbbi bakım işlemlerinde, bilgisayar ağlarındaki sızmaların tespitinde, görüntü ve video işlemede aykırılıkların tespitinde, endüstriyel hata tespitinde kullanılmaktadır.

Anormallik tespitinde, anormallikler üç türde incelenmektedir; küresel anormallikler, içeriksel anormallikler ve toplu anormallikler. Eğer anormal nesne, verinin kalanından belirgin bir şekilde uzaklaşıyorsa bu durumda bu nesneye küresel anormal nesne denir. Bu anormallikler zaman zaman noktasal anormallik olarak da ifade edilir ve en temel anormallik türüdür. Eğer anormal nesne, bu nesnenin özel bir içeriği sebebiyle diğer nesnelerden ayrılıyorsa bu durumda bu nesneye içeriksel anormal nesne denir. İçeriksel anormallikler, nesnelerin seçilmiş bir özelliğine göre belirlendikleri için şartlı anormallik olarak da bilinirler. Eğer anormal nesneler, bir grup olarak veri kümesinden farklılık gösteriyor iseler bu durumda bu anormal nesnelere toplu anormal nesneler denir. Toplu olarak anormal olan nesneler kendi içlerinde anormal olarak görülmeyebilirler, ancak grup olarak anormal davranış sergiledikleri de gözlemlenebilir.

Anormallik tespiti metotları, en genel olarak üç türe ayrılmaktadır; danışmanlı metotlar, yarı-danışmanlı metotlar ve danışmansız metotlar. Danışmanlı metotlarda, normal ve anormal veriler modellenerek normal ve anormal veriler, veri kümesinden işaretlenmiştir. Sınıflandırmadaki modeller kullanılarak anormallik tespiti yapılabilir. Yarı-danışmanlı metotlarda, normal veri veya da anormal veri modellenmiş olarak bulunmaktadır. Eldeki bilinen modelden yola çıkılarak bilinmeyen parça da tahmin edilebilir. Danışmansız öğrenme metotlarında da başlangıçta normal verilerin kümelendiği varsayımından yola çıkılmaktadır. Bu şekilde kümeleme tabanlı algoritmalar yardımıyla anormallik tespiti yapılmaktadır.

1.6. Mekân-Zamansal Birliktelik Analizi

Veri madenciliği, pek çok şekillerde alt dallara bölümlendirilebilir. Çalışılan veri kümesinin karakteristiği, uygulanacak olan yöntemin içeriği ve geçirilecek olan işlem adımları bu alt dallara bölümlendirmede etkili olmaktadır. Bu tez çalışması kapsamında yapılan çalışmalar, veri madenciliğinin bir alt dalı olan birliktelik analizinin içerisinde mekân-zamansal birliktelik analizi olarak kendine yer bulmaktadır.

Mekân-zamansal veri madenciliği, eldeki veri kümesi içerisindeki kayıtların hem mekân hem de zaman bilgisi olduğu ve bu bilgilere bağlı olarak sonuç çıkarımı yapılması gerektiği durumlarda kullanılan bir veri madenciliği alanıdır. Mekân-zamansal birliktelik analizi de benzer şekilde birliktelik analizinin mekân ve zaman bilgilerine bağlı olarak sonuçlar çıkarılması gereken durumlar için kullanılmaktadır. Mekân-zamansal birliktelik analizine ihtiyaç duyulmasının en temel sebebi mekân ve zaman bilgilerini de veriye dahil ettikten sonra elde edilen sonuçlarla verinin sadece kendisini kullanarak elde edilen sonuçlardan çok daha fazla bilgi içerebilmesidir.

Market sepeti örneği üzerinden bir karşılaştırma yapılacak olunursa, örneğin günün belli saatlerinde veya ayın belli günlerinde, farklı bölgelerdeki market şubelerinden yapılan alışverişler incelendiğinde belli ürünler daha fazla ilgi görüyorsa, bu ürünler için daha çok ilgi gördüğü günler göz önünde bulundurularak kampanyalar düzenlenebilir. Daha farklı bir yaklaşımla, örneğin bayram günlerinde, farklı semtlerdeki şeker ve tatlı satışları üzerinden bir mekân-zamansal birliktelik analizi ile bu ürünlerle birlikte örneğin kuruyemişlerin veya kolonya satışlarını da içerdiği bulunabilirse bu birliktelik de mekân-zamansal birliktelik analizi ile daha kârlı kampanyalar yapılmasını sağlayabilmektedir.

Mekân-zamansal birliktelik analizi genel olarak mekân ve zaman bilgilerini birlikte içeren alanlarda kullanılmaktadır. Market sepetinin mekân ve zaman bilgileri içeren geliştirilmiş versiyonunda, internet kayıtlarının incelenmesinde, satın alma ve satış geçmişlerinin analizinde, iklimsel olayların tespitinde mekân-zamansal birliktelik analizi sıklıkla kullanılmaktadır.

İnternet geçmişleri incelenerek bir internet kullanıcısına ait gezinme alışkanlığı tespit edilebilir. Ulaşılabilen tüm internet kullanıcılarının alışkanlıkları tespit edilerek buna göre reklamlar yerleştirerek daha fazla dikkat çekecek şekilde bir yapı kurulabilir. Yine bunu sağlamak için kullanıcılara ait tekrarlı kayıtların zamansal analizine ihtiyaç duyulmaktadır.

İklimsel olaylar, sadece iklimle ilgili çalışanlar için değil incelenmesi neredeyse herkes için bir ihtiyaçtır. Örneğin, hava durumuna bakarak buna göre gün içerisinde yapacağımız işleri düzenleriz. Hidrolojik olaylar ise, doğadaki su döngüsü ile nehirlerdeki su oranları ile ilgili olayları kapsar. Örneğin, akarsu akımları incelenerek

seller veya kuraklıkların ortaya çıkarılması sağlanabilir. Hidrolojik olayları tetikleyen bazı meteorolojik faktörler bulunmaktadır. Bu çalışmada akım parametresi ve başta yağış olmak üzere meteorolojik parametreler arasındaki mekân-zamansal birliktelikler incelenmektedir. Böylece ortaya çıkacak olan kurallar yardımıyla Türkiye'de akımı etkileyen meteorolojik faktörler tespit edilmektedir.

1.7. Mekân-Zamansal Birliktelik Analizi Değerlendirme Ölçütleri

Değerlendirme ölçütleri veri madenciliği için çok önemli bir yer tutmaktadır. Değerlendirme ölçütleri, algoritmalar tarafından ortaya çıkarılan örüntülerin ilginç, manalı veya yaygın olup olmadığını değerlendirmek için kullanılmaktadır. Çalışılacak olan veri kümesi üzerinde hangi algoritmanın çalıştırılacağı kadar hangi değerlendirme ölçütü ile ortaya çıkarılan örüntülerin değerlendirileceği de önem taşımaktadır. Çünkü ortaya çıkarılan örüntülerin kalitesini, kullanılan değerlendirme ölçütü belirlemektedir. Değerlendirme ölçütlerine ait ortaya konulmuş tam bir tanımlama bulunmamaktadır. Şimdiye kadar yapılan çalışmalardan derlenecek olunursa bir değerlendirme ölçütünde bulunması gereken özellikler; kısalık, geniş kapsamlılık, güvenilirlik, alışılmamışlık, farklılık, tuhaflık, şaşırtıcılık, yararlılık ve uygulanabilirlik olarak sayılabilir [3].

Mekân-zamansal birliktelik analizi için özel bir değerlendirme ölçütü kullanılmamaktadır. Klasik birliktelik analizine ait ölçütler mekân-zamansal birliktelik analizi için de geçerli olabilmekte ve etkin bir şekilde çalışabilmektedir. Bu tez çalışması kapsamında, mekân-zamansal birliktelik analizi için gereken değerlendirme ölçütünün hangi özelliklerde olması gerektiği de tartışılacak ve mekân-zamansal birliktelikleri daha etkili bir şekilde ortaya çıkarabilecek bir değerlendirme ölçütü önerilecektir.

Birliktelik analizinde sıklıkla kullanılan değerlendirme ölçütlerinden biri destek (support) ölçütüdür. Destek ölçütü, birliktelik analizi algoritması ile elde edilen aday örüntülerin, veri kümesi içerisindeki yaygınlığını ortaya çıkarmaktadır. Bu şekilde veri içerisinde yaygın olmayan örüntüler elenerek, algoritma sonucu olarak sadece alakalı örüntüler ortaya çıkmaktadır. Destek ölçütü olasılık tabanlı bir ölçüttür ve veri içerisinde örüntünün geçme sıklığını tespit etmektedir. Örnek bir {X, Y} örüntüsü için destek ölçütünün formülü elde edilecek olunursa:

$$destek(XY) = \frac{XY \text{ örüntüsünü içeren kayıtların sayısı}}{Veri \text{ kümesindeki toplam kayıt sayısı}} \quad (1.1)$$

şeklinde ifade edilebilir. Daha kısa bir şekilde formül olasılıksal olarak ortaya konulacak olunursa $destek(XY) = P(X \cup Y)$ olarak da gösterilebilir. Destek ölçütü az sayıda parametreye ihtiyaç duyması ve yaygın örüntüleri ortaya çıkarmadaki başarısından dolayı tercih edilmektedir.

Destek ölçütünü kullanmaktaki sakınca, sadece yaygın örüntüleri ortaya çıkarmasıdır. Yaygın olmayan ancak ilginç örüntülerin tamamını alakasız olarak belirleyerek elemektedir. Bu durumda tuhaf ancak veri kümesi içerisinde yaygın olmayan örüntülerin ortaya çıkarılamamasına sebep olmaktadır.

Destek ölçütü ile birlikte kullanılan ve elde edilmiş olan örüntülerden X→Y tarzında birliktelik kuralları elde edilebilmesini sağlayan bir diğer ölçüt ise güven (confidence) ölçütüdür. Örnek X→Y örüntüsünde öncül parametre X, ardıl parametre ise Y parametresi olarak alınacaktır. Yani kuralın **eğer** kısmındaki parametre öncül, **ise** kısmındaki parametre ardıl olarak isimlendirilmektedir. Bu ölçüt, eldeki örüntüyü kural haline getirebilmek için gereken şartı ifade etmektedir. Yine örnek bir XY örüntüsünden X→Y kuralını elde ederken kullanılan güven ölçütünün formülü elde edilecek olunursa:

$$güven(XY) = \frac{XY \text{ örüntüsünü içeren kayıtların sayısı}}{\text{Öncül örüntüyü } (X'i) \text{içeren kayıtların sayısı}} \quad (1.2)$$

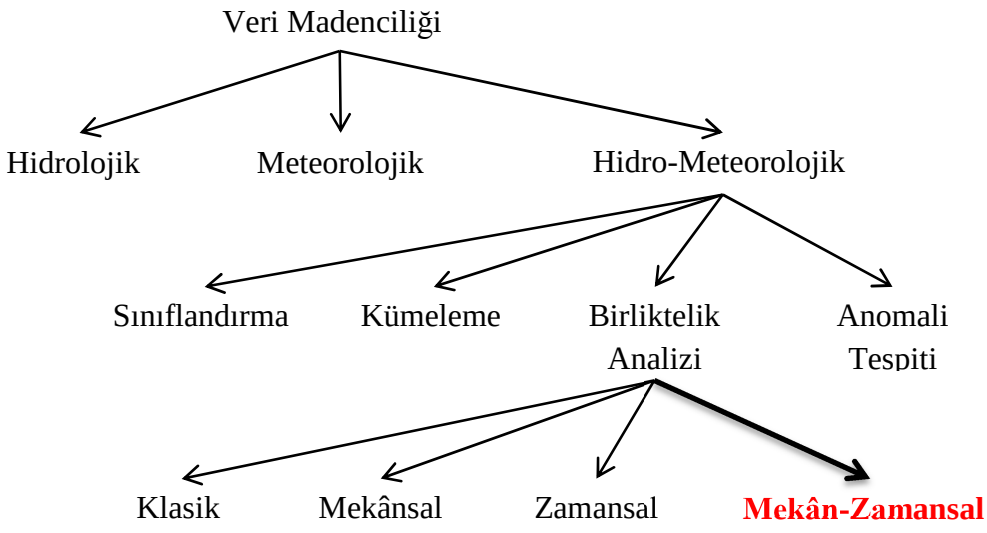
şeklinde ifade edilir. Daha kısa bir şekilde formül olasılıksal olarak ifade edilecek olunursa $güven(XY) = P(X \cup Y)/P(X)$ olarak ifade edilebilir. Güven ölçütü, destek ölçütünü tamamlayan bir şekilde çalışması sebebiyle çoğunlukla birlikte kullanılmaktadır.

Bu tez çalışması kapsamında destek ve güven ölçütlerine ek olarak mekânsal yaygınlıkları ortaya çıkaracak bir ölçüt olan istasyon yaygınlık ölçütü önerilmektedir. Bu ölçüt kullanılarak elde edilen kuralların yeterli sayıda istasyonda (mekânda) geçerli olup olmadığı ortaya konulmakta ve bu sayede kuralların yaygınlıkları daha genel hale getirilebilmektedir. Bu ölçütle ilgili detaylar ilerleyen bölümlerde verilmiştir.

1.8. Literatür Taraması

Bu tez çalışması kapsamında hem mekân-zamansal birliktelik analizi incelenmiş bu çalışmaların hidrolojik ve iklimsel veritabanlarının analizinde uygulanabilirliği incelenmiştir. Önerilen algoritmanın performansı ve ürettiği sonuçların kullanılabilirliği değerlendirilmiştir. Bu sebeple aşağıda mekân-zamansal birliktelik analizi çalışmaları ve hidrolojik ve iklimsel veritabanlarında veri madenciliği uygulamaları incelenmiştir.

Şekil 1.4'de bu tez çalışmasına ait çalışma konuları gösterilmiştir. Öncelikli olarak çalışılan veri ile ilgili olarak bir sınıflandırma söz konusu olmaktadır. Bu ayrımdan sonra, kullanılan veri madenciliği yöntemi değerlendirilmiştir. Ardından bu çalışmada kullanılan birliktelik analizi ile ilgili olarak alt çalışma alanları sıralanmış ve bu çalışmanın hangi kapsamda yer aldığı ifade edilmiştir.



Şekil 1.4. Literatür taramasında tez çalışmasının yeri

1.8.1. Mekân-Zamansal Birliktelik Analizi

Artan boyutlardaki verilerin güvenilir bir şekilde analizi, bu verilerden doğru sonuçlar elde etme ihtiyacı ve veri madenciliğinin istatistiksel yöntemlere göre daha hızlı oluşu, birliktelik örüntülerinin tespiti için veri madenciliği algoritmalarının geliştirilmesini zorunlu hale getirmiştir. Birliktelik analizi, ilk olarak Agrawal ve arkadaşları [4] tarafından önerilmiştir. Birliktelik analizi, hızlı algoritmaları ve ortaya koyduğu etkili

sonuçları ile kendini ispatlamış bir alandır. Birliktelik analizi konusunda, birbirinden farklı yaklaşımlar (Apriori, FP-Tree [2]) ve bunlara bağlı algoritmalar (ör., Apriori, OCD, CLOSET, CHARM, [2]) geliştirilmiştir.

