CSE454 – DATA MINING

ÖDEV 2

Isolated Forest'a Dayalı Akış Verisi Anormallik Algılamasının Paralel Algoritması

Yue Liu, Yuansheng, Lou Sipei Huang

2020 Uluslararası Yapay Zeka ve Elektromekanik Otomasyon Konferansı (AIEA)

Can BEYAZNAR

161044038

Konferans Hakkında:

AIEA2020, Yapay Zeka ve Elektromekanik Otomasyon alanındaki yenilikçi akademisyenleri ve endüstri uzmanlarını ortak bir forumda bir araya getirmektedir. Konferansın birincil amacı, gelişmiş Yapay Zeka ve Elektromekanik Otomasyon alanındaki araştırma ve geliştirme faaliyetlerini teşvik etmektir; tüm dünyada çalışan araştırmacılar, geliştiriciler, mühendisler, öğrenciler ve uygulayıcılar arasında bilimsel bilgi alışverişini teşvik etmektir.

Makalenin Özeti:

Bu makalede, isolated forest algoritması geliştirilerek, hidrolojik alana uygulanmıştır. Paralel anormallik algılama algoritması (Flink-iForest) önerilmiştir. Aynı zamanda, kmeans algoritması, FlinkiForest threshold division problemini çözmek ve anormallik algılama sonuçlarının kararlılığını iyileştirmek için birleştirilir. Çeşitli deneyler ve gerçek hidrolojik veriler aracılığıyla, öncelikle Flink-iForest algoritması doğruluk, verimlilik ve ölçeklenebilirlik açısından doğrulanır ve standart SKlearn-iForest ve PIFH algoritması ile karşılaştırılır; son olarak, FlinkiForest algoritmasının etkinliği ve verimliliği deneylerle kanıtlanmıştır.

Giriş:

Hidroloji alanında kullanılan hidrolojik veriler, gitgide artan bir zaman serisi verileridir. Hidrolojik veriler oldukça yararlı bilgiler içermekte ve çeşitli bölgelerde kullanılmaktadır. Hidrolojik verilerin tahimini, erken uyarı, gizli anormal verilerin analizi ve karar verilmesi gibi alanlarda da kullanılmaktadır. Ancak verilerin hızla artmasıyla hidroloji alanında anomali tespit teknolojisi için yeni gereksinimler ortaya çıkmaktadır. Son yıllarda, çok sayıda hidrolojik verinin gerçek zamanlı olarak nasıl tespit edileceğine dair bazı araştırma sonuçları olmuştur. Wu Yafei, geleneksel algılama algoritmasının yerel çözümleri bütünleştirememesi sorununu çözen ve ardından bellek darboğazı sorununu çözen Hadoop anormallik algılama dağıtılmış hesaplamaya dayalı bir zaman serisi anormallik algılama algoritması önerdi. Bununla birlikte, algoritmanın verimliliğinin hala iyileştirilmesi gerekmektedir ve akış verisi ortamı için uygun değildir; Tian Lu, ilk kez spark streaming'e dayalı bir akış veri madenciliği algoritması önerdi. Mevcut kullanıcı veri akışını seçmek için kayan pencerenin boyutunu ayarlayarak ve kullanıcının mevcut güç tüketimi davranış modunu elde etmek için akış kümeleme algoritmasını kullanarak hızlı algılama gerçekleştirilebilir. Bununla birlikte, spark streaming, konveksiyon verilerinin gerçek zamanlı işlenmesini gerçekten gerçekleştirmeyen ve konveksiyon verilerini bir kez doğru bir şekilde tüketemeyen, kaçınılmaz olarak sonuçların doğruluğuna yol açacak olan mikro parti işlemine dayanmaktadır; Song Po, ilk olarak, özellik seçimi sorununu çözen ve yeni bir değerlendirme standardı ekleyen ve aynı zamanda akış verilerinin paralel işlenmesini sağlayan fırtınaya dayalı gerçek zamanlı bir ağ veri akışı anormallik algılama algoritması önermiştir. Bununla birlikte, akış verilerinin tek seferlik doğru tüketimini garanti edemeyen aynı problem hala mevcuttur ve algoritmanın doğruluğu diğer paralel algoritmalardan biraz daha düşüktür. Bu nedenle, bu makale hidrolojik zaman serilerindeki aykırı değerleri daha fazla kazmayı ve Apache Flink platformuna dayalı paralel anormallik algılama algoritmasını keşfetmeyi önermektedir.

