Büyük Veri Analizinde Veri Madenciliği Araçlarının Performansı

Gamze Özçelik
Dokuz Eylül Üniversitesi
Bilgisayar Mühendisliği
İzmir/Türkiye
gamze.ozcelik@ceng.deu.edu.tr

Prof. Alp Kut
Dokuz Eylül Üniversitesi
Bilgisayar Mühendisliği
İzmir/Türkiye
alp.kut@ceng.deu.edu.tr

ÖZET

Günümüzde veri artışına bağlı olarak müşteri ilişki yönetimi, envanter yönetimi veya promosyonların belirlenmesi zorlaşmıştır. Bununla birlikte her alanda olduğu gibi giyim sektöründe de karar destek sistemlerine olan ihtiyaç artmıştır. Çalışmamızda Spark ve Weka araçları kullanılarak giyim sektörüne ait veri seti analiz edilmiş, sonuçlar değerlendirilerek hata oranı en az olacak şekilde karar destek sistemi oluşturulmuştur. Geliştirilen karar destek sistemi kapsamında mağazanın müşteri profilleri ve birlikte satılan ürünler belirlenmiştir. Müşteri profilleri belirlenirken K-Means algoritması, birlikte satılan ürünlerin belirlenmesinde FP-Growth algoritması kullanılmıştır. Fp-Growth ve K-Means algoritmaları aynı veri seti üzerinde her iki veri madenciliği aracında da çalıştırılmış ve üretilen sonuçlar karşılaştırılmıştır. Üretilen sonuçlar yorumlanarak mağazanın yararlanabileceği bilgi setleri elde edilmiştir. Aynı zamanda yapılan çalışma, Weka ve Spark'ın çalışma performanslarının net bir şekilde incelenmesini mümkün kılmıştır.

Anahtar Kelimeler

Veri Madenciliği; Apache Spark; Weka; Fp-Growth; K-Means

ABSTRACT

Today, depending on the data growth, it has become difficult to determine customer relationship management, inventory management or promotions. However, there is a growing need for decision support systems in the clothing sector as well as in all areas. In our study, the data set of the clothing sector was analyzed by using Spark and Weka tools and a decision support system was established to evaluate the results and to minimize the error rate. Within the scope of the decision support system developed, the customer profiles of the store and the products sold together were determined. The K-Means algorithm was used to determine the customer profiles while the FP-Growth algorithm was used to identify the products sold together. The Fp-Growth and K-Means algorithms were run on both data mining tools on the same dataset and the results produced were compared. The produced results are interpreted and information sets are obtained which can be used by the store. At the same time, it was possible to compare clearly the performances of Weka and Spark..

Keywords

Data Mining; Apache Spark; Weka; Fp-Growth; K-Means

GİRİŞ

Günümüzde bilişim sistemlerinin hayatın hemen hemen her alanında aktif bir rol oynuyor olması ile birlikte veri boyutlarında ciddi artışlar meydana gelmiştir. Veri boyutundaki bu büyük artışla birlikte veri madenciliği de doğal olarak gelişim göstermiş ve hala gelişmeye devam etmektedir. Google, Facebook, Twitter, LinkedIn gibi popüler ticari şirketlerin ve kamusal güvenlik ile ilgili kurumların ciddi yatırımlar yapmalarıyla bu alana ilgi daha da artmıştır. [1], [2], [3]

Veri madenciliği, büyük veri sistemlerinde gizli örüntülerin elde edilebilmesi amacıyla kullanılan istatistiksel ve matematiksel yöntemler bütünüdür. Bu yöntemler karar verme sürecinde oldukça etkili rol oynamaktadırlar. [1], [2], [3]

Bu çalışmanın ana hedefi veri madenciliği aşamasında kullanılan iki farklı aracın farklı algoritmalar üzerinde performans ve doğruluk analizini yapmaktır. Elde edilen analiz sonuçları doğrultusunda giyim sektörüne yönelik karar destek sistemi elde edilecektir.