Birliktelik analizi geniş bir ölçekte birbirinden farklı alanlarda kullanım imkanı bulmuştur. Her türlü alanda kolay bir şekilde uygulanabilirliği sayesinde popülerliği artan bir yaklaşım olmaktadır. Birliktelik analizini kullanan alanlara örnek olarak sağlık [5], tıp [6] ve ticareti [7, 8] gösterebiliriz. Bu alanlar dışında da çok farklı alanlarda kullanılabilmektedir.

Mekânsal ve zamansal verilerin gittikçe artması ve çok büyük ölçekli hale gelmesi, klasik birliktelik analizi algoritmalarının bu veriler üzerinde etkisiz kalması ve klasik sonuçların da tam olarak istenenleri verememesi olarak sonuçlanmıştır. Bu eksiklik araştırmacılar tarafından görülerek mekânsal, zamansal ve mekân-zamansal birliktelik analizi yöntemleri üzerinde çalışılmaya başlanmıştır. Veri kümelerinin içerdiği mekânsal ve zamansal parametreler de dikkate alınarak bu parametrelere bağlı kurallar ortaya çıkarılmaktadır. Bu şekilde klasik birliktelik analizi yöntemleri ile çözümlenemeyecek problemler ve çıkarılamayacak olan sonuçlar, verilerin mekânsal ve zamansal parametreleri de dikkate alınarak daha net bir şekilde çözümlenmesi sağlamıştır.

Mekân-zamansal veri madenciliği iki şekilde görülebilir: birincisi mekânsal veri üzerinde zamansal verinin ilave edilmesiyle elde edilen mekân-zamansal veri madenciliği, ikincisi ise zamansal veri üzerine mekânsal verinin ilave edilmesiyle elde edilen mekân-zamansal veri madenciliğidir. Genel kabul gören, bu çalışmada da tercih edilen, yaklaşım ikinci yaklaşım olmaktadır.

Mekânsal ve mekân-zamansal veri madenciliği, geniş bir şekilde incelenmiş, çeşitli algoritmalar önerilmiş [9-12] ve farklı yönlerini inceleyen kitaplar yazılmıştır [13, 14]. Mekân-zamansal birliktelik analizi konusu da yine detaylı bir şekilde incelenen bir alan olmuştur [9, 15-18]. Birliktelikler ve birlikte oluşumlar, kullanılabilirlik açısında araştırmacılara geniş bir perspektif vermektedir.

Mekân-zamansal algoritmalar geniş bir şekilde kendine çalışma alanı bulmasına rağmen, hidrolojik ve meteorolojik veriler üzerinde uygulaması çok azdır. Var olan algoritmaların hidrolojik ve meteorolojik verilerin karakteristiği sebebiyle doğrudan

uygulanamaması da bu alandaki çalışmaları zorlaştırmaktadır. Bu tez çalışması kapsamında, hidrolojik ve meteorolojik verilere rahatlıkla uygulanabilecek, yeni bir mekân-zamansal birliktelik analizi algoritması önerilmiştir.

1.8.2. Hidrolojik ve Meteorolojik Veritabanlarında Veri Madenciliği Uygulamaları

Hidrolojik ve meteorolojik veritabanlarında veri madenciliği, bugüne kadar öncelikli olarak tercih edilen bir yöntem olmamıştır. Ancak istatistik ve yapay zeka yöntemlerine göre ortaya koyduğu sonuçlar açısından faydaları göz önüne alındığında geniş bir şekilde kullanılabileceği görülmektedir. Aşağıda literatürde mevcut olan sınırlı sayıda çalışma özetlenmektedir.

İklimsel parametrelerle yerleşim alanları ve orman alanlarının arasındaki ilişkiler [19-22] tarafından çalışılmıştır. Bu çalışmalarda asıl amaç, yerleşim alanlarındaki değişimin iklime olan etkisini ortaya koymaktır. Bu kapsamda yapılan çalışmalarda, etkin algoritmalar geliştirmiş ve bu algoritmaların sonuçlarını incelemişlerdir. Yapılan çalışmalar uydu verileri ve iklimsel verilerin ortak incelenmesini gerektirdiği için uygulama aşamasında hesaplama karmaşıklığı yüksektir.

Karbon başta olmak üzere atmosferdeki gazların döngüleriyle iklimsel parametreler arasındaki ilişkiler [23-24] tarafından çalışılmıştır. Bu konuda yapılan çalışmaların temel yaklaşımı, küresel ısınma üzerinde etkisi olduğu düşünülen gazların gerçekten iklim değişikliklerine sebep olup olmadıkları üzerine yoğunlaşmıştır. Yapılan çalışmaların büyük çoğunluğu iklim değişikliklerinde, denizlerin yükselmesinde ve kutup buzullarının erimesinin birbiriyle ilişkili olduğu sonucuna ulaşmıştır.

Okyanus iklim indeksleri, okyanusların belirli bölgelerinin karakteristiğini ifade eden zaman serileridir. Bu indeksler, okyanusların karalardaki iklime olan etkisini kestirmede etkili bir araçtır. Veri madenciliği yöntemleri, var olan okyanus iklim indekslerine ek olarak, yeni okyanus iklim indeksleri ortaya çıkarmada kullanılabilir. Bu konuda yapılan çalışmalarda [25-27] faydalı ve etkili yeni okyanus iklim indeksleri elde edilmiştir.

Ürünlerin yetiştirilmesi ve yıl içerisinde ne kadar ürün toplanacağı iklimsel parametrelerin değişimine bağımlıdır. Dolayısıyla üretimdeki tahminler de iklimsel

parametrelere göre şekillendirilir. Bu konuda yapılan çalışmalarda [28-30] iklimsel parametreler kullanılarak ne kadar ürün elde edilebileceği tespit edilmeye çalışılmıştır.

Kuraklıklar ve seller, ülke ekonomisi başta olmak üzere her türlü yapıya çok büyük zararlar veren iklim olaylarıdır. Bu olaylar sebebiyle her yıl dünyada çok büyük ölçüde maddi hasarlar meydana gelmektedir. Gelişen bilişim teknolojisinin de yardımıyla bu kuraklık ve sel olayları belli bir ölçüde önlenebilir. Bu kapsamda veri madenciliği yöntemleri de kullanılabilir. İklimsel parametrelerin kuraklığa veya sele sebep olacak değişimlerinin tespiti için çeşitli veri madenciliği algoritmaları geliştirilmiştir [31-34].

İklim indeksleri dünyanın belirli bölgelerindeki iklimsel parametreler değerlendirilerek hesaplanmış, bölgesel veya küresel ölçekte etkili olabilen iklim parametreleridir. Bu parametreler kullanılarak belli bölgelerdeki iklim değişiklikleri tespit edilebilir. Sıralı veri madenciliği teknikleri kullanılarak iklim indekslerinin iklimsel parametreler üzerindeki etkisi belirlenebilir [31, 34-36].

Veri madenciliği yöntemleri meteorolojik parametrelerin hidrolojik parametrelere olan etkisini ortaya çıkarmak için kullanılabilir. Birliktelik analizi, meteorolojik parametreler arasındaki ilişkileri, birlikte oluşumları net bir şekilde ortaya koyabilir [37- 40]. Bu tez kapsamında yapılan çalışmalar bu kategoride değerlendirilmektedir ve var olan yaklaşımlara ek yeni yaklaşımlar geliştirilmiştir.

2. BÖLÜM

HİDRO-METEOROLOJİK BİRLİKTELİKÖRÜNTÜ MADENCİLİĞİ

Bu tez çalışması kapsamında, mekân-zamansal birliktelik örüntü madenciliği algoritmaları incelenmiş ve iki adet mekân-zamansal birliktelik örüntü madenciliği algoritması önerilmiştir. Ayrıca algoritmalar tarafından elde edilen sonuçların değerlendirilmesi için bir değerlendirme ölçütü önerilmiş ve kullanılmıştır.

Hidrolojik bir parametre olan akarsu akımları su yönetimi konusunda kullanılan önemli parametrelerden biridir. Akımın gelecek dönemlerdeki muhtemel davranışının tahmin edilmesi zordur. Ancak akım parametresinin tahmin edilebilmesi pek çok açıdan fayda sağlayabilecek bir durumdur. Akımın düşük veya yüksek olacağının öngörülebilmesi, seller veya kuraklıklar için önceden önlem alınmasını sağlayarak ülke ekonomisine büyük zararlar veren bu tür doğa olaylarının yıkıcı etkisi belli bir ölçüde engellenebilecektir. Yağış, sıcaklık gibi meteorolojik parametrelerin akım üzerinde etkisi olduğu bilinmektedir. Dolayısıyla bu parametrelerle akım parametresi arasındaki ilişkiler tespit edildiği takdirde akımın davranışı belli bir yanılma payıyla ortaya konulabilecektir. Bu tez çalışmasında, akım parametresine etki eden meteorolojik parametreler içerisinden en yüksek etkisi olduğu bilinen yağış, sıcaklık, rüzgar hızı ve bağıl nem parametreleri ile akım arasındaki ilişkiler incelenmiş ve meteorolojik parametrelerin değişiminden akım parametresinin nasıl etkilendiği incelenmiştir.

Hidrolojik ve meteorolojik veritabanlarında çalışmak, uzun zaman almaktadır ve sonuçlarının yorumlanması ise karmaşıktır. Bu açıdan veri madenciliği yöntemleri

hidrolojik ve meteorolojik konularda çalışmalara farklılık katmakta, işlemleri hızlandırmakta ve sonuçlarının anlamlandırılması aşamasında daha iyi yaklaşım sergileyebilmektedir. Bu sebeple veri madenciliği bu alandaki açığı kapatmada etkin bir şekilde uygulanabilmektedir. Bu tez çalışmasında önerilen her iki algoritma da, meteorolojik parametrelerin hidrolojik parametre üzerindeki etkisini ortaya çıkarmada etkin olmakla beraber performanslı çalışması sebebiyle tercih edilen ve önerilen algoritma HMP-Miner algoritması olmaktadır.

Bu bölümde, öncelikle hidro-meteorolojik birliktelik analizi ile ilgili temel tanımlar ve kavramlar verilmiştir. Daha sonra hidro-meteorolojik birliktelik örüntü madenciliğine ait problem tanımlaması yapılmıştır. Ardından önerilmiş olan hidro-meteorolojik örüntü madenciliği algoritmalarına ait kaba kodlar ve algoritmaların açıklamaları verilmiştir. Son olarak daha etkin ve avantajlı olduğu ortaya konulan HMP-Miner algoritmasının çalışılmış olan veri kümesinin küçük bir parçası üzerindeki örnek çalışması gösterilmiştir.

2.1 Temel Tanımlar

2.1.1. Birliktelik Kuralı

Birliktelik kuralı, şu şekilde tanımlanabilir. $I=\{i_1, i_2, \ldots, i_n\}$, n adet özellikler kümesi olsun ve bu kümeye *elemanlar(items)* diyelim. $D=\{t_1, t_2, \ldots, t_m\}$ de veri tabanı adı verilen kayıtlar dizisi olsun. D'deki her bir kayıt, kendine ait bir kayıt numarası ve I'nın bir alt kümesini içermektedir. **Birliktelik kuralı**, $X \rightarrow Y$ formunun bir versiyonu olarak tanımlanmaktadır. Bu formda $X, Y \subseteq I$ ve $X \cap Y = \emptyset$ olmak zorundadır. X eleman seti birliktelik kuralının öncül, Y eleman seti ise ardıl kısmı olarak tanımlanmaktadır. Bir birliktelik örüntüsü P ise {X, Y} şeklindeki bir seti ifade etmektedir.

Birliktelik kuralında, kuralın şart kısmını ifade eden öncül X değeri, tek bir parametre olabileceği gibi birden fazla parametre de içerebilmektedir. Benzer şekilde kuralın sonuç kısmını ifade eden ardıl Y değeri de tek bir parametre olabileceği gibi birden fazla parametre de içerebilir. Bu çalışmada meteorolojik parametrelerin hidrolojik parametreye olan etkisi incelendiği için kuralın öncül (şart) kısmında meteorolojik parametreler, ardıl (sonuç) kısmında ise hidrolojik parametre bulunmaktadır.

2.1.2. Destek Ölçütü

Verilen bir P örüntüsü ve S_i istasyonuna ait D_i veri kümesinde, ***P* örüntüsünün S_i istasyonundaki destek değeri**, P örüntüsünü içeren kayıtların sayısının S_i istasyonuna ait D_i veri kümesindeki toplam kayıt sayısına oranıdır. S_i istasyonundaki P örüntüsüne ait destek değeri aşağıdaki şekilde formulize edilebilir:

$$destek(P, S_i) = (P'yi\ içeren\ kayıtlar)/(S_i\ istasyonundaki\ toplam\ kayıtlar) \quad (2.1)$$

Destek ölçütü, birliktelik analizine ait pek çok algoritmada kullanılabilen ve örüntülerin yaygınlığını belirlemede etkin bir ölçüttür. Kullanıcı tarafından belirlenmektedir ve eldeki örüntünün veri kümesi içerisindeki yaygınlığını test etmektedir.

2.1.3. İstasyon içi Örüntü

P örüntüsü S_i istasyonuna ait veri kümesi içerisinde bulunuyor ise, bu örüntü S_i istasyon içi örüntü olarak isimlendirilir.

2.1.4. İstasyon içi Yaygın Örüntü

Eğer S_i istasyonuna ait P örüntüsü kullanıcı tarafından verilen minimum destek değeri (*min_destek*) kadar veya daha yüksek ise P örüntüsü **yaygın (destek yaygın)** olarak ifade edilir.

2.1.5. İstasyonlar arası Örüntü

P örüntüsü, istasyonlar kümesi S içerisinde en az bir istasyonda destek yaygın olarak bulunuyor ise, bu örüntü istasyonlar arası örüntü olarak isimlendirilir.

2.1.6. İstasyon Yaygınlık Ölçütü

Verilen bir P örüntüsü ve $D=\{D_1, D_2, \ldots, D_n\}$ veri tabanında, ***P* örüntüsünün istasyon yaygınlığı**, P örüntüsünü içeren istasyonlarının sayısının D veri tabanındaki toplam istasyon sayısı S'e oranıdır. P örüntüsüne ait istasyon yaygınlık değeri aşağıdaki şekilde formulize edilebilir:

$$istasyon(P, D) = (P'yi\ içeren\ istasyonlar/(Toplam\ istasyon\ sayısı(S)) \quad (2.2)$$

İstasyon yaygınlık ölçütü bu tez çalışmasında önerilmiş bir ölçüttür ve elde edilen kuralın bölgesel bir kural olabilmesi için sadece belirli bir istasyonda yaygın olmasının yeterli olmadığı ve ayrıca bölgedeki diğer istasyonların büyük bir kısmında da yaygın

olması gerektiği düşüncesiyle önerilmiştir. İstasyon yaygınlık ölçütünün bölgesel kural çıkarımında çok etkin bir şekilde kullanılabildiği ve etkin bir yaygınlık ölçütü olduğu yapılan çalışmalarda ortaya konulmuştur.

2.1.7. İstasyonlar arası Yaygın Örüntü

Eğer *P* örüntüsü, *S* istasyonlar seti içerisinde, en az kullanıcı tanımlı minimum istasyon değeri (*min_istasyon*) kadar veya daha fazla istasyonda destek yaygın ise, *P* örüntüsü **istasyonlar arası yaygın** olarak ifade edilir.