Apache Flink Nedir? :

Apache Flink, Apache Software Foundation tarafından geliştirilen açık kaynaklı bir akış işleme çerçevesidir. Çekirdeği Kumaş veri işleme motorudur. Flink, paralel veri ve ardışık düzen modunda rastgele akış verisi programları yürütür ve Flink'in ardışık düzen çalışma zamanı sistemi, toplu işleme ve akış işleme programlarını yürütebilir. Ek olarak, Flink'in çalışma zamanının kendisi yinelemeli algoritmaların yürütülmesini destekler.

Isolated Forest Algoritması :

Isolated Forest algoritması, temel olarak tüm veri setindeki anormal noktaların küçük bir kısmını yakalar ve bu noktalar veri merkezinin özelliklerinden sapacaktır. Bu izole noktalar için iForest'in çok verimli bir stratejisi var. Bir veri uzayında, rastgele hiper düzlem, onu bölümlemek için kullanılır ve tek bir bölümleme, iki veri alt uzayına bölünebilir, aynı adım, her boşlukta sadece bir veri noktası olana kadar her alt alanı bölümlere ayırmaya devam eder. Her bölümleme tamamen rastgele olduğundan, Monte Carlo yöntemi bir yakınsama değeri elde etmek için kullanılabilir. iForest n ağaçtan (ikili ağaç) oluşur.

Algılama aşamasında, her verinin ormandaki iTree'den geçmesine izin verin, her ağaçtaki her verinin düğüm konumunu alabilir ve düğüm konumuna göre her ağaçtaki her verinin yüksekliğini hesaplayabiliriz. Şekil 1'de gösterildiği gibi, düğüm kök düğümden (root node) daha uzaktaysa, bu verilerin büyük olasılıkla normal bir değer olduğu anlamına gelir, aksi takdirde anormal bir değer olma olasılığı çok yüksektir ve kırmızı nokta çok daha fazladır.

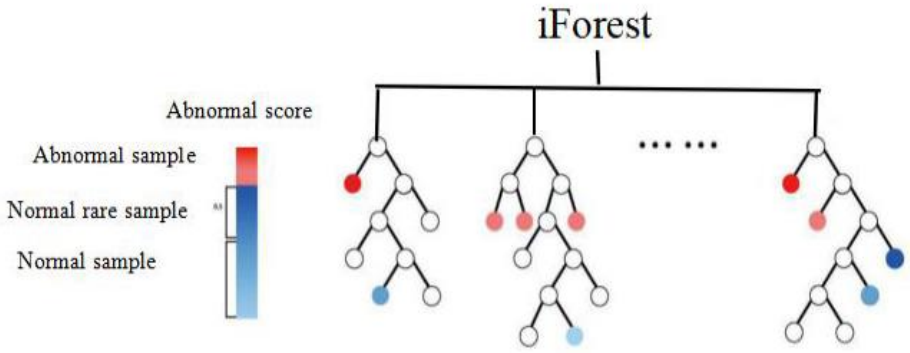


Figure 1. abnormal scores of isolated forest

Isolated Forest Algoritmasının İyileştirilmesi :

A) Hidrolojik alanda Isolated Forest algoritmasının sorunları

- iTree'yi oluşturmak için eğitim setinin yalnızca bir alt kümesinin rastgele seçilmesi gerekir.

- iForest algoritmasının doğrusal zaman karmaşıklığı, düşük sabit ve düşük bellek gereksinimleri vardır.