Çalışma sırasında kullanılan veri seti, giyim sektörüne ait alışveriş hareketlerini kapsamaktadır. Müşterilerin satın aldıkları ürünler baz alınarak, müşteri verileri üzerinde kümeleme ve birliktelik analizi yapılmıştır. Birliktelik analizi yapılarak mağazada birlikte satılan ürünlerin belirlenmesi hedeflenirken kümeleme operasyonu ile müşteri profillerini elde etmek temel hedef haline gelmiştir.

Çalışmada Spark ve Weka araçlarının ürettiği sonuçlar değerlendirilmiştir. Değerlendirme sonucunda, mağazaya yönelik ürün öneri sistemi ve müşteri profilleri sunulmaktadır.

İLGİLİ ÇALIŞMALAR

Büyük veri analizinde;

- Spark'ın kullanıldığı,
- Farklı veri madenciliği araçlarının performanslarının karşılaştırıldığı,
- Aynı veri madenciliği aracı üzerinde farklı algoritmaların performanslarının karşılaştırıldığı

çok sayıda çalışma olmasına rağmen Spark'ın diğer veri madenciliği araçlarından daha iyi performans ile çalıştığını, sonuçlarla sergileyen bir çalışma bulunamamıştır. Hazırlanan çalışmanın sonuçları, sadece veri madenciliği araçlarının performanslarını karşılaştırmak için değil, mağaza öneri sisteminin en iyi şekilde geliştirilmesini sağlamaktadır.

İncelenen makalelerden birinde, hamile kadınların düşük yapma durumlarını önceden belirleyebilmek ve mobil cihaz üzerinden bilgilendirme yapmak hedeflenmiştir. Kullanılan veri seti mobil cihazlar üzerinden toplanmış, Spark aracı üzerinde K-Means algoritması çalıştırılarak hamile kadınlar üç ayrı kümeye ayrılmıştır. Düşük yapma riski taşıyan kadınlar kategorisinde yer alan bayanların önceden bilgilendirilmeleri sağlanmıştır.[4]

İncelenen bir diğer makalede birliktelik analizi algoritmalarından Eclat algoritması Spark üzerinde dört ayrı veri seti üzerinde çalıştırılmış ve farklı minimum destek sürelerinde hız değerlendirmesi yapılmıştır.[5]

İncelenen bir diğer çalışmada ise iris veri seti kullanılarak RapidMiner, Orange ve Weka araçları üzerinde Naive Bayes algoritması çalıştırılmıştır. Üç farklı veri madenciliği aracının ürettiği sonuçların doğruluk oranları karşılaştırılmıştır. Çalışma sonucunda Weka aracının ürettiği sonuçların hata oranının en düşük olduğu tespit edilmiştir.[6]

WEKA / SPARK KULLANILARAK BÜYÜK VERİ ANALİZİ

Veri analizi, büyük çaplı veriler arasından bilgiye ulaşma işlemidir. Ya da bir anlamda büyük veri yığınları içerisinden gelecek ile ilgili tahminde bulunabilmemizi sağlayabilecek bağıntıların bilgisayar programı aracığıyla bulunmasıdır.[3] Veri analizinde kullanılan bu programlar ürettikleri sonuçlara göre performans farklılıkları göstermektedir. Çalışmamızda aynı veri seti üzerinde hedefimize uygun olan algoritmalar çalıştırılmıştır. Elde edilen en ideal sonuç mağazaya ait ürün önerilerini ve müşteri etiketlerini belirlemiştir.

1. Çalışma Ortamı

Veri analiz sürecinde sürecinde bir çok araç kullanılmıştır.