2.1.8. Hidro-meteorolojik Birliktelik Örüntüsü

Verilen bir *P* örüntüsü ve *D* veri tabanı içerisinde, **hidro-meteorolojik birliktelik örüntüsü**, birliktelik kuralının öncül ve/veya ardıl kısmında en az bir hidrolojik veya meteorolojik parametre içeren, hidrolojik ve meteorolojik parametrelere ait eleman kümelerinin(*itemsets*) bir alt kümesi olarak tanımlanabilir.

Bu çalışmanın esas amacı, meteorolojik parametrelerin hidrolojik parametreler üzerindeki etkisini ortaya çıkarmak olduğu için, bu çalışmada, hidro-meteorolojik örüntülerin öncül (şart) kısmı meteorolojik parametreleri ve ardıl (sonuç) kısmı ise hidrolojik parametreleri içermektedir. Örneğin; "**Eğer** yağış belli bir değerde oluşuyor **ise** akım da belli bir değerde oluşur." şeklinde kurallar elde edilir.

2.1.9. Hidro-meteorolojik Yaygınlık İndeksi

Verilen bir *P* örüntüsü, $S=\{S_1, S_2, \ldots, S_n\}$ istasyonları ve *S* istasyonlarına ait $D=\{D_1, D_2, \ldots, D_n\}$ veri tabanları üzerinde, *P* örüntüsüne ait **hidro-meteorolojik yaygınlık indeksi**, istasyon ve destek yaygınlık ölçütlerinin birleşimiyle aşağıdaki şekilde formulize edilebilir:

$$istasyon_{S_i \in S}(destek(P, S_i) \geq \min_destek) \qquad (2.3)$$

Hidro-meteorolojik yaygınlık indeksi, sadece *P* örüntüsünün *min_destek* eşik değerini sağladığı istasyonları hesaba katmaktadır.

2.1.10. Hidro-meteorolojik Yaygın Örüntü

Verilen bir *P* örüntüsü, destek ve istasyon yaygınlık eşik değerleri (sırasıyla *min_destek* ve *min_istasyon*) olmak üzere, eğer *P* örüntüsü kullanıcı tanımlı destek ve istasyon yaygınlık eşik değerlerini sağlıyorsa, bu örüntü **hidro-meteorolojik olarak yaygındır**. Hidro-meteorolojik yaygın örüntülerin hesaplanmasında kullanılan formül aşağıdaki şekilde ifade edilebilir:

$$istasyon_{S_i \in S}(destek(P, S_i) \geq min_destek) \geq min_istasyon \qquad (2.4)$$

2.1.11. Güven Ölçütü

Eldeki bir $X \rightarrow Y$ kuralı ($P=\{X, Y\}$ olmak üzere) üzerinde, $X \rightarrow Y$ kuralının **güven** değeri, *Y* içerisindeki parametrelerin, *X* içerisindeki parametreleri içeren kayıtlarda ne sıklıkla geçtiğini belirlemektedir. Güven ölçütü aşağıdaki şekilde formülize edilebilir:

$$güven(X \rightarrow Y) = destek(X \cup Y)/destek(X) \qquad (2.5)$$

Güven ölçütü, $P(Y \mid X)$ olasılığını bulmak için kullanılmaktadır. Bu çalışma için *Y* hidrolojik parametreyi, *X* de meteorolojik parametreler setini ifade etmektedir.

Elde edilen hidro-meteorolojik kural, eğer kuralın güven değeri, minimum güven (*min_güven*) eşik değerine eşit veya bu değerden yüksekse, kural anlamlıdır denir.

2.2. Hidro-meteorolojik Birliktelik Örüntü Madenciliği Problem Tanımlaması

Verilenler:

- Hidro-meteorolojik veri tabanı, *D*.
- Minimum destek eşik değeri, *min_destek*.
- Minimum istasyon yaygınlık eşik değeri, *min_istasyon*.
- Minimum güven eşik değeri, *min_güven*.

Algoritma Çıktısı:

- *min_destek*, *min_istasyon* ve *min_güven* eşik değerlerini sağlayan hidro-meteorolojik birliktelik örüntü.

Hedef:

- Hesaplama maliyetini düşürmek.

Sınırlamalar:

- Verilen parametrelerle doğru ve tam hidro-meteorolojik kurallar bulunması.
- Kuralın ardıl (sonuç) kısmı hidrolojik parametre olması.

2.3. Hidro-meteorolojik Birliktelik Örüntü Madenciliği

Hidro-meteorolojik örüntü madenciliği veriler içerisindeki mekân ve zaman bilgisini de dikkate alarak hidrolojik ve meteorolojik parametreler arasında birliktelik örüntüleri elde etmeyi hedeflemektedir. Mekân ve zaman bilgileri de örüntü keşfi işlemine dahil edildiği için algoritmaların karmaşıklığı ve hesaplama maliyetleri artmaktadır. Bu durumda mekân-zamansal bir veri tabanı üzerinde etkili bir hidro-meteorolojik örüntü madenciliği algoritması geliştirme ihtiyacı ortaya çıkmaktadır.

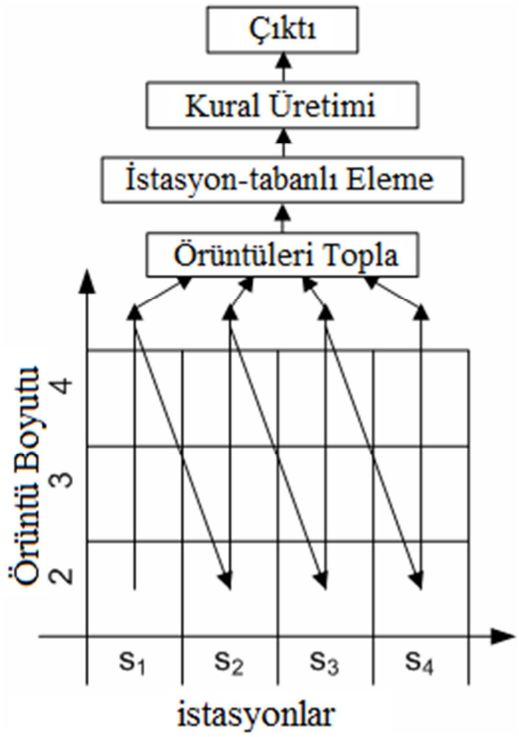
Bu tez çalışması kapsamında, hidrolojik ve meteorolojik parametreler arasındaki mekân-zamansal ilişkiler dikkate alınarak örüntüler elde edilmektedir. Diğer bir deyişle meteorolojik parametrelerin hidrolojik parametre üzerindeki etkisi incelenmektedir. Meteorolojik parametrelerin hidrolojik parametre üzerindeki etkisi, verilerin büyük boyutlu olması ve veri kümesinin çok boyutlu olması sebebiyle uzun süren bir işlem olmaktadır. Önerilen HMP-Miner algoritması, bu zorlukların üstesinden gelecek şekilde geliştirilmiştir.

Bu kısımda hidro-meteorolojik birliktelik örüntülerini keşfetmede kullanılabilecek olan iki algoritma gösterilmiştir: Toy (Naive) yaklaşım ve HMP-Miner algoritması. Aşağıda, her iki algoritmaya ait kaba kodlar ve algoritmaların çalışma mekanizması açıklanmıştır.

2.3.1. Toy (Naive) Yaklaşım

Toy (naive) yaklaşım, birden fazla istasyon üzerinde hidro-meteorolojik birliktelik örüntülerinin madenciliği için A-priori algoritmasının genişletilmiş bir versiyonudur. En başta, bir bölgedeki her bir istasyon için ayrı ayrı tüm boyutlardaki yaygın örüntüleri keşfetmekte ve son işlem adımı olarak yeterli sayıda istasyonda yaygın olmayan hidro-meteorolojik örüntüleri elemektedir. Yaygın olmayan hidro-meteorolojik örüntüler

elendikten sonra, minimum güven eşik değeri *min_güven*'i sağlayan anlamlı örüntüler keşfedilmektedir. Bu algoritmanın sınırlaması, yaygın olmayan hidro-meteorolojik aday örüntülerinin son işlem adımından önce üretilmesidir. Ancak eğer bir *k*-boyutlu örüntü hidro-meteorolojik olarak yeterli sayıda istasyonda yaygın değilse, bu örüntü *k+1* boyutlu adaylar örüntülerin üretiminde kullanılmaması gerekmektedir. Yaygın olmayan örüntülerin gereksiz üretimi algoritmanın karmaşıklığını artırmaktadır.



Şekil 2.1. Toy (naive) yaklaşımın çalışma mantığı

Şekil 2.1'de toy yaklaşımın çalışma mekanizması verilmiştir. Şekilde de görülebildiği üzere, her bir istasyon için tüm hidro-meteorolojik aday örüntüler üretilmektedir. Ardından üretilen tüm örüntüler toplanmakta ve daha sonra bu örüntüler üzerinde istasyon tabanlı eleme işlemi yapılmaktadır. Eleme işleminden sonra kural üretimi aşamasında *min_güven* değerini sağlayan kurallar seçilmektedir. Çıktı aşamasında ise elde edilen bu kurallar algoritma sonucu olarak verilmektedir.

Toy yaklaşıma ait kaba kod ve açıklaması Tablo 2.1'de verilmiştir.

Tablo 2.1. Toy yaklaşıma ait kaba kod

Girdiler:

min_destek: Minimum destek eşik değeri

min_istasyon: Minimum istasyon yaygınlık indeksi eşik değeri

min_güven: Minimum güven eşik değeri

D: Hidro-meteorolojik veritabanı

Çıktı:

min_destek, min_istasyon ve *min_güven* eşik değerlerini sağlayan yaygın ve anlamlı hidro-meteorolojik örüntüler.

Değişkenler:

İS: İstasyon sayısı.

A_k: *k* boyutlu aday örüntüler seti

Y_k: *k* boyutlu destek yaygın örüntü seti

AHM: aday hidro-meteorolojik örüntü seti

HM: hidro-meteorolojik örüntü seti

HM_Örüntüler: Anlamlı(güven yaygın) hidro-meteorolojik örüntü seti

Algoritma:

1. başlangıç; $k=0$, $L_0=D$
2. **for** $s=1$ **to** *İS*
3. **while** L_k boş olmadığı sürece
4. **if** $k=0$ **then**
5. $A_{k+1}(s)=birli_üret(D(s))$
6. **else**
7. $A_{k+1}(s)=aday_üret(Y_k(s))$
8. **end if**
9. $A_{k+1}(s)=destek_hesapla(A_{k+1}(s), D(s))$
10. $Y_{k+1}(s)=destek_eleme(A_{k+1}(s), min_destek)$
11. $k++$
12. **end while**
13. **end for**
14. *AHM=istasyon_hesapla(L)*
15. *HM=istasyon_eleme(AHM, min_istasyon)*
16. *HM_Örüntüler=anlamlı_örüntüleri_keşfet(HM, min_güven)*
17. **return** *HM_Örüntüler*

2. adım ile 13. adım arasında, toy yaklaşım her bir istasyon için hidro-meteorolojik örüntüleri üretmektedir. 14 ve 15. adımlarda yaygın hidro-meteorolojik örüntüler keşfedilmektedir ve 16. adımda da hidro-meteorolojik kurallar üretilmekte ve 17. adımda bu kurallar çıktı olarak verilmektedir. Algoritmada kullanılan fonksiyonların açıklaması aşağıda verilmiştir.

***birli_üret* fonksiyonu (Adım 5):** Bu fonksiyon bir *S* istasyonuna ait veri kümesini girdi olarak alarak istasyona ait *1* boyutlu aday örüntüleri çıktı olarak vermektedir. *1* boyutlu aday örüntüler, *S* istasyonunda bulunan parametrelerin özellikleri olmaktadır.

***aday_üret* fonksiyonu (Adım 7):** Bu fonksiyonda yaygın *k* boyutlu Y_k örüntülerden *k+1* boyutlu aday A_{k+1} örüntüler üretilmektedir. 2. ve 12. adımlar arasındaki **for** döngüsü, her bir istasyon için destek yaygın hidro-meteorolojik örüntüler üretmek için çalıştırılmıştır.

***destek_hesapla* fonksiyonu (Adım 9):** Bu fonksiyon *A* aday örüntülerine ait destek değerlerini hesaplamaktadır. Fonksiyonun ait girdiler, aday örüntüler ve *S* istasyonuna ait veri kümesidir. Fonksiyona ait çıktı ise aday örüntülerin destek değerleridir.

***destek_eleme* fonksiyonu (Adım 10):** *destek_hesapla* fonksiyonu tarafından üretilen destek değerleri arasında minimum destek eşik değeri *min_destek* değerini sağlamayan aday örüntüler, bu fonksiyon tarafından elenmektedir.

***istasyon_hesapla* fonksiyonu (Adım 14):** Bu fonksiyonda her bir hidro-meteorolojik örüntü için istasyon destek değerleri hesaplanmaktadır.

***istasyon_eleme* fonksiyonu (Adım 15):** Bu fonksiyonda *istasyon_hesapla* fonksiyonu tarafından hesaplanan istasyon destek değerleri arasından minimum istasyon eşik değeri *min_istasyon* değerini sağlamayan hidro-meteorolojik örüntüler elenmektedir.

***anlamlı_örüntüleri_keşfet* fonksiyonu (Adım 16):** Bu fonksiyonda *HM* içerisindeki örüntüler arasından minimum güven eşik değeri *min_güven* değerini sağlayan hidro-meteorolojik örüntüler keşfedilmektedir. Bu fonksiyon öncelikle her bir zaman dilimi için ilgilenilen kuralları üretmektedir. Bu çalışmada, ilgilenilen kurallar kuralın ardıl (sonuç) kısmında hidrolojik parametreyi içeren kurallar olmaktadır. Kural üretiminden sonra, minimum güven eşik değeri *min_güven* değerini sağlamayan kurallar elenmektedir. Ardından, kuralların istasyon yaygınlıkları hesaplanarak minimum güven

eşik değeri tabanlı eleme işlemi sonrası halen *min_istasyon* eşik değerini sağlayıp sağlamadığı tekrar kontrol edilmektedir. Eğer kurallar *min_istasyon* eşik değerini sağlamıyorsa, bu hidro-meteorolojik kurallar elenmektedir.

Son olarak 17. adımda yaygın ve anlamlı hidro-meteorolojik kurallar algoritma tarafından döndürülmektedir.

Toy yaklaşımın sınırlaması istasyonlar arası yaygın olmayan hidro-meteorolojik örüntülerin son işlem adımından önce gereksiz üretimidir. Bu sebeple toy yaklaşım uzun sürmekte ve algoritmanın hesaplama karmaşıklığını artırmaktadır.