- iForest algoritması entegrasyon fikrine dayanmaktadır. Bazı ağaçların verimliliği çok yüksek olmasa bile, entegrasyon algoritması her zaman zayıf algoritmayı güçlü bir algoritmaya dönüştürebilir.

iForest algoritmasının, gerçek ortam gereksinimlerinde, isolated forest algoritmasının hala birkaç iyileştirmeye ihtiyacı vardır:

- iForest algoritmaları, tek bir makinenin statik veri setine dayanır ve veri ölçeği sınırlıdır ve kontrol edilebilir. Gerçek ortamda, hidrolojik test istasyonunun sensör verileri her 5 saniyede bir güncellenecektir, bu nedenle veri kaynağı sürekli, kontrol edilemez ve öngörülemez.

- Bazı özel uygulama senaryolarında, deney yalnızca bağımsız bir ortamda gerçekleştirilebiliyorsa, birçok işlemin sistemin yük dengesini ve kararlılığını koruması gerekir. O zaman makinenin hafızası veya diski kesinlikle bir gün dolacak ve tek bir makinenin hesaplama verimliliği sınırlı olacak, bu nedenle yeni bir gereksinim öneriliyor ve deney yapılırsa sistemin yük dengesini ve kararlılığını korumak için iForest yalnızca bağımsız bir ortamda gerçekleştirilebilir.

- Tam rastgelelik (Complete randomness), iForest algoritmasına avantajlar getirir, ancak aynı zamanda bazı kusurlara da yol açar. Örneğin, her bir hesaplama istisnasının eşik değeri değişebilir ve bu da tespit sonuçlarının doğruluğunun azalmasına veya yanlış değerlendirme olgusuna yol açacaktır.

B) Flink'e dayalı Isolated Forest'ın paralel algoritması üzerine araştırma

1) Hidrolojik akış verilerinin anormallik tespiti için genel çerçeve

Hidrolojik akış verilerinde anormallik tespitinin genel çerçevesi Flink platformunda, ilk iki modül şunlardır:

1) **Akış işleyici**
2) **Pencere işleyici**

Akış işleyici, her akış nesnesine bir eşleme işlevi uygular ve bunu pencereye gönderir. Pencere işlemcisi, her slaytta bir aykırı değer algılama algoritması çalıştırır. Akış

işleyicinin harita işlevinde, her hidrolojik veri noktası tek bir bölme penceresine gönderilir. Pencerenin, slaytta kalıcı olan kendi depolama kaydı durumu vardır.

Hidrolojik akış verileri kayan pencereye ulaştığında, ilk detektör, çoklu pencere verilerine dayalı iForest algoritması ile eğitilir. Test aşamasında, istisna algılayıcı, kayan penceredeki her bir durumu tespit etmek ve örneğin istisna puanına göre bunun bir istisna noktası olup olmadığına karar vermek için kullanılır. Bir kayan pencerede bir örnek tamamlandıktan sonra, istatistiksel sonuçlar, kayan pencerenin anormallik oranının u eşik değerinden daha düşük olması durumunda, kavram sapmasının bu zamanda meydana gelmediğini ve eğitimli anormallik detektörünün değişmeyeceğini göstermektedir. Aksi takdirde, kavramın değiştiği, eğitimli anormallik dedektörünün modifiye edilmesi ve yeniden eğitilmesi gerektiği ve ardından dedektörün, önceki dedektörü atmak için mevcut kayan penceredeki tüm durumlara göre güncellenip yeniden eğitildiği anlamına gelir.

2) İstisnaların uyarlanabilir eşik belirlemesi

Bu makale, iForest algoritmasını iyileştirmek için k-ortalamalı kümelemeye dayalı bir yöntem önermektedir ve anormal eşiği, gerçek zamanlı olarak eğitilen iForest modeline göre uyarlamalı olarak bölebilir. Flink-IForest algoritması ile hesaplanan anormallik skoru ikili bir sınıflandırma problemidir, daha sonra deneydeki k parametresi değeri 2 olarak alınmıştır.