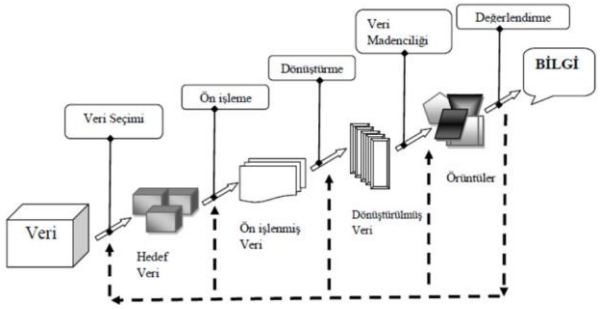
- Veri Temizleme: Visual Studio 2017, Mssql Server
- Veri Madenciliği: Apache Spark, Weka
- Değerlendirme: Excel

2. Veri seti

Kullanmış olduğumuz veri seti giyim sektörüne ait ERP sisteminden temin edilmiştir. Geliştirilecek karar destek sistemi çerçevesinde kullanılabilecek müşteri, satış ve ürün bilgilerini belirlenmesi ve 01-04-2016 - 09-02-2017 tarih aralığına ait verilerin sistemimize aktarımı sağlanmıştır. Toplamda 4.275.910 satış hareketi, 3.119.918 müşteri ve 462.856 ürün üzerinde çalışma gerçekleştirilmiştir. Verilerin temizlenmesi aşamasında bazı müşteri ve satış hareketleri algoritmanın çalışma mantığı göz önünde bulundurularak veri madenciliği sürecine dahil edilmemiştir.

3. Çalışma Adımları

Veri madenciliği sürecine ait adımları aşağıdaki resimde görmek mümkündür. Kulladığımız veri seti üzerindeki bilgi keşfi süreci, şekilde yer alan başlıklar altında anlatılmıştır.



Şekil 1 Veri Madenciliği Süreci

3.1 Veri Hazırlama

Veri temizleme operasyonu; eksik, gürültülü ve tutarsız olan verileri iyileştirmeyi amaçlamaktadır.[1] Veri madenciliği algoritmalarının çalıştırılabilmesi için veri belli aşamalardan geçirilerek analize hazır hale getirilmektedir. Verilerin temizlenmesi operasyonu aşağıda belirtilen adımları içermektedir. [1], [2], [3]

-Veri seçimi: Yapılacak analizde ihtiyaç duyulacak verilerin belirlenmesi

-Veri ön işleme: Eksik verilerin tamamlanması, aykırı ve gürültülü verilerin ayıklanması

-Veri dönüştürme: Veri formatlarının değiştirilmesi, kesikleştirilmesi ve normalize edilmesi

Yapmış olduğumuz çalışmada ilk adım olarak, karar destek sisteminde kullanılabilecek müşteri, ürün ve satış bilgileri belirlenmiştir. 01/04/2016 – 09/02/2017 tarih aralığına ait belirlenen alanların verileri sistemimize aktarılmıştır.

Veri temizleme aşamasında belirlenen zaman aralığında satış hareketi bulunmayan müşteriler analize dahil edilmemiştir. Müşteri verileri incelendiğinde cinsiyet, yaş, medeni durumu, doğum yeri gibi alanlarda hatalı

bilgiler olduğu ve boş alanların çok sayıda olduğu tespit edilmiştir. Verilerin temizlenerek analize dahil edilmesi doğruluk oranını olumsuz etkileyeceğinden müşteri kümelemesinin sadece aldığı ürünler bazında yapılmasının doğru olacağı düşünülmüştür. Bu kapsamda müşteriler mağazadan aldıkları ürün hiyerarşisi kullanılarak küme etiketi atanmıştır. Kümeleme işleminde 90 ürün kategorisi(Ceket, pantolon vs.) kullanılmıştır.

Birliktelik analizi operasyonu sadece satış hareketlerini kapsamaktadır. Kullanılan satış hareketleri verisinde gürültülü alanların az olmasından dolayı karar destek sürecine satış verilerinin tamamı dahil edilmiştir. Birliktelik analizinde ürün kodu bazında analiz yapıldığında eşik değerini geçen birlikteliğin az olacağı düşünülerek ürün ifadesi "Cinsiyet-Renk-Ürün Kategorisi" olarak ifade edilmiştir. 2133 ürün tarzı içerisinden birliktelik analizi yapılmıştır.

Veri hazırlama aşamasının öncesi ve sonrasına ait veri setlerinin özellikleri tabloda yer almaktadır.