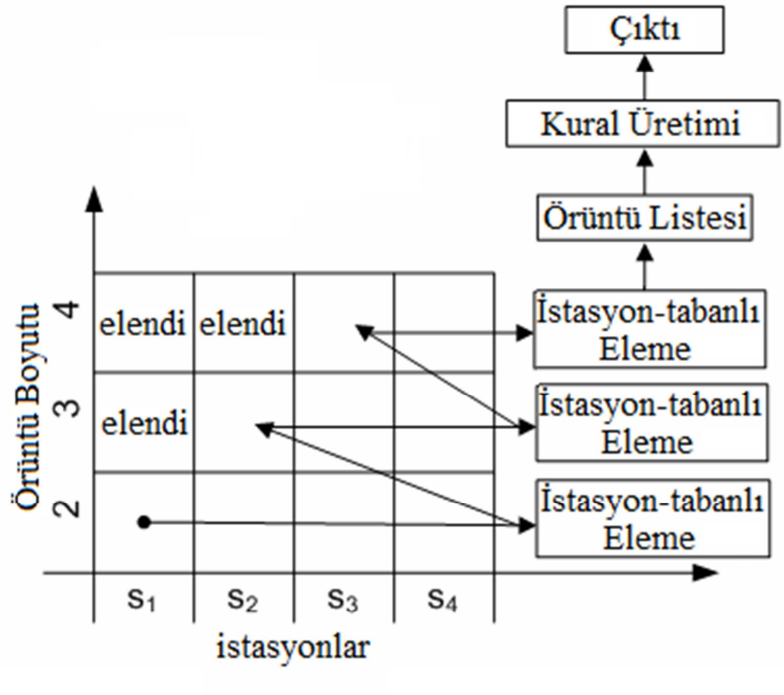
2.3.2. HMP-Miner Algoritması

Toy yaklaşımın gereksiz hidro-meteorolojik örüntü üretimi sınırlamasının üstesinden gelmek için HMP-Miner algoritması önerilmiştir. Bu amaçla, her bir *k* boyutlu örüntü üretimi sonrasında tüm istasyonlar için istasyonlar arası yaygın olmayan örüntülerin eleme işlemi uygulanmıştır. Toy yaklaşımın aksine, HMP-Miner algoritması hidro-meteorolojik örüntüleri keşfetmek için son işlem adımını beklememektedir.

HMP-Miner algoritmasının temel stratejisi, istasyon yaygınlık eleme işlemini mümkün olduğu kadar erken algoritmaya dahil etmektir. Algoritma, ilk olarak, her bir istasyon için *k* boyutlu yaygın örüntüleri üretmekte ve daha sonra üretilen bu örüntüler üzerinde istasyonlar arası yaygınlık eleme adımını uygulayarak *k* boyutlu istasyonlar arası yaygın olmayan hidro-meteorolojik örüntüleri elemektedir. Elde edilen istasyonlar arası yaygın *k* boyutlu hidro-meteorolojik örüntüler kullanılarak *k+1* boyutlu aday örüntüler üretilmektedir. Bu işlem, tüm boyutlardaki yaygın hidro-meteorolojik örüntüler keşfedilinceye kadar devam ettirilir. Bu algoritma yardımıyla, istasyonlar arası yaygın olmayan hidro-meteorolojik örüntülerin üretimi engellenmiş oldu ve bu sayede algoritmanın hesaplama maliyetini artıran bir etmen ortadan kaldırıldı.

Şekil 2.2'de HMP-Miner algoritmasına ait temel çalışma mekanizması verilmiştir. Şekilde görülebileceği üzere, algoritmanın her boyuttaki aday üretimi aşamasında istasyon tabanlı eleme işlemi eklenmiştir. Bu şekilde yeterli sayıda istasyonda yaygın olarak bulunmayan bir örüntü üst boyutlara taşınmamakta ve bu sayede bölgesel kural çıkarımı aşamasında algoritmanın etkinliği ve çalışma hızı artmaktadır. Son olarak elde edilen kurallar üzerinde minimum güven değeri tabanlı bir eleme işlemi yapılmakta ve

bu adımı da geçen hidro-meteorolojik örüntüler kural olarak değerlendirilmektedir. Elde edilen bu kurallar son olarak algoritma çıktısı olarak verilmektedir.



Şekil 2.2. HMP-Miner algoritmasına ait çalışma mantığı

HMP-Miner algoritmasına ait kaba kod ve algoritmanın açıklaması Tablo 2.2'de verilmiştir.

HMP-Miner algoritmasına ait kaba kodda da açıkça görülebileceği üzere öncelikle *k* boyutlu aday hidro-meteorolojik örüntülerin incelenmesi için bir döngü kurulmuş daha sonra bu döngü içerisinde istasyonlara ait örüntü çıkarımı döngüsü yerleştirilmiştir. Bu yapı sayesinde her adımda aday hidro-meteorolojik örüntüler incelenmekte ve yaygın olan hidro-meteorolojik örüntülerden bir üst aday örüntülerin üretimi yapılmaktadır.

Algoritmaya ait kaba kodda kullanılan fonksiyonlar ve işlevleri toy yaklaşıma ait fonksiyonlarla aynı olduğu için, bu kısımda ayrıca fonksiyonlara ait açıklama yapılmamıştır.

Tablo 2.2. HMP-Miner algoritmasına ait kaba kod

Girdiler:

min_destek: Minimum destek eşik değeri

min_istasyon: Minimum istasyon yaygınlık indeksi eşik değeri

min_güven: Minimum güven eşik değeri

D: Hidro-meteorolojik veritabanı

Çıktı:

min_destek, *min_istasyon* ve *min_güven* eşik değerlerini sağlayan yaygın ve anlamlı hidro-meteorolojik örüntüler.

Değişkenler:

İS: İstasyon sayısı

A_k: *k* boyutlu aday örüntüler seti

Y_k: *k* boyutlu destek yaygın örüntü seti

AHM_k: *k* boyutlu aday hidro-meteorolojik örüntü seti

HM_k: *k* boyutlu hidro-meteorolojik örüntü seti

HM_Örüntüler: Anlamlı(güven yaygın) hidro-meteorolojik örüntü seti

Algoritma:

1. başlangıç; $k=0$, $AHM_0=D$
2. **while** AHM_k boş olmadığı sürece
3. **for** $s=1$ **to** *İS*
4. **if** $k=0$ **then**
5. $A_{k+1}(s)=birli_üret(D(s))$
6. **else**
7. $A_{k+1}(s)=aday_üret(AHM_k(s))$
8. **end if**
9. $A_{k+1}(s)=destek_hesapla(A_{k+1}(s), D(s))$
10. $Y_{k+1}(s)=destek_eleme(A_{k+1}(s), min_destek)$
11. **end for**
12. $AHM_{k+1}=istasyon_hesapla(Y_{k+1})$
13. $HM_{k+1}=istasyon_eleme(AHM_{k+1}, min_istasyon)$
14. k++
15. **end while**
16. $HM_Örüntüler=anlamlı_örüntüleri_keşfet(HM, min_güven)$
17. **return** *HM_Örüntüler*

2.4. HMP-Miner Algoritmasına Ait Örnek Çalışma

Bu kısımda, önerilen algoritmanın daha iyi anlaşılması için HMP-Miner algoritmasına ait örnek çalışma yer almaktadır. Öncelikle örnek çalışmada, çalışılacak olan veri kümesi tanıtılmış ve ardından algoritmanın çalışması bu veri kümesi üzerinde gösterilmiştir. Veri kümesinde 5 istasyona ait hidrolojik ve meteorolojik parametreler bulunmaktadır. Her bir istasyon için hidrolojik ve meteorolojik parametreler, 6 aylık orijinal ve ayrıklaştırılmış verileri içermektedir. Elde edilen bu veriler, parametrelerin günlük ölçüm değerlerinin aylık ortalamaları (yağış için aylık toplam) alınarak hesaplanmıştır. Parametre değerleri düşük (D), ortalama (O) ve yüksek (Y) olarak ayrıklaştırılmıştır. Tablo 2.3'de örnek veri kümesi gösterilmektedir. Veri kümesi hazırlama işlemleri ile ilgili detaylar 3. bölümde verilmiştir.

Buradan sonraki adımlarda, kurallarda, Rüzgar Hızı parametresi yerine *RH*, Nem parametresi yerine *N*, Yağış parametresi yerine *Y*, Sıcaklık parametresi yerine *S*, ve Akım parametresi yerine *A* kısaltmaları kullanılacaktır. Ayrıca D, O, Y kısaltmaları da parametrelerin sınıfını göstermek için düşük, orta ve yüksek yerine kullanılacaktır. Parametrelerin girdi değerleri olarak *min_destek* = 0.3, *min_istasyon* = 0.4 ve *min_güven*=0.3 olarak belirlenmiştir.

Tablo 2.3. Örnek çalışma için kullanılacak olan veri kümesi

	Aylar	İstasyon #1203	İstasyon #1221	İstasyon #1222	İstasyon #1224	İstasyon #1226
Yağış	1	1.49–D	3.02–O	3.60–Y	1.00–D	3.61–Y
	2	1.88–D	3.78–Y	3.67–Y	0.77–D	1.15–D
	3	3.18–Y	2.40–O	2.79–O	0.98–D	1.56–D
	4	1.84–D	0.73–D	1.31–D	1.78–O	2.43–O
	5	3.54–Y	4.64–O	1.78–D	4.94–Y	5.49–Y
	6	2.29–O	3.74–Y	2.73–O	3.61–Y	2.38–O
Sıcaklık	1	-0.99–D	6.04–D	6.34–D	-2.01–D	-2.58–D
	2	-0.17–D	4.81–D	5.76–D	0.14–D	-0.49–D
	3	7.45–O	11.74–O	10.70–O	7.40–O	6.18–O
	4	11.52–O	15.03–0	14.02–O	12.15–O	11.70–O
	5	13.78–Y	17.65–Y	17.18–Y	13.75–Y	12.91–Y

	6	17.83–Y	21.54–Y	21.61–Y	18.67–Y	17.90–Y
Bağıl Nem	1	86.55–Y	73.11–Y	78.89–D	83.72–Y	84.17–Y
	2	82.99–Y	77.05–Y	79.29–D	74.43–Y	78.12–Y
	3	66.58–D	63.20–D	78.75–D	56.59–D	68.19–O
	4	66.68–D	62.35–D	80.77–O	59.84–D	61.94–D
	5	76.04–O	71.27–O	84.22–Y	72.34–O	72.24–O
	6	71.84–O	71.64–O	84.78–Y	59.68–D	60.41–D
Rüzgar Hızı	1	1.41–O	1.70–O	3.73–O	1.91–D	1.83–O
	2	1.47–O	1.61–D	5.56–Y	3.08–Y	2.53–Y
	3	1.74–Y	1.65–O	2.51–D	2.47–O	2.01–O
	4	1.64–Y	1.99–Y	3.24–O	3.01–Y	1.69–D
	5	0.88–D	1.62–D	2.96–D	1.95–D	1.69–D
	6	1.33–O	1.73–O	3.24–O	2.61–O	1.85–O
Akım	1	5.78–D	118.00–D	22.00–D	7.41–D	4.67–D
	2	7.25–D	191.00–Y	35.40–O	7.71–D	6.31–D
	3	12.10–Y	147.00–O	48.20–Y	12.40–Y	14.50–O
	4	10.10–O	134.00–D	43.30–Y	7.88–O	16.50–O
	5	11.90–Y	194.00–Y	38.40–O	11.70–Y	46.30–Y
	6	8.83–O	132.00–D	19.50–D	6.90–D	12.20–O

HMP-Miner *1* boyutlu örüntülerin keşfedilmesi ile başlar. *1* boyutlu aday örüntüler Tablo 2.4'de gösterilmektedir. Her bir aday örüntünün tüm istasyonlardaki destek değerleri hesaplanır. İstasyonlar içerisinde minimum destek yaygınlık eşik değeri (*min_destek*) olan 0.3 değerini sağlamayan örüntüler elenir. Daha sonra kalan *1* boyutlu aday örüntülerin istasyon yaygınlık değerleri hesaplanır. Tablo 2.4'ün son sütunu örüntülere ait istasyon yaygınlık değerlerini göstermektedir. Tablo 2.4'den de görülebileceği tüm *1* boyutlu aday örüntüler minimum istasyon yaygınlık değeri (*min_istasyon*) olan 0.4 değerini sağladığı için *1* boyutlu aday örüntülerden hiçbiri elenmemektedir.

Tablo 2.4. *1* boyutlu örüntülerin üretimi (*min_destek*=0.3 ve *min_istasyon*=0.4)

Örüntü	Her bir istasyondaki örüntülere ait destek değerleri					İstasyon Yaygınlığı
	#1203	#1221	#1222	#1224	#1226	
{RH-D}	1/6-Elendi	2/6	2/6	2/6	2/6	4/5
{RH-O}	3/6	3/6	3/6	2/6	3/6	5/5
{RH-Y}	2/6	1/6-Elendi	1/6-Elendi	2/6	1/6-Elendi	2/5
{N-D}	2/6	2/6	3/6	3/6	2/6	5/5
{N-O}	2/6	2/6	1/6-Elendi	1/6-Elendi	2/6	3/5
{N-Y}	2/6	2/6	2/6	2/6	2/6	5/5
{Y-D}	3/6	1/6-Elendi	2/6	3/6	2/6	4/5
{Y-O}	1/6-Elendi	3/6	2/6	1/6-Elendi	2/6	3/5
{Y-Y}	2/6	2/6	2/6	2/6	2/6	5/5
{S-D}	2/6	2/6	2/6	2/6	2/6	5/5
{S-O}	2/6	2/6	2/6	2/6	2/6	5/5
{S-Y}	2/6	2/6	2/6	2/6	2/6	5/5
{A-D}	2/6	3/6	2/6	3/6	2/6	5/5
{A-O}	2/6	1/6-Elendi	2/6	1/6-Elendi	3/6	3/5
{A-Y}	2/6	2/6	2/6	2/6	1/6-Elendi	4/5

İkinci adımda, yaygın *1* boyutlu örüntüler birleştirilerek *2* boyutlu aday örüntüler elde edilmektedir. Bu çalışmada meteorolojik parametrelerin hidrolojik parametrelere olan etkisi incelendiği için, *2* boyutlu aday örüntülerden bir tanesi hidrolojik parametreyi, akım parametresini, içermektedir. Tablo 2.5'te *2* boyutlu örüntüler ve her bir istasyon için bu örüntülerin destek değerleri verilmiştir.