Flınk Platformunda Geliştirilmiş Isolated Forest Algoritmasının Paralel Uygulaması:

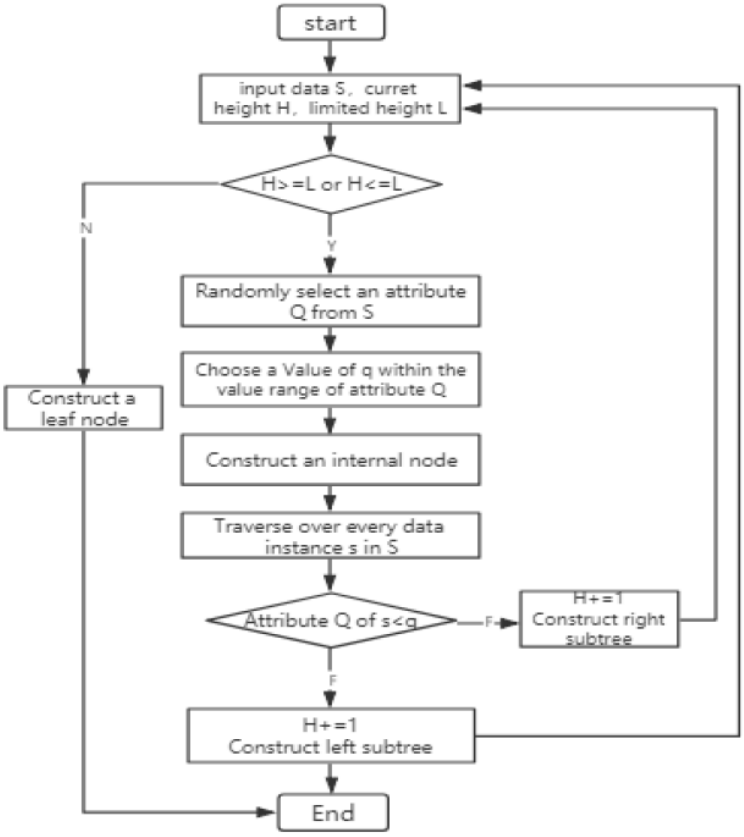
A) Paralel eğitim modelinin oluşturulması

Örneklenen verilerde ParameterMap işleminin süreci paraleldir ve Flink'in map () operatörü paralel hesaplama yapar. map()'te karakteristik örnekleme, iTree yükseklik sınırı hesaplama, yinelemeli iTree yapısı vb. Vardır.

Flink-IForest, özellik örneklemesi için bir parametre (maxFeatures) kabul eder. Bir sonraki adım daha önemlidir, çünkü ağacın yüksekliği sonraki anormallik tespitinin performansını doğrudan etkileyecektir. Ağacın yüksekliği, algoritmanın erken bitmesi için sınırlandırılmalıdır. Ağacın yüksekliği maxDepth tarafından kontrol edilecektir. Kod uygulandığında, her örnek alt uzaydaki veri sayısına göre tam sayı log2 (aşağı doğru) alınır. Bu sayı, ağacın maksimum yükseklik sınırı Yoludur ve son yükseklik, ikisinden küçük olanıdır.

İTree'yi inşa etmek, Flink-iForest'i gerçekleştirmenin temel parçasıdır. Yinelemeli bir süreçtir. Bir yaprak düğüm oluşturulduğunda, özyinelemeli döngüden çıkar ve yaprak düğüm, belirli veri örneklerini depolar. Daha da önemlisi, eğitim verileri bölünür ve iTree'nin dahili düğümlerine

kapsüllenir ve dahili düğümlerin sol ve sağ alt ağaçları yinelemeli olarak oluşturulur. Şekil 2'de



gösterildiği gibi. Figure 2. iTree Construction flow chart

B) Paralel anormallik tespitinin uygulanması

Bu bölümde uygulanan Flink-iForest algoritması, paralel istisna algılama ve algılama aşamasında Flink 1.8.0'dan sonraki sürümün kullanımını yansıtabilir. Paralel algılama sürecinde, iForest modeli yayın yoluyla her bir TaskManager'a iletilecek ve kayan penceredeki her veri parçası, istisna değerlendirmesi için tüm ağaçların üzerinden geçecek ve anormallik tespiti, hesaplanan anormallik puanı ve uyarlanabilir eşik ile gerçekleştirilecektir.