Tablo 1 Veri Seti Özellikleri

	Veri seti	**Temizlenmiş Veri seti**
Müşteri Sayısı	3.119.918	718.903
Satış Hareketi Sayısı	4.275.910	4.275.910

3.2 Veri Madenciliği

Veri madenciliği aşaması, veri madenciliği araçları kullanarak temizlenmiş veri seti üzerinde algoritmaların çalıştırılması ve model oluşturulmasını kapsamaktadır. Çalışmamızda iki farklı veri madenciliği aracı kullanılmıştır. Veri madenciliği araçlarının kullandığı algoritmalar aynı olmalarına rağmen model oluşturma aşamasını tekrarlama sayıları farklı olabilmektedir. Bu kapsamda ürettikleri sonuçların doğruluk oranları da farklılık göstermektedir. Hazırlanan sistemde iki farklı aracın sonuçlarından en ideal olan karar destek sisteminin kurallarını oluşturmaktadır.

- K-Means Algoritması

K-Means algoritması en yaygın kullanılan kümeleme algoritmalarındandır. K değeri elimizde bulunan veri setinin kaç kümeye ayrılacağını belirtmektedir. Çalışmamızda birbiriyle benzerlik gösteren müşterilerin aynı etiket ile ifade edilmesi K-Means algoritması kullanılarak sağlanmıştır. Algoritma, müşterilerin mağazada bulunan 90 ürün kategorisinden aldığı ve almadığı ürünlere göre kümesini belirlemektedir.

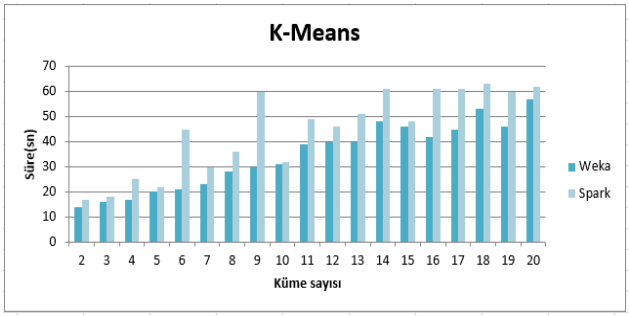
- Fp-Growth Algoritması

Birliktelik analizinde temel hedef büyük veri yığınlarında sıklıkla tekrarlanan desenleri, ilişkileri belirlemektir. Mağaza verilerinde birlikte satılan ürünleri belirlerken FP-Growth algoritması kullanılmıştır. Algoritma farklı en düşük destek değeri(Minimum support) ile çalıştırılmış ve anlamlı sayıda kural seti oluşturup oluşturmadığı incelenmiştir. 2133 ürün kategorisi içerisinden birlikte satılan ürünlerin belirlenmesi hedeflenmiştir.