Tablo 2.5. *2* boyutlu aday örüntüler (*min_destek*=0.3 ve *min_istasyon*=0.4)

Örüntü	Her bir istasyondaki örüntülere ait destek değerleri					İstasyon Yaygınlığı
	#1203	#1221	#1222	#1224	#1226	
{RH-D, A-D}	-	0/6-Elendi	0/6-Elendi	1/6-Elendi	0/6-Elendi	0/5-Elendi
{RH-D, A-O}	-	-	1/6-Elendi	-	1/6-Elendi	0/5-Elendi
{RH-D, A-Y}	-	2/6	1/6-Elendi	1/6-Elendi	-	1/5-Elendi
{RH-O, A-D}	2/6	2/6	2/6	1/6-Elendi	1/6-Elendi	3/5

{RH-O, A-O}	1/6-Elendi	-	0/6-Elendi	-	2/6	1/5-Elendi
{RH-O, A-Y}	0/6-Elendi	0/6-Elendi	1/6-Elendi	1/6-Elendi	-	0/5-Elendi
{RH-Y, A-D}	0/6-Elendi	-	-	1/6-Elendi	-	0/5-Elendi
{RH-Y, A-O}	1/6-Elendi	-	-	-	-	0/5-Elendi
{RH-Y, A-Y}	1/6-Elendi	-	-	0/6-Elendi	-	0/5-Elendi
{N-D, A-D}	0/6-Elendi	1/6-Elendi	0/6-Elendi	0/6-Elendi	0/6-Elendi	0/5-Elendi
{N-D, A-O}	1/6-Elendi	-	1/6-Elendi	-	2/6	1/5-Elendi
{N-D, A-Y}	1/6-Elendi	0/6-Elendi	1/6-Elendi	1/6-Elendi	-	0/5-Elendi
{N-O, A-D}	0/6-Elendi	1/6-Elendi	-	-	0/6-Elendi	0/5-Elendi
{N-O, A-O}	1/6-Elendi	-	-	-	1/6-Elendi	0/5-Elendi
{N-O, A-Y}	1/6-Elendi	1/6-Elendi	-	-	-	0/5-Elendi
{N-Y, A-D}	2/6	1/6-Elendi	1/6-Elendi	2/6	2/6	3/5
{N-Y, A-O}	0/6-Elendi	-	1/6-Elendi	-	0/6-Elendi	0/5-Elendi
{N-Y, A-Y}	0/6-Elendi	1/6-Elendi	0/6-Elendi	0/6-Elendi	-	0/5-Elendi
{Y-D, A-D}	2/6	-	0/6-Elendi	2/6	1/6-Elendi	2/5
{Y-D, A-O}	1/6-Elendi	-	1/6-Elendi	-	1/6-Elendi	0/5-Elendi
{Y-D, A-Y}	0/6-Elendi	-	1/6-Elendi	1/6-Elendi	-	0/5-Elendi
{Y-O, A-D}	-	1/6-Elendi	1/6-Elendi	-	0/6-Elendi	0/5-Elendi
{Y-O, A-O}	-	-	0/6-Elendi	-	2/6	1/5-Elendi
{Y-O, A-Y}	-	1/6-Elendi		-	-	0/5-Elendi
{Y-Y, A-D}	0/6-Elendi	1/6-Elendi	1/6-Elendi	1/6-Elendi	1/6-Elendi	0/5-Elendi
{Y-Y, A-O}	0/6-Elendi	-	1/6-Elendi	-	0/6-Elendi	0/5-Elendi
{Y-Y, A-Y}	2/6	1/6-Elendi	0/6-Elendi	1/6-Elendi	-	1/5-Elendi
{S-D, A-D}	2/6	1/6-Elendi	1/6-Elendi	2/6	2/6	3/5
{S-D, A-O}	0/6-Elendi	-	1/6-Elendi	-	0/6-Elendi	0/5-Elendi
{S-D, A-Y}	0/6-Elendi	1/6-Elendi	0/6-Elendi	0/6-Elendi	-	0/5-Elendi
{S-O, A-D}	0/6-Elendi	1/6-Elendi	0/6-Elendi	0/6-Elendi	2/6	1/5-Elendi
{S-O, A-O}	1/6-Elendi	-	0/6-Elendi	-	0/6-Elendi	0/5-Elendi
{S-O, A-Y}	1/6-Elendi	0/6-Elendi	2/6	1/6-Elendi	-	1/5-Elendi
{S-Y, A-D}	0/6-Elendi	1/6-Elendi	1/6-Elendi	1/6-Elendi	0/6-Elendi	0/5-Elendi
{S-Y, A-O}	1/6-Elendi	-	1/6-Elendi	-	1/6-Elendi	0/5-Elendi
{S-Y, A-Y}	1/6-Elendi	1/6-Elendi	0/6-Elendi	1/6-Elendi	-	0/5-Elendi

Tablo 2.5'te, *2* boyutlu aday örüntüler ve bu örüntülerden seçilen ve elenenler gösterilmektedir. Tablonun daha iyi anlaşılabilmesi için *Y-D* ve *A-D* elemanları, (bu

elemanlar *{Y-D, A-D}* örüntüsünü oluşturmaktadır) birlikte, 1203 numaralı istasyonda, 6 ay içerisinden 2 ayda (1. ve 2. ayda) bulunmaktadır. *2* boyutlu örüntüler arasından destek değeri 0.3 olan *min_destek* eşik değerini sağlamayanlar elenmektedir. 1222 ve 1226 numaralı istasyonlarda örnek örüntü *{Y-D, A-D}* örüntüsüne ait destek değeri *min_destek* değerini sağlamamaktadır. 1221 numaralı istasyonda bu örüntü oluşmadığı için – işareti ile bu istasyon için bu örüntünün değerinin oluşmadığı gösterilir. Ardından örüntüye ait istasyon yaygınlık değeri hesaplanır. *{Y-D, A-D}* örüntüsüne ait istasyon yaygınlık değeri, 5 istasyon arasından 2'sinde (1203 ve 1224 numaralı istasyonlar) destek yaygın olduğu için 2/5'tir. Eğer istasyon yaygınlık değeri, *min_istasyon* eşik değeri olan 0.4 veya daha büyükse, örüntü hidro-meteorolojik olarak yaygın örüntü olur ve bu örüntü algoritmanın bir sonraki *k* boyutlu aday üretimi aşamasında kullanılır. *{Y-D, A-D}* örüntüsüne ait 2/5 olan istasyon yaygınlık değeri, *min_istasyon* eşik değeri olan 0.4 değerini sağladığı için bu örüntü hidro-meteorolojik olarak yaygın olarak seçilmektedir. *min_istasyon* eşik değerini sağlamayan örüntüler Tablo 2.5'te görüldüğü üzere elenmektedir. Tablo 2.6'da *2* boyutlu yaygın hidro-meteorolojik örüntüler verilmiştir.

Tablo 2.6. *2* boyutlu yaygın örüntüler (*min_destek*=0.3 ve *min_istasyon*=0.4)

Örüntü	Her bir istasyondaki örüntülere ait destek değerleri					İstasyon Yaygınlığı
	#1203	#1221	#1222	#1224	#1226	
{RH-O, A-D}	2/6	2/6	2/6	1/6-Elendi	1/6-Elendi	3/5
{N-Y, A-D}	2/6	1/6-Elendi	1/6-Elendi	2/6	2/6	3/5
{Y-D, A-D}	2/6	-	0/6-Elendi	2/6	1/6-Elendi	2/5
{S-D, A-D}	2/6	1/6-Elendi	1/6-Elendi	2/6	2/6	3/5

2 boyutlu yaygın hidro-meteorolojik örüntüler birleştirilerek *3* boyutlu aday örüntüler üretilmektedir (Tablo 2.7). *3* boyutlu yaygın örüntüler birleştirilerek *4* boyutlu aday örüntüler üretilmektedir (Tablo 2.8). Bu işlem daha fazla aday hidro-meteorolojik örüntü üretilemeyinceye kadar devam ettirilmektedir.

Tablo 2.7. *3* boyutlu aday örüntüler (*min_destek*=0.3 ve *min_istasyon*=0.4)

Örüntü	Her bir istasyondaki örüntülere ait destek değerleri					İstasyon Yaygınlığı
	#1203	#1221	#1222	#1224	#1226	
{N-Y, Y-D, A-D}	2/6	-	-	2/6	-	2/5
{N-Y, S-D, A-D}	2/6	-	-	2/6	2/6	3/5
{Y-D, S-D, A-D}	2/6	-	-	2/6	-	2/5

Tablo 2.8. *4* boyutlu aday örüntüler (*min_destek*=0.3 ve *min_istasyon*=0.4)

Örüntü	Her bir istasyondaki örüntülere ait destek değerleri					İstasyon Yaygınlığı
	#1203	#1221	#1222	#1224	#1226	
{N-Y, Y-D, S-D, A-D}	2/6	-	-	2/6	-	2/5

Örüntüler elde edildikten sonra, bir sonraki adımda kural üretimi aşaması gelmektedir. Bu aşamada, *min_güven* eşik değeri olan 0.3 değerini sağlayan hidro-meteorolojik örüntülerden anlamlı kural üretimi yapılmaktadır (Tablo 2.9). Her bir istasyonda *min_güven* eşik değerini sağlamayan tüm örüntüler elenmektedir. Tablo 2.9'da verilmiş olan kuralların tamamı *min_güven* eşik değerini sağlamaktadır ve güven değerine göre bir eleme işlemi yapılmamaktadır. Ardından kuralların istasyon yaygınlık değerleri kontrol edilmektedir. Kuralların istasyon yaygınlık değerleri Tablo 2.9'un son sütununda görülebilir. Bu aşamada da kuralların tamamı istasyon yaygınlık eşik değeri *min_istasyon* değerini (0.4'ü) sağladıkları için hiçbir kural elenmemektedir. Son olarak, hidro-meteorolojik olarak yaygın kurallar, *min_destek*, *min_istasyon* ve *min_güven* değerlerini sağlayan kurallar, algoritma tarafından döndürülmüştür. Üretilen sonuçlar Tablo 2.9'da gösterilmektedir.

Tablo 2.9. Güven-tabanlı eleme ve HMP-Miner algoritmasının çıktısı

Kural	Her bir istasyon için kuralların güven değerleri					İstasyon Yaygınlığı
	#1203	#1221	#1222	#1224	#1226	
RH-O→A-D	2/3	2/3	2/3	-	-	3/5
N-Y→A-D	2/2	-	-	2/2	2/2	3/5
Y-D→A-D	2/3	-	-	2/3	-	2/5
S-D→A-D	2/2	-	-	2/2	2/2	3/5
N-Y, Y-D→A-D	2/2	-	-	2/2	-	2/5
N-Y, S-D→A-D	2/2	-	-	2/2	2/2	3/5
Y-D, S-D→A-D	2/2	-	-	2/2	-	2/5
N-Y, Y-D, S-D→A-D	2/2	-	-	2/2	-	2/5

Tablo 2.9'daki şekilde algoritmanın çıktıları olarak 8 adet hidro-meteorolojik kural elde edilmiştir. Bu kuralları elde etmek için öncelikle destek ve istasyon yaygınlık hesaplaması yapılmış ardından kalan örüntüler için güven ve istasyon yaygınlık hesaplaması yapılmıştır. Destek ölçütü *X,Y* tarzında kurallar elde ederken, güven ölçütü *X, Y* örüntüsünden *X→Y* kuralları elde etmiştir. Bu şekilde **eğer-ise** yapısında kurallar ortaya çıkmış olmaktadır.

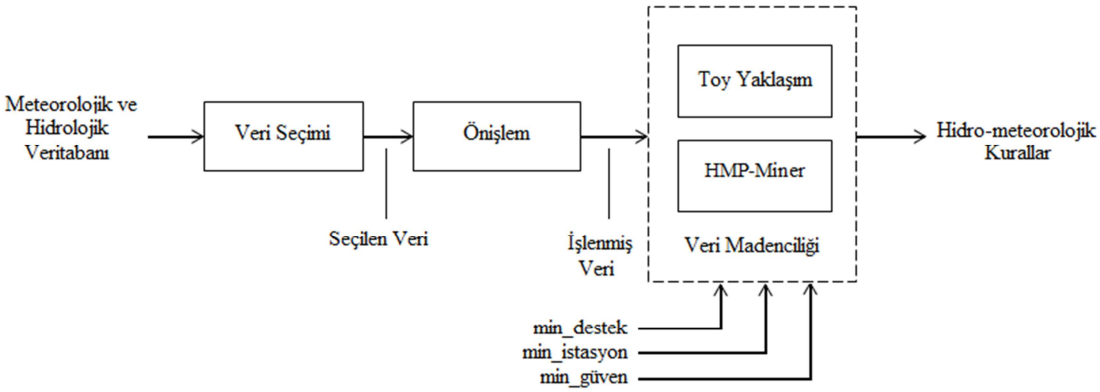
3. BÖLÜM

DENEYSEL ÇALIŞMA

3.1 Deney Düzeneği

Bu bölümde, tez çalışmasında önerilen algoritmalarla ilgili deneysel çalışmalar bulunmaktadır. Toy yaklaşım ile HMP-Miner algoritmasının performansı gerçek veriler üzerinde karşılaştırılmış ve bu karşılaştırma sonucu verilmiştir. Her iki algoritmada da 3 adet kullanıcı tanımlı değişken bulunmaktadır, *min_destek, min_istasyon* ve *min_güven*, ve bu üç değişkenden *min_destek* ve *min_istasyon* eşik değerlerinin algoritmaların performansı üzerinde etkisi bulunmaktadır. *min_güven* eşik değeri, örüntü üretimi aşamasında bulunmadığı ve sadece kural üretimi aşamasında her iki algoritmada da aynı şekilde kullanıldığı için bu değişken için performans karşılaştırması yapmaya gerek kalmamaktadır. Dolayısıyla bu değişkenlerin değerleri değiştirilerek algoritmaların performans değişimi gözlenmiştir. Deneyler aşağıdaki soruları cevaplamak için yapılmıştır:

- Farklı minimum destek eşik değerlerinde algoritmaların çalışma zamanları nasıl değişmektedir?
- Farklı minimum istasyon yaygınlık eşik değerlerinde algoritmaların çalışma zamanları nasıl değişmektedir?
- Veri kümesinde kullanılan farklı sayıdaki istasyonlar olması durumunda algoritmaların çalışma zamanları nasıl değişmektedir?



Şekil 3.1. Deney düzeneği

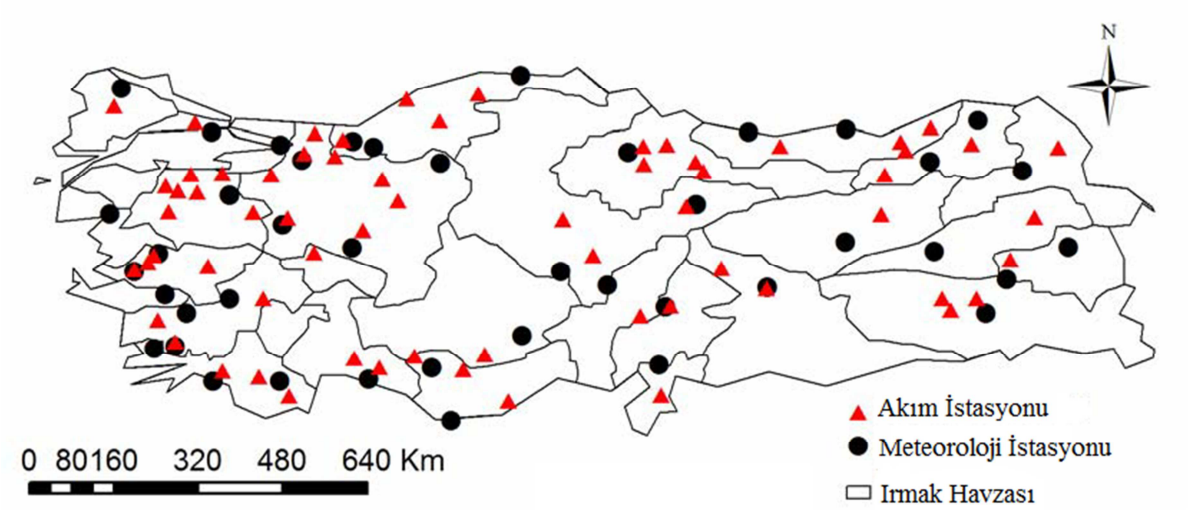
Deneyler, Intel® Core™ 2 Quad 2.66 GHz işlemci ve3 GB RAM'e sahip olan bir bilgisayar üzerinde yapılmıştır.