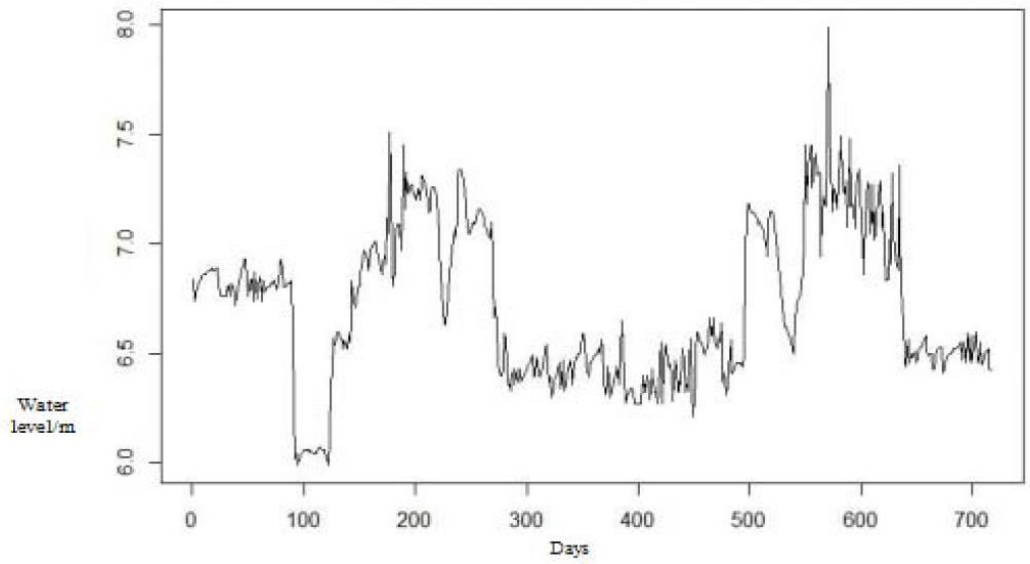
Burada, farklı özdeğerlere sahip örnekleri alabileceğimiz Flink Tablosunun ScalarFunction'ı kullanıyoruz ve ardından iForestModel aracılığıyla, ağaçtaki örneklerin avgHeight'ını alabilir ve istisnaları avgHeight'a göre puanlayabiliriz.

Vektör kabı, veri örneğinin öz vektörünü alır. İlk olarak, iTree'nin ortalama yüksekliği, örnek sayısı ile hesaplanır; daha sonra ormandaki her ağacın yüksekliğinin toplamının ortalama değeri iForestModel tarafından hesaplanır. Ağacın yaprak düğümüyle çapraz geçiş sürecinde karşılaşılırsa, ağaçtaki verilerin yüksekliği doğrudan döndürülür; yaprak düğümle karşılaşılmazsa, karakteristik değeri yargılanacaktır; iç düğümün karakteristik indeksinden

büyükse, sağ çocuğa özyinelemeli olarak erişilecektir; aksi takdirde, sol çocuğa özyinelemeli olarak erişilecektir. Son olarak, transform () yöntemini uygulamamız gerekiyor.

Deneysel sonuçlar ve analiz :

Şekil 3, Chuhe Nehri Havzasındaki Liuhe hidroloji istasyonunun günlük ortalama su seviyesini, belirgin anormal noktalar ile göstermektedir.