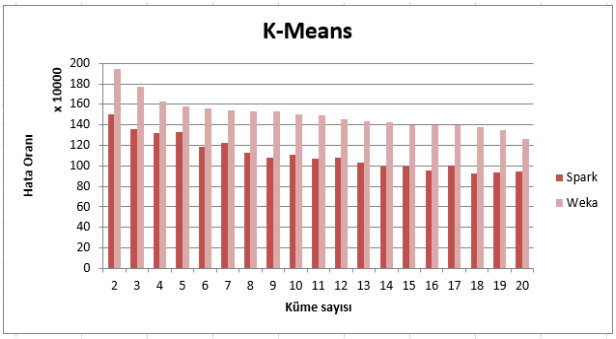
3.3 Sonuçların Değerlendirilmesi

- Kmeans Algoritması Sonuçları

Spark ve Weka'da temizlenmiş veri seti üzerinde farklı küme sayıları (2 ile 20 aralığında) ile K-Mmeans algoritması çalıştırılmıştır. Algoritmanın çalışma süreleri ve hata oranları göz önünde bulundurulduğunda paralel yapıda çalışıyor olmasına rağmen Spark'ın daha uzun sürede çalıştığı ve hata oranının Weka'dan az olduğu gözlemlenmiştir. Aynı algoritma olmasına rağmen kullanılan araçların analizi sonlandırma noktaları(tekrar sayıları) ve bununla birlikte hata oranları farklılık göstermektedir. Yani bir diğer ifadeyle; Spark küme dağılımını sabitleyerek analizi sonlandırdığı noktaya ulaşıncaya kadar çok daha fazla kez analizi tekrarlamaktadır. Bunun sonucu olarak da Spark'ın çalışma süresi daha uzun olurken analiz sonucunun doğruluğu artmaktadır. Her iki veri madenciliği aracınında sonuç üretme süresi saniye bazında olduğundan karar destek tarafından kullanılacak sonucun hata oranı en düşük olan araç ve küme sayısı olarak belirlenmesi anlamlı olacaktır. Sonuçları incelediğimizde müşterileri Spark tarafından küme sayısı 18 olarak gerçekleştirilen analiz sonuçlarının kullanılmasının, karar verme sürecine etkisinin daha olumlu olacağı belirlenmiştir.

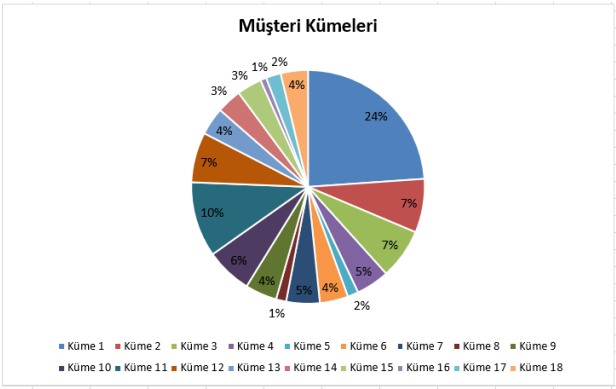


Şekil 2 K-Means Algoritması Çalışma Süresi



Şekil 3 K-Means Algoritması Hata Oranı

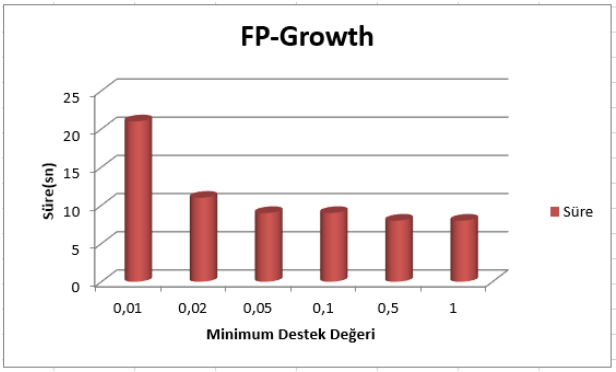
Spark tarafından 18 ayrı kümeye ayrılan müşterilerinin dağılımı grafik ile gösterilmiştir.



Şekil 4 Müşteri Kümeleme Sonucu

- Fp-Growth Algoritması Sonuçları

Weka'da veri boyutunun büyüklüğünden dolayı FP-Growth algoritması sonuç üretememiştir. Spark üzerinde çalıştırılan FP-Growth algoritmasına ait farklı minimum destek değerlerinde algoritmanın çalışma süreleri tabloda yer almaktadır.



Şekil 5 Fp-Growth Çalışma Süresi

Fp Growth algoritması kullanılarak 0,01 minimum destek ile 111, 0,02 minimum destek ile 40 adet birliktelik kuralı elde edilmiştir. 0,05 ve daha büyük değerler için kural seti bulunamamıştır. Elde edilen kural setlerine bir kaç örnek Tablo 3 de gösterilmektedir. Mağazanın bu sonuçlardan anlaması gereken; siyah kemer alan bir erkeğe siyah çorap ve siyah ayakkabı önermesi olmalıdır.