3.2 Kullanılan Veri Kümesi

Bu çalışmada kullanılan veriler hidrolojik ve meteorolojik parametrelerden oluşmaktadır. Kullanılan parametreler; yağış, sıcaklık, bağıl nem, rüzgâr hızı ve akım parametreleridir. Parametrelere ait ölçüm verileri iki grup halinde görülebilir. Birinci grupta meteorolojik parametreler – yağış, sıcaklık, bağıl nem ve rüzgâr hızı – bulunmaktadır ve bu parametreler meteorolojik gözlem istasyonlarından toplanan verilerden oluşmaktadır. İkinci grupta da hidrolojik parametre – akım parametresi – bulunmaktadır ve bu parametre akım ölçüm istasyonlarından toplanan verilerden oluşmaktadır.

Şekil 3.2'de, bu çalışmada kullanılan verilerin toplandığı istasyonlar ve bu istasyonların Türkiye haritasındaki konumları gösterilmektedir. Kullanılan veriler 64 adet akım ölçüm istasyonu ve her bir akım istasyonu ile en yüksek ilişki gösteren meteorolojik gözlem istasyonuna aittir. İstasyonların seçimi ve ilişki analizi ile ilgili detaylı bilgi 3.3 numaralı başlıkta bulunmaktadır. Kırmızı üçgen şekilli noktalar akım ölçüm istasyonlarını ve siyah daire şekilli noktalar da meteorolojik gözlem istasyonlarının konumlarını göstermektedir.

Veri kümesinde bulunan parametrelerin tamamına ait veriler günlük gözlem verilerinden oluşmaktadır. Tüm veriler 1975 yılından başlayıp 2000 yılına kadar devam etmektedir.



Şekil 3.2. Akım ölçüm ve meteoroloji istasyonların konumu

3.3 Veri Üzerinde Yapılan İşlemler

Bu kısımda veri kümesi üzerinde gerçekleştirilmiş olan ön işlem aşaması anlatılmaktadır. Genel olarak alt başlıklara bölünecek olunursa, öncelikle veri tabanı içerisinden veri seçme işlemi gerçekleştirilmiş ve ilgilenilen istasyonlara ait veriler seçilmiştir. Ardından verilerin günlük formattan aylık hale getirilmesi işlemi gerçekleştirilmiştir. Son olarak, sürekli veri ayrıklaştırılarak ayrı veri kümesi elde edilmiştir. Aşağıda bu adımlarla ilgili detaylar verilmiştir.

3.3.1. Veri Seçme

Bu adımda veri tabanı üzerinde bulunan akım ve meteoroloji istasyonlarından hangilerinin kullanılacağı ve bu istasyonlara ait hangi parametrelerin seçileceği belirlenmiştir. Çalışmanın amacı hidrolojik parametre üzerindeki etki incelendiği için öncelikle akım ölçüm istasyonlarının seçimi yapılmıştır. Bu istasyonların seçimi aşamasında, 300'den fazla istasyon arasından istasyonlara ait verilerin tamlık, homojenlik ve uzunluğu incelenerek, tam, homojen ve yeterli uzunlukta olan 64 adet akım istasyonu seçilmiştir. Ardından meteorolojik gözlem istasyonları için de aynı şekilde tamlık, homojenlik ve uzunluk incelemesi yapılmıştır. Meteorolojik istasyonlar arasından da tam, homojen ve yeterli uzunlukta olan istasyonlar seçilmiştir.

Akım ölçüm istasyonları ile meteoroloji istasyonlarına ait veriler birleştirilerek veri kümesi elde edilecektir. Ancak bir akarsu havzasında birden fazla meteoroloji istasyonu bulunmaktadır. Bu sebeple bu aşamada bir ara işlem yaparak her bir akım istasyonu

için, aynı havzadaki hangi meteoroloji istasyonunun en uygun olduğuna karar vermek gerekmektedir. Bu işlem için de meteorolojik istasyon verileri ile akım istasyon verileri arasında korelasyon analizi yapılmıştır. Bir akım istasyonu ile en çok ilişki gösteren meteoroloji istasyonu akım istasyonu ile ilişkilendirilmiş ve istasyonlara ait veriler birleştirilerek veri kümesi oluşturulmuştur.

3.3.2. Verilerin Aylık Forma Dönüştürülmesi

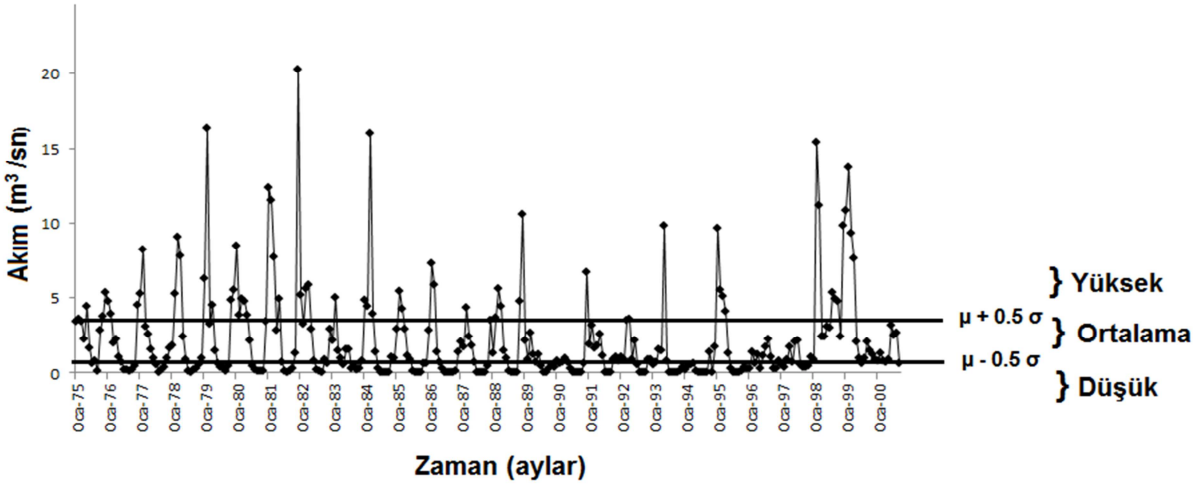
Veri kümesi istasyonlar bazında oluşturulmuştur ancak ortaya çıkan verilerin tamamı günlük gözlem verilerini içermektedir. Bu çalışmada verilerin aylık formları kullanılacak olduğundan dolayı günlük gözlem verilerinin aylık hale getirilmesi gerekmektedir. Bu amaçla, sıcaklık bağıl nem, rüzgar hızı ve akım parametrelerinin aylık ortalamaları alınarak ve yağış verisinin ise aylık toplamı alınarak veri kümesi aylık hale getirilmiştir.

Yağış verisinin aylık hale getirilmesinde toplamının alınmasının sebebi, yağış parametresinin toplam yağış şeklinde incelenmesinin, örüntülerin elde edilmesi aşamasında daha doğru sonuçlar vermesidir.

3.3.3. Ayrıklaştırma

Birliktelik analizi algoritmalarının özelliklerinden biri de ayrık veri ile çalışmalarıdır. Bu çalışmada önerilen HMP-Miner algoritması da ayrık verilerle çalışabilmektedir. Dolayısıyla seçilmiş ve aylık hale getirilmiş olan sürekli hidrolojik ve meteorolojik verilerin ayrıklaştırılması gerekmektedir. Bu kısımda, bu çalışmada yapılmış olan ayrıklaştırma işlemi anlatılmaktadır.

Verilerin doğru bir şekilde ayrıklaştırılabilmesi için istatistiksel bir yöntem kullanmak gerekmektedir. Bu çalışmada ayrıklaştırma işlemi ortalama ve standart sapmaya bağlı bir şekilde yapılmıştır. Her bir parametre ortalama (μ) ve standart sapmaya (σ) bağlı olarak 3 gruba ayrılmıştır. Parametre değeri, eğer "μ - 0.5σ" ile "μ + 0.5σ" arasında ise "Ortalama (O)" olarak isimlendirilmiştir. Parametre değeri, eğer "μ - 0.5σ"'dan düşükse "Düşük (D)" ve eğer "μ + 0.5σ"'den yüksekse "Yüksek (Y)" olarak isimlendirilmiştir. Şekil 3.3'te akım parametresine ait örnek bir ayrıklaştırma işlemi görülmektedir.



Şekil 3.3. Akım parametresine ait örnek bir ayrıklaştırma

Şekil 3.3'te görülebileceği üzere "$\mu - 0.5\sigma$" değeri çizgisi altında kalan akım değerleri düşük (D), "$\mu - 0.5\sigma$" ile "$\mu + 0.5\sigma$" çizgileri arasında kalan akım değerleri ortalama (O) ve "$\mu + 0.5\sigma$" çizgisi üzerinde kalan akım değerleri de yüksek (Y) olarak etiketlenmektedir.

Ayrıklaştırma işlemi sonucunda eldeki veri kümesi, gerçek veri değerlerinin ayrıklaştırılmış karşılıkları olan D, O ve Y etiketleri ile etiketlenmiş olmaktadır. Ayrıklaştırma işleminden sonra veri ön işlem aşamaları tamamlanmış olmaktadır ve veri kümesi artık algoritmaların çalışmasına uygun hale gelmektedir.

3.4. Deneyler

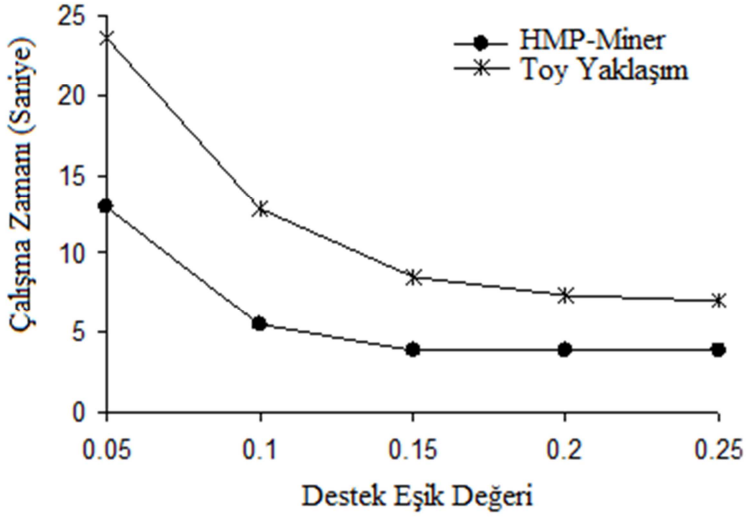
Deneyler yapılırken, her iki algoritmanın da (HMP-Miner ve Toy Yaklaşım) değişkenleri aynı olduğu için, değişkenlerden ikisi sabit tutulurken, bir değişkenin değeri değiştirilerek algoritmaların çalışma zamanları bu değişkenin değerleri için incelenmiştir. Sonuç olarak elde edilen sonuçlar değerlendirilerek, algoritmaların performansı bu sonuçlar altında incelenmiştir. Yapılan deneylerde, destek eşik değerinin etkisi nedir, istasyon eşik değerinin etkisi nedir ve istasyon sayısının etkisi nedir sorularına cevap aranmaktadır.

3.4.1. Minimum Destek Eşik Değerinin Etkisi

Bu deneyde, minimum destek eşik değerinin algoritmaların çalışma zamanlarına etkisi incelenmiştir. Deney sonucu Şekil 3.4'te gösterilmektedir. İstasyon eşik değeri 0.8 ve

güven eşik değeri de 0.1 olarak belirlenmiştir. Minimum destek eşik değeri 0.05'ten 0.25'e kadar 0.05'lik artışlarla algoritmalar çalıştırılmıştır. Algoritmaların çalışma zamanları incelendiğinde destek eşik değeri arttıkça algoritmaların çalışma zamanları azalmaktadır. Şekil 3.4'te de açıkça görülebildiği üzere HMP-Miner algoritması, toy yaklaşımdan daha az zaman almaktadır. Bunun sebebi ise, HMP-Miner algoritmasının istasyonlar arası yaygın olmayan örüntülerin mümkün olduğunca erken elenmesidir ve aday örüntülerin potansiyel olarak başarılı örüntülerden üst aday üretimi yapmasıdır.

Şekil 3.4'te de görülebileceği gibi, her iki algoritma da azalan bir eğilim göstermektedir. Bununla beraber başlangıç aşamasında bile HMP-Miner algoritması, toy yaklaşımdan yaklaşık 2 kat daha hızlı çalışmaktadır. Minimum destek eşik değeri arttıkça algoritmaların çalışma zamanları birbirine yaklaşmaktadır. Sonuç olarak HMP-Miner algoritmasının, minimum destek eşik değeri parametresi üzerinde daha performanslı olarak çalıştığı ortaya konulmuştur.



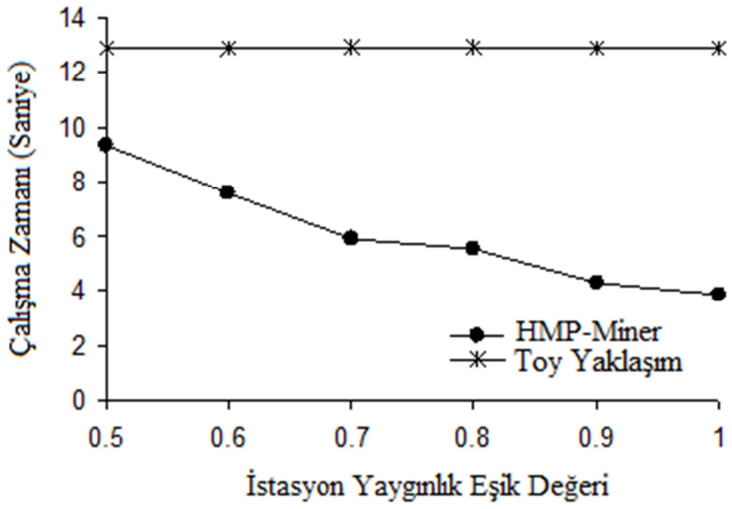
Şekil 3.4. Minimum destek eşik değerinin etkisi

3.4.2. Minimum İstasyon Eşik Değerinin Etkisi

Bu deneyde minimum istasyon eşik değerinin algoritmaların çalışma zamanlarına etkisi incelenmiştir. Deney sonucu Şekil 3.5'te gösterilmektedir. Destek eşik değeri 0.1 ve güven eşik değeri de 0.1 olarak belirlenmiştir. Minimum istasyon eşik değeri 0.5'ten 1.0'a kadar 0.1'lik artışlarla algoritmalar çalıştırılmıştır. Algoritmaların çalışma zamanları incelendiği zaman, minimum istasyon eşik değeri arttıkça HMP-Miner

algoritmasının çalışma zamanı azalmaktadır. Ancak toy yaklaşımın minimum istasyon eşik değerinden etkilenmediği ve çalışma zamanının sabit kaldığı görülmektedir. Sonuçlar incelendiğinde HMP-Miner algoritmasının minimum istasyon eşik değerine hassas olduğu görülmektedir. Ancak toy yaklaşımın çalışma zamanı minimum istasyon eşik değerine bağlı olarak değişmemektedir. Dolayısıyla toy yaklaşımın minimum istasyon eşik değerine hassas olmadığı söylenebilir. Bu durumda istenen yaklaşımı HMP-Miner algoritması göstermektedir çünkü bir girdi parametresinin algoritmanın çalışma zamanını ve performansını etkilemesi gerekmektedir.