1) Performans karşılaştırma deneyi

İlk olarak, algoritmanın AUC değeri, algoritmanın doğruluğu ile karşılaştırılır. Tek çekirdekli yapılandırma ortamında, orijinal belgede bağımsız iForest, PIFH, SKlearn-iForest ve Flink-iForest'i karşılaştırın ve belirli sonuçlar Şekil 4'te gösterilmektedir.

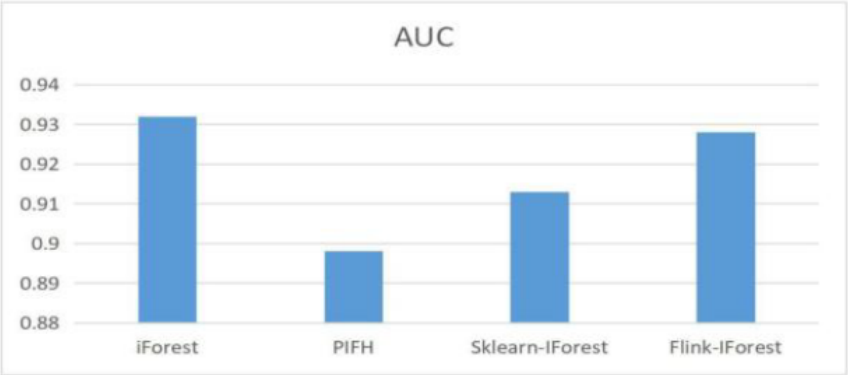


Figure 4. AUC value comparison of different algorithms

Deneysel sonuç karşılaştırma tablosundan, bu makaledeki geliştirilmiş Flink-IForest algoritmasının, hidrolojik verilerin spesifik senaryosunda orijinal makaleden daha düşük bir AUC değerine sahip olduğu, ancak PIFH ve SKlearn-IForest'inkinden daha düşük olduğu görülebilir. Daha yüksek AUC değerleri, algoritmanın kendisinde yetersiz anormallik tespiti nedeniyle sınırlanabilir. Bu yazıda uygulanan algoritma, doğruluk açısından anormallik tespitinin gereksinimlerini karşılamaktadır.

O zaman algoritmanın uygulama zamanını karşılaştırmamız gerekiyor. İlk deney grubu şu şekildedir: aynı CPU koşulu altında, SKlearn-iForest'in çalışma belleği 64G ile sınırlıdır; ikinci deney grubu şudur: Flink-iForest algoritması bir görev gönderdiğinde, tek bir düğümü 2G belleğe ayarlar ve CPU çekirdek sayısı 1 çekirdekten 4 çekirdeğe kadardır. Üçüncü deney grubu: PIFH algoritması, Flink-iForest algoritması ile aynı deneysel işlemi ve ortamı kullanır ve deneysel sonuçlar Şekil5'te gösterilmiştir.