Tablo 2 Birlikte Satılan Ürünler

1)	Erkek Siyah Düz Takım Elbise, Erkek Beyaz Cvc Gömlek
2)	Erkek Mavi Gömlek, Erkek Beyaz Gömlek
3)	Erkek Lacivert Takım Elbise, Erkek Beyaz Gömlek
4)	Erkek Siyah Gömlek, Erkek Beyaz Gömlek
5)	Erkek Siyah Kemer, Erkek Siyah Çorap
6)	Erkek Siyah Ayakkabı, Erkek Siyah Kemer

SONUÇ

Weka ve Spark araçları kullanılarak, müşteri satış hareketleri analiz edilmiş ve sonuçlar yorumlanarak mağaza sistemi için doğruluk oranı en yüksek karar destek sistemi çalışmamız kapsamında geliştirilmiştir. Weka ve Spark üzerinde iki farklı algoritmanın sonuçları incelendiğinde; Spark üzerinde çalıştırılan K-Means algoritmasının her küme sayısı için hata oranının Weka'ya oranla daha düşük olduğu gözlemlenmiştir. Birliktelik analizi aşamasında kullanılan Fp-Growtth algoritmasını veri büyüklüğünden kaynaklı olarak Weka'da çalıştırmak mümkün olmamasına karşın Spark üzerinde kısa sürede sonuç üretilmiştir. Bu sonuçlar da bize büyük boyuttaki verilerin analizinin Spark ile mümkün olduğu ve Spark ile üretilen sonuçların doğruluk oranlarının daha yüksek olduğunu göstermiştir. Büyük veri setleri analiz edilirken Spark ile analiz etmenin performans açısından olduğu gibi doğruluk oranı olarak da avantaj sağladığı belirlenmiştir.

KAYNAKÇA

[1] "Veri Madenciliği" Url: http://www.sonsuz.us/veri-madenciligi/

[2] "Veri Madenciliği" http://mail.baskent.edu.tr/~20410964/DM_1.pdf

[3] "Veri Madenciliği" https://burakisikli.wordpress.com/2009/02/15/veri-madenciligidata-mining-nedir-ve-nerelerde-kullanilir-1/

[4] "Real-time Miscarriage Prediction with SPARK", H.Asri & H. Mousannif & H. Moatassime, The 7th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare (2017)

[5] “An implementation of Eclat on spark”, W.Mohamed & M.Abdel-fattah & S. El-Gaber, International Journal of Computer Science and Information Security (2017)

[6] “Analysis of Data Mining Tools for Disease Prediction”, K. Ahmed P, Journal of Pharmaceutical Sciences and Research (2017)

[7] “R-Apriori: An Efficient Apriori based Algorithm on Spark”, S. Rathee & M. Kaul & A.Kashyap, Conference on Information and Knowledge Management (2015)

[8] “Network Anomaly Detection by Cascading K-Means Clustering and C4.5 Decision Tree algorithm” A.Prabakar Muniyandi & R.Rajeswari & R.Rajaramca International Conference on Communication Technology and System Design (2012)

[9] http://www.sonsuz.us/wp-content/uploads/2017/05/veri-madenciligi-1-768x397.jpg

[10] ”Apache Spark ”

Url: https://spark.apache.org/docs/latest/index.html

[11] “Integration of Spark framework in Supply Chain Management”, H. Jaggi & S. Kadam, 7th International Conference on Communication, Computing and Virtualization (2016)

[12] ”Weka”

Url: https://www.cs.waikato.ac.nz/ml/weka/

ÖZGEÇMİŞLER

Gamze ÖZÇELİK

Dokuz Eylül Üniversite Bilgisayar Mühendisliği Bölümünde yüksek lisans eğitime devam etmektedir. Bir yazılım firmasında çalışmaktadır. Araştırma alanları arasında Büyük Veri, Veri Madenciliği ve Bulut Bilişim yer almaktadır.

Prof. Alp KUT

Dokuz Eylül Üniversitesi Bilgisayar Mühendisliğinde profesördür. 2003 sonbaharından beri Bölüm Başkanı olarak görev yapmaktadır. Veri Madenciliği, Veritabanı Yönetim Sistemleri ve Dağıtık Sistemlerle ilgili çalışmaları bulunmaktadır. Web Tabanlı Sistemler ve Paralellik de dahil olmak üzere çeşitli konularda birçok yayını bulunmaktadır.