Şekil 3.5'te de görülebileceği üzere toy yaklaşım minimum istasyon eşik değeri artmasına rağmen sabit kalmakta, HMP-Miner algoritması ise minimum istasyon eşik değerinin artışına tepki vererek çalışma zamanını azaltmaktadır. Bunun sebebi, toy yaklaşımın istasyon yaygınlık değerlendirmesini son işlem adımı olarak yapması ve gereksiz örüntülerin üretimini engelleyememesidir. Buna rağmen HMP-Miner algoritması, istasyon yaygınlık değerlendirmesini örüntü üretimi aşaması ile paralel bir şekilde ilerlettiği için minimum istasyon eşik değerinin değişimine daha iyi adapte olabilmektedir. Sonuç olarak, HMP-Miner algoritması, toy yaklaşımdan daha az örüntü üreterek daha az örüntü ile ilgilendiği için toy yaklaşımdan daha iyi sonuç vermektedir.

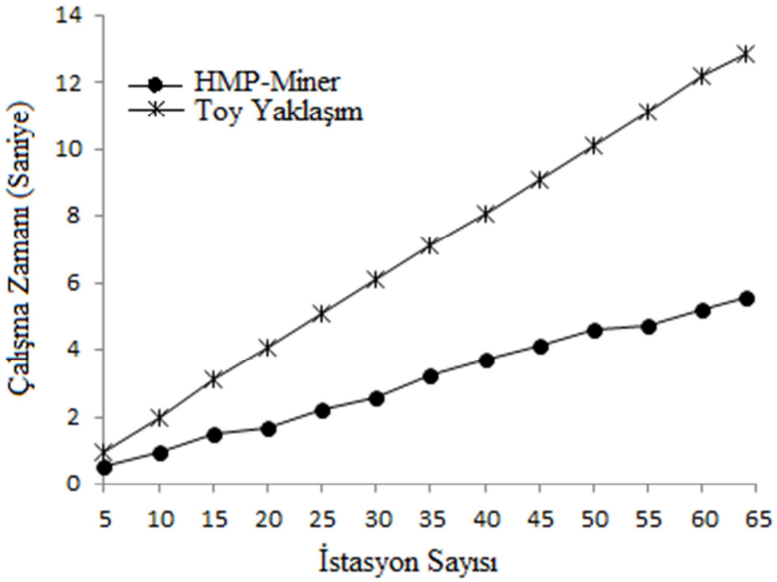


Şekil 3.5. Minimum istasyon eşik değerinin etkisi

3.4.3. İstasyon Sayısının Etkisi

Son olarak istasyon sayısının algoritmaların performansı üzerinde etkisi olup olmadığı ve varsa nasıl bir etkisi olduğu incelenmiştir. Deney sonucu Şekil 3.6'da gösterilmektedir. Destek eşik değeri 0.1, istasyon eşik değeri 0.8 ve güven eşik değeri de 0.1 olarak belirlenmiştir. İstasyon sayısı 5'ten 64'e kadar 5 istasyonluk artışlarla algoritmalar çalıştırılmıştır. Algoritmaların çalışma zamanları incelendiği zaman, her iki algoritmanın da istasyon sayısı arttıkça çalışma zamanları artmaktadır. Ancak artış yönleri incelendiği zaman HMP-Miner algoritması daha yavaş bir artış göstermekle birlikte toy yaklaşım daha hızlı bir artış göstermektedir. Bu durunda da HMP-Miner algoritması istasyon sayısının artışı konusunda da toy yaklaşıma göre daha performanslı çalıştığını ortaya koymaktadır.

Şekil 3.6'da algoritmaların istasyon sayısındaki değişime göre çalışma zamanları görülmektedir. HMP-Miner algoritması daha performanslı çalışmasıyla dikkat çekmektedir. Bu durumda HMP-Miner toy yaklaşıma göre veri kümesinin artışına daha iyi yanıt vermekte ve daha kısa sürede sonuç üretebilmektedir.



Şekil 3.6. İstasyon sayısının etkisi

3.5. Deney Sonuçları

Bu kısımda hidro-meteorolojik birliktelik analizi sonucu Türkiye hidrolojik ve meteorolojik verileri kullanılarak elde edilen sonuçlar verilmektedir. Bu tez

çalışmasının amaçlarından biri algoritmik bir yapı ortaya koymak olduğu kadar, elde edilen sonuçların da anlamlı olması gerekmektedir. Bu sebeple geliştirilen HMP-Miner algoritması ve değerlendirme ölçütünün kullanılmasıyla elde edilen sonuçların yeterli ve anlamlı olup olmadığı test edilmiştir.

Algoritma sonucunun anlamlı olması, değerlendirme ölçütlerinin belli değerleri için daha önceki yapılmış olan çalışmalarla tutarlı sonuçlar elde edilmesi demektir. Bu amaçla HMP-Miner algoritması, Türkiye bazında hidro-meteorolojik kuralları ortaya çıkarmak için kullanılmıştır (Tablo 3.1 ve Şekil 3.7). Bu deneyde 64 akım ölçüm istasyonu ve bu istasyonlarla ilişkili meteoroloji istasyonları kullanılmıştır. Minimum destek eşik değeri olarak 0.15, istasyon yaygınlık eşik değeri olarak 0.5 ve minimum güven eşik değeri olarak 0.5 değerleri kullanılmıştır. Algoritmanın hedefi de düşük (D), ortalama (O) ve yüksek (Y) akım değerleridir.

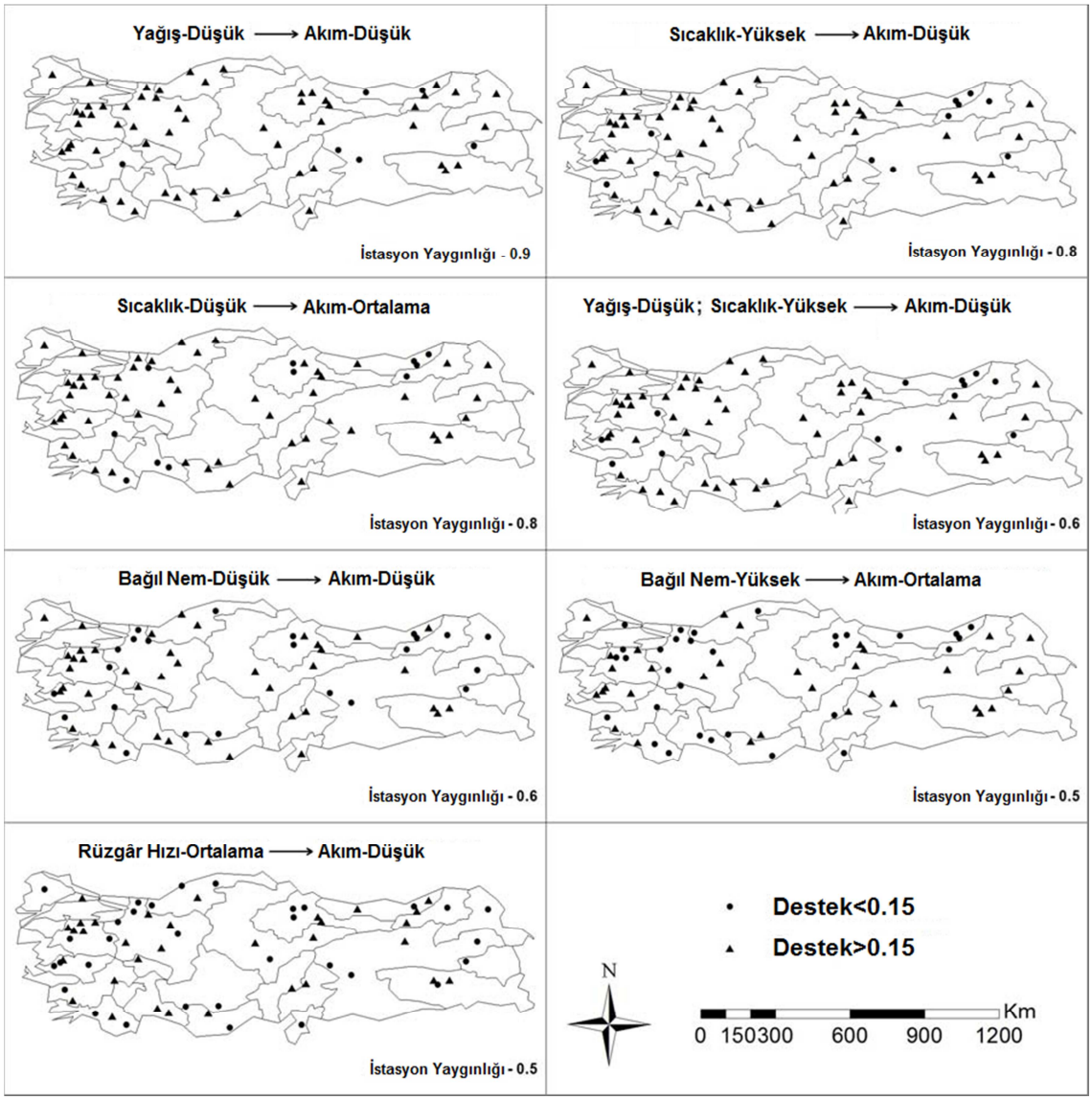
Bu kısıtlar altında HMP-Miner algoritması 7 adet hidro-meteorolojik kural çıkarmıştır Elde edilen kurallar Tablo 3.1'de gösterilmiştir. Elde edilen sonuçlar akım ile yağış ve sıcaklık arasında güçlü bir ilişki olduğunu ortaya çıkarmıştır. Özellikle düşük akım değerleri için bu ilişki geçerli olmaktadır. İstasyonların büyük çoğunluğunda yağış düşük ve sıcaklık yüksek olduğu durumlarda akımın düşük olduğu gözlenmiştir.*{Yağış-Düşük→Akım-Düşük}* ve *{Sıcaklık-Yüksek→Akım-Düşük}* kuralları, Türkiye'nin kuzeydoğusundaki küçük bir bölge hariç tutulursa, mekânsal olarak yaygın olduğu görülmüştür. Bu kurallar için istasyon yaygınlıkları, sırasıyla 0.9 ve 0.8'dir. Bu demektir ki bu kurallar, tüm istasyonların en az %80'inde bulunmaktadır. Aynı zamanda rüzgar hızı ve bağıl nem parametreleriyle ilgili ilişkiler de verilmiştir. Rüzgar hızı yüksek ve bağıl nem düşük olduğu durumlarda akım değeri düşük olarak gözlenmiştir. Akım değerinin ortalama olması ise yüksek bağıl nem ve düşük sıcaklık değerleriyle ilişkili olduğu gözlenmiştir.

Tablo 3.1'de *Yağış* yerine *Y*, *Sıcaklık* yerine *S*, *Rüzgâr Hızı* yerine *RH*, *Bağıl Nem* yerine *N* ve *Akım* yerine *A* kısaltması kullanılmıştır.

Tablo 3.1. HMP-Miner algoritmasının bulduğu kurallar (*min_destek*=0.15, *min_istasyon*=0.5, *min_güven*=0.5) (D: Düşük, O: Ortalama, Y: Yüksek)

Kural	Destek Yaygınlığı (64 istasyon için)	İstasyon Yaygınlığı (64 istasyon için)
{Y-D→A-D}	>=0.20	0.9
{S-Y→A-D}	>=0.20	0.8
{S-D→A-O}	>=0.15	0.6
{Y-D, S-Y→A-D}	>=0.15	0.6
{N-D→A-D}	>=0.15	0.6
{N-Y→A-O}	>=0.15	0.5
{RH-O→A-D}	>=0.15	0.5

Bu çalışmada bulunan sonuçların doğru veya ilginç olup olmadıkları da ayrıca değerlendirilmiştir. Bu amaçla, bu çalışmada elde edilen sonuçlarla, aynı başlıkta yapılan çalışmaların sonuçları karşılaştırılmıştır. Değerlendirme sonucuna göre, akım ve meteorolojik parametreler arasındaki ilişki daha önceki pek çok çalışma ile uyumludur [39, 42]. Sonuçlarımızla benzer olarak, daha önceki çalışmalar ayrıca yağış ve sıcaklığın akımı en çok etkileyen iki parametre olduğu görülmüştür. Daha yeni bir çalışmada, Dadaşer-Çelik ve Cengiz [38], Türkiye'de akım ile çeşitli meteorolojik parametre arasındaki ilişkiler incelenmiştir ve tüm Türkiye üzerinde akım ile sıcaklık arasında güçlü ilişki olduğu gösterilmiştir.



Şekil 3.7. HMP-Miner algoritmasının bulduğu kurallar (*min_destek*=0.15, *min_istasyon*=0.5, *min_güven*=0.5)

4. BÖLÜM

SONUÇ, TARTIŞMA VE ÖNERİLER

Bu tez çalışmasında mekân-zamansal birliktelik örüntülerini ortaya koyacak bir algoritma önerilmiş ve bu algoritma toy yaklaşımla karşılaştırılmıştır. Önerilen bu algoritmanın hidrolojik ve meteoroloji veri kümesi üzerindeki çalışma performansı gözlenmiş ve etkin bir şekilde çalıştığı ortaya konulmuştur. Ayrıca bu tez çalışması kapsamında hidro-meteorolojik örüntülerin ortaya çıkarılması için bir değerlendirme ölçütü (hidro-meteorolojik yaygınlık indeksi) önerilmiştir. Deneysel çalışmalarda da görüldüğü üzere HMP-Miner algoritması başarılı bir şekilde hidro-meteorolojik örüntüleri keşfedebilmekte ve klasik yönteme (toy yaklaşım) göre avantajlar sağlamaktadır.

4.1 Sonuç

Bu çalışmada, hidro-meteorolojik kurallar elde etmek için kullanılan HMP-Miner algoritması ve toy yaklaşım performans açısından karşılaştırılmıştır. Yapılan 3 farklı deneyde de HMP-Miner algoritmasının toy yaklaşıma göre çok daha iyi performans gösterdiği gözlenmiştir. İlk deneyde, minimum destek eşik değeri artırılarak, bu artışın ne tür bir etkisi olacağı gözlenmiştir. Bu etki beklendiği şekilde algoritmaların çalışma zamanlarının azalması ile sonuçlanmıştır. Çünkü minimum destek eşik değeri artırıldığı zaman daha az *k* boyutlu örüntü, eşik değerini sağlamakta ve dolayısıyla daha az örüntü ile çalışılmaktadır. Bu durum da algoritmaların daha hızlı sonuç vermesini ve çalışma zamanlarının azalmasını sağlamaktadır. Bu deneyde HMP-Miner algoritmasındaki azalma toy yaklaşımdaki azalmadan daha fazla olmuş ve neredeyse iki katı kadarlık bir performans farkı ortaya konulmuştur.

İkinci deneyde, minimum istasyon eşik değeri artırılarak, bu değişkenin algoritmaları nasıl etkileyeceği gözlenmiştir. Bu değişkenin artışı HMP-Miner algoritmasının performansına, minimum destek eşik değerinde olduğu gibi, çalışma zamanının azalması şeklinde yansımıştır. Ancak toy yaklaşım istasyon yaygınlık şartını son işlem olarak uyguladığı için toy yaklaşımın çalışma zamanı sabit kalmakta ve bu değişkenin değişiminden etkilenmemektedir. Bu deneyde de HMP-Miner, istasyon yaygınlık değerini mümkün olduğu kadar erken işleme koyduğu için toy yaklaşıma göre daha performanslı olduğu, bu deney yardımıyla da ortaya konulmuştur.

Üçüncü deneyde, istasyon sayısının algoritmaların çalışmasındaki etkisi incelenmiştir. Bu değerin artışına toy yaklaşım lineere yakın bir grafik ile tepki vermiş ve daha dik bir artış göstermiştir. HMP-Miner algoritmasında da istasyon sayısının artışı çalışma zamanını artırmakla birlikte daha yatık bir eğimle bu sonuç ortaya konulmuştur. Bu durumun sebebi de yine HMP-Miner algoritmasının istasyon yaygınlığını daha erken işleme koyması sebebiyle daha az örüntüyle çalışması yatmaktadır.

4.2 Tartışma

Mekân-zamansal birlikteliklerin ortaya konulması, gün geçtikçe daha çok önem kazanmaktadır. Çünkü bir örüntünün içerisinde bulunduğu mekân ve zamandan bağımsız olarak doğru bir şekilde değerlendirilmesi mümkün olmamaktadır.

Bu tez çalışmasında kullanılan tekniklerin, akım ve meteorolojik parametreler arasındaki ilişkileri ortaya koymakta kullanılan korelasyon veya regresyon analizi gibi tekniklere göre çeşitli avantajlar sağladığı gözlenmiştir. Regresyon analizi veri için belli bir hata payıyla bir model ortaya koymaktadır. Korelasyon analizi iki parametre arasında birlikteliklerin derece ve yönünü ortaya çıkarmaktadır. Önerilen HMP-Miner algoritması, bu yaklaşımlara göre daha iyi ve performanslı çalışmaktadır. Çünkü HMP-Miner algoritması ile akım ve çeşitli meteorolojik parametreleri arasındaki ilişkiler ortaya konulmuş, ayrıca bu parametreler arasındaki farklı kombinasyonları da değerlendirmektedir. Üretilen sonuçların kalitesi açısından incelendiğinde HMP-Miner algoritmasının daha iyi olduğunu ortaya koymaktadır. Aynı zamanda HMP-Miner algoritması çok sayıda parametreye sahip geniş veri tabanları üzerinde de başarılı bir şekilde çalışabilmektedir. Algoritmanın veri kümesine entegrasyonu ise çok daha kolay bir şekilde gerçekleştirilebilmektedir.

Bu avantajları sebebiyle HMP-Miner algoritması her açıdan hidro-meteorolojik örüntülerin keşfedilmesi için uygun ve tercih edilebilir bir algoritma olmaktadır. Ayrıca önerilen değerlendirme ölçütü de üretilen sonuçların kalitesinin artmasını sağlamıştır.

4.3 Gelecekteki Çalışmalar

Gelecekteki çalışmalarda, önerilen HMP-Miner algoritmasını farklı zaman ve mekân ölçeklerinde çalışacak şekilde geliştirilmesine çalışılacaktır.

KAYNAKÇA

1. Han, J., Kamber, M., and Pei, J., 2011. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers Inc., 3rd Edition, 744 pp.
2. Tan, P.N., Steinbach, M., and Kumar, V., April 2005. Introduction to Data Mining. Addison-Wesley, 769 pp.
3. Geng, L., and Hamilton, H.J., 2006. Interestingness measures for data mining: A survey. **ACM Computing Surveys, 38** (3): Makale No: 9
4. Agrawal, R., Imielinski, T., and Swami, A., 1993. Mining association rules between sets of items in large databases, pp. 207 – 216. *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, May 1993, Washington DC, USA.*
5. Kob, H.C., and Tan, G., 2005. Data mining application in healthcare. **Journal of Healthcare Information Management, 19**: 64-71.
6. Chen, H., Fuller, S.S., Friedman, C., and Hersh, W., 2005. Medical informatics: Knowledge management and data mining in biomedicine. Springer-Verlag Inc., 647 pp.
7. Giudici, P., and Figini, S., 2009. Applied data mining for business and industry. John Wiley & Sons, 2nd Edition., 376 pp.
8. Olson, D., and Shi, Y., 2005. Introduction to business data mining. McGraw-Hill/Irwin. 273 pp.
9. Koperski, K., and Han, J., 1995. Discovery of spatial association rules in geographic information databases, pp. 47-66. *Proceedings of the 4th International Symposium on Advances in Spatial Databases, 1995, London, UK.*
10. Mamoulis, N., Cao, H., Kollios, G., Hadjieleftheriou, M., Tao, Y., and Cheung, D.W., 2004. Mining, indexing, and querying historical spatiotemporal data, pp. 236-245. *Proceedings of the 10th ACM SIGKDD international conference on Knowledge discovery and data mining, 2004, Seattle, WA, USA.*
11. Tsoukatos, I., and Gunopulos, D., 2001. Efficient mining of spatiotemporal patterns, pp. 425-442. *Proceedings of the 7th International Symposium on Advances in Spatial and Temporal Databases, 2001, London, UK.*

12. Hadjieleftheriou, M., Kollios, G., Gunopulos, D., and Tsotras, V., 2003. On-line discovery of dense areas in spatio-temporal databases, pp. 306-324. *In*: Advances in Spatial and Temporal Databases (Hadzilacos, T., Manolopoulos, Y., Roddick, J., and Theodoridis, Y.. Springer, Heidelberg Berlin.
13. Shekhar, S., and Chawla, S., 2003. Spatial Databases: A Tour. Prentice-Hall. 262 pp.
14. Hsu, W., Lee, M.L., and Wang, J., 2008. Temporal and Spatio-Temporal Data Mining. IGI Global. 292 pp.
15. Celik, M., Shashi, S., Rogers, J.P., and Shine, J.A., 2008. Mixed-drove spatiotemporal co-occurrence pattern mining. **IEEE Transactions on Knowledge and Data Engineering, , 20** (10): 1322-1335.
16. Celik, M., Shekhar, S., Rogers, J.P., and Shine, J.A., 2006. Sustained emerging spatio-temporal co-occurrence pattern mining: A summary of results, pp. 106-115. *Proceedings of the 18th IEEE International Conference on Tools with Artificial Intelligence, 2006, Washington DC, USA.*
17. Celik, M., Kang, J.M., and Shekhar, S., 2007. Zonal co-location pattern discovery with dynamic parameters, pp. 433-438. *Proceedings of the 7th IEEE International Conference on Data Mining, 2007, Omaha, Nebraska, USA.*
18. Mamoulis, N., 2009. Spatio-temporal data mining, pp. 2725-2730. *In:* Encyclopedia of Database Systems (Liu, L., and ÖZsu, M.T.). Springer US.
19. Boriah, S., Kumar, V., Steinbach, M., Potter, C., and Klooster, S., 2008. Land cover change detection, pp. 857-865. *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2008, New York NY, USA.*
20. Mithal, V., Garg, A., Boriah, S., Steinbach, M., Kumar, V., Potter, C., Klooster, S., and Castilla-Rubio, J.C., 2011. Monitoring global forest cover using data mining. **ACM Transactions on Intelligent Systems and Technology , 2** (4): 1-24.
21. Boriah, S., Mithal, V., Garg, A., Steinbach, M., Kumar, V., Potter, C., Klooster, S., and Castilla-Rubio, J.C., 2010. Automated detection of forest cover changes, pp. 44-47. *2010 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), 25-30 July 2010, Honolulu HI, USA.*

22. Potter, C., Kumar, V., Klooster, S., and Nemani, R., 2007. Recent history of trends in vegetation greenness and large-scale ecosystem disturbances in Eurasia. **Tellus B, 59** (2): 260-272.

23. Potter, C., Klooster, S., Steinbach, M., Tan, P.N., Kumar, V., Shekhar, S., and Reis de Carvalho, C., 2004. Understanding global teleconnections of climate to regional model estimates of amazon ecosystem carbon fluxes. **Global Change Biology, 10** (5): 693-703.

24. Potter, C., Klooster, S., Steinbach, M., Tan, P., Kumar, V., Shekhar, S., Nemani, R., and Myneni, R., 2003. Global teleconnections of climate to terrestrial carbon flux. **Journal of Geophysical Research: Atmospheres, 108** (D17): 45-56.

25. Steinbach, M., Tan, P.-N., Kumar, V., Klooster, S., and Potter, C., 2003. Discovery of climate indices using clustering, pp. 446-455. *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2003, Washington, DC, USA.*

26. Steinbach, M., Tan, P.-N., Kumar, V., Klooster, S., and Potter, C., 2002. Data mining for the discovery of ocean climate indices, *2nd SIAM International Conference on Data Mining, 11-13 April 2002, Arlington, VA, USA.*

27. Steinbach, M., Tan, P.-N., Kumar, V., Klooster, S., and Potter, C., 2002. Temporal data mining for the discovery and analysis of ocean climate indices. *Proceedings of KDD Temporal Data Mining Workshop, 2002.*

28. Romani, L.A.S., Ávila, A.M.H.d., Jr., J.Z., Jr., C.T., and Traina, A.J.M., 2009. Mining climate and remote sensing time series to discover the most relevant climate patterns, pp. 181-195. *XXIV Simpósio Brasileiro de Banco de Dados, 2009, Outubro, Fortaleza, Ceará, Brasil.*

29. Romani, L.A.S., de Avila, A.M.H., Chino, D.Y.T., Zullo, J., Chbeir, R., and Traina, A.J.M., 2013. A new time series mining approach applied to multitemporal remote sensing imagery. **IEEE Transactions on Geoscience and Remote Sensing, , 51** (1): 140-150.

30. Romani, L.A.S., Ávila, A.M.H., Jr., J.Z., Jr., C.T., and Traina, A.J.M., 2010. Mining relevant and extreme patterns on climate time series with clips miner. **Journal of Information and Data Management, 1** (2): 245-260.

31. Tadesse, T., Wilhite, D.A., and Hayes, M.J., 2004. Discovering associations between climatic and oceanic parameters to monitor drought in nebraska using data-mining techniques. **Journal of Climate, 18**: 1541-1549.
32. Tadesse, T., Wilhite, D., Harms, S., Hayes, M., and Goddard, S., 2004. Drought monitoring using data mining techniques: A case study for nebraska, USA. **Natural Hazards, 33** (1): 137-159.
33. Kumar, D.N., Ish, M., and Dhanya, C.T., 2009. Data mining and it's applications for modelling rainfall extremes. **Journal of Hydraulic Engineering, 15** (1): 25-51.
34. Dhanya, C.T., and Nagesh Kumar, D., 2009. Data mining for evolution of association rules for droughts and floods in india using climate inputs. **Journal of Geophysical Research: Atmospheres, 114** (D2): D02102.
35. Kahya, E., and Karabork, M.C., 2001. The analysis of el nino and la nina signals in streamflows of Turkey. **International Journal of Climatology, 21**: 1231-1250.
36. Lin, F., Jin, X., Hu, C., Gao, X., Xie, K., and Lei, X., 2007. Discovery of teleconnections using data mining technologies in global climate datasets. **Data Science Journal, 6**: 749-755.
37. Celik, F.D., Celik, M., and Dokuz, A.S., 2012. Associations between stream flow and climatic variables at kızılırmak river basin in Turkey. **Global Nest Journal, 14** (3): 354-361.
38. Celik, F.D., and Cengiz, E., 2012. Correlations of stream flow and climatic variables in Turkey. *BALWOIS 2012 - 5.International Conference on Water, Climate, & Environment, Ohris, Macedonia 2012.*
39. Changnon, S.A., and Kunkel, K.E., 1995. Climate-related fluctuations in midwestern floods during 1921-1985. **Journal of Water Resources Planning and Management, 121**: 326-334.
40. Kletti, L., and Stefan, H., 1997. Correlations of climate and streamflow in three minnesota streams. **Climatic Change, 37** (4): 575-600.
41. Groisman, P.Y., Knight, R.W., and Karl, T.R., 2001. Heavy precipitation and high streamflow in the contiquous united states: Trends in the twentieth century. **Bulletin of the American Meteorological Society, 82**: 219-246.

42. Cayan, D.R., Riddle, L.G., and Aguado, E., 1993. The influence of precipitation and temperature on seasonal streamflow in california. **Water Resources Research, 29** (4): 1127-1140.

ÖZGEÇMİŞ

Adı Soyadı: Ahmet Şakir Dokuz

Uyruğu: Türkiye Cumhuriyeti (T.C.)

Doğum Tarihi ve Yeri: 15.08.1988 - Yozgat

Medeni Durumu: Evli

Tel: 507 533 0256

E-posta: adokuz@nigde.edu.tr, dokuz.a.sakir@gmail.com

Adres: Niğde Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Merkez/NİĞDE

I. Eğitim Durumu:

2006 – 2010: Erciyes Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü (**Derece: 2.96 / 4.00**)

2002 – 2006: Nuh Mehmet Baldöktü Anadolu Lisesi / Kayseri (**Derece: 4.96 / 5.00**)

II. İş Tecrübeleri

2011 – Halen: Niğde Üniversitesi Bilgisayar Mühendisliği Bölümü Araştırma Görevlisi

2010 – Halen: TÜBİTAK Proje Asistanı (Veri Madenciliği)

III. Bilimsel Çalışmalar ve Yayınlar

SCI, SCI-Expanded, SSCI veya AHCI Kapsamındaki Dergi Yayınları:

1. M. Celik, F. Dadaser Celik and **A. S. Dokuz**, Discovery of Hydro-meteorological Patterns, *Turkısh Journal Of Electrıcal Engıneerıng & Computer Scıences* (Kabul edildi, yayınlanmadı)
2. F.Dadaser-Celik, M. Celik, and **A. S. Dokuz**, Associations Between Stream Flow and Climatic Variables at Kizilirmak River Basin in Turkey, *Global NEST Journal*, 14 (3), 354-361, 2012

Konferans Yayınları:

1. M. Çelik, F. Dadaşer-Çelik, A. Ü. Kömüşçü, M. Arslan ve **A. Ş. Dokuz**, "Yağış ve İklim İndeksleri Arasındaki Sıralı Birliktelik Örüntülerinin Tespit Edilmesi", *2012 ASYU - Akıllı Sistemlerde Yenilikler ve Uygulamaları*, pp. 134-138, 3-4 Temmuz 2012 Trabzon Türkiye
2. F. Dadaser-Celik, M.Celik and **A. S. Dokuz**, "Associations Between Stream Flow and Climatic Variables at Kizilirmak River Basin in Turkey", *in 2011 CEST - Proceedings of the 12th International Conference on Environmental Science and Technology*, pp. 367-374, 8-10 September 2011, Rhodes Greece
3. M. Celik, F. Dadaser-Celik and **A. S. Dokuz**, "Anomaly Detection in Temperature Data Using DBSCAN Algorithm", in *2011 INISTA - International Symposium on Innovations in Intelligent Systems and Applications*, pp. 91-95, 15-18 June 2011, Istanbul Turkey

IV. Katıldığı Konferanslar

1. ASYU 2012 – Trabzon Konferansı – Makale Sunumu

2. INISTA 2011 – İstanbul Konferansı – Makale Sunumu

V. Yabancı Dil

İngilizce: İyi, Almanca: Orta