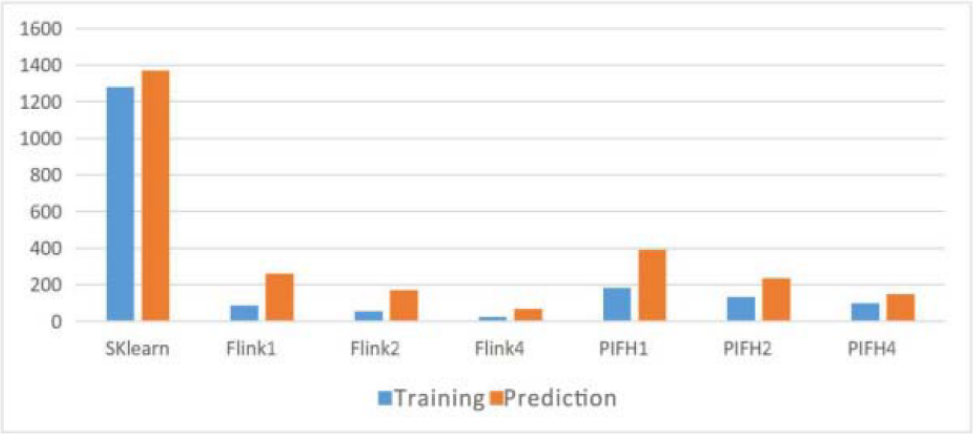


Figure 5. Algorithm performance comparison chart

Deneysel sonuçlar, tek çekirdekli bir ortamda, eğitim aşamasındaki SKlearn-IForest algoritmasının zaman maliyetinin Flink-IForest algoritmasının 15 katı olduğunu göstermektedir. Tahmin aşamasında, zaman ek yükü SKlearn-IForest algoritması Flink-IForest'in 5 katıdır. PIFH algoritması açıkça SKlearn-IForest algoritmasından daha fazla zaman kazandırıyor, ancak Flink-IForest ile hala bir boşluk var. Paralellik derecesi artmaya devam ederken, Flink-IForest'in avantajları artıyor. 4 çekirdekli durumda, Flink-IForest'in tahmin aşaması, tek çekirdekli durumdan üç kat daha hızlı olan 68 saniye sürüyor. PIFH algoritmasının tespit süresi, tek çekirdekten yalnızca 2 kat daha hızlıdır ve Flink-IForest'ten daha fazla zaman alır. Özetlemek gerekirse, Flink-iForest algoritmasının tek sürümü, standart SKlearn-IForest algoritmasından ve PIFH algoritmasından daha iyi performans gösterir. Algoritmaların artan paralelliği ile Flink-iForest algoritmasının avantajları daha da ön plana çıkmaktadır.

2) Genişletilebilirlik deneyi

Flink-IForest algoritmasının paralelliği, TaskManager sayısı arttıkça artabilir. Aynı akış veri setini kullanmaya devam edin, Flink platformundaki bellek her çalışan düğüm için 2G'ye ayarlandı ve çekirdek sayısı (paralellik) 1 çekirdekten 4 çekirdeğe yükselmeye devam ediyor. Deneysel parametreler: 100 iTrees, örnek alt uzay boyutu 256'dır. Paralelliğin artması durumunda, eğitim aşamasında Flink-IForest algoritmasının süresi ve anormallik algılama aşamasının yürütme süresi (saniye) hızla azalır ve sonuçlar Şekil 6'da gösterilmiştir.

Figure 6. expansion curve of Flink-iForest algorithm

Grafikten eğitim ve tespit süresinin çok hızlı düştüğü ve bunun logaritmik düşüş hızına ulaşabildiği görülebilir. Flink-i Forest algoritmasının artan paralellik durumunda gittikçe daha fazla performans avantajına sahip olduğunu ve büyük veri ortamında algoritmanın gereksinimlerini karşıladığını göstermektedir.

Sonuç:

Bu yazıda hidrolojik zaman serisi verileri için anormallik tespiti yapılmıştır. Bu makale esas olarak Flink'e dayalı iForest algoritmasının paralel uygulamasını tanıtmaktadır. İlk olarak, Flink'in özelliklerini ve iForest algoritmasının hidroloji alanında geliştirilmesini tanıtıyor. Daha sonra eğitim modelini ve paralel süreci gerçekleştirir. Son olarak, Flink-iForest algoritmasının geçerliliği ve verimliliği deneylerle kanıtlanmıştır.

Yayının Literatüre Katkısı:

Okumuş olduğum makalenin literatüre katkısı, hidroloji alanında yapılan çalışmaların iyileştirilmesine yardımcı olmaktadır. Ve anormallik tespiti yöntemini kullanarak, yeni çözümler sunmaya katkı sağlamıştır. Ayrıca okuduğum makale oldukça anlaşılır ve bilgiler gayet açıktır. Bu yüzden bu konu ile ilgilenmek isteyen kişiler için uygun bir başlangıç yazısı olabilir.

Deneylerinde elde ettiği sonuçları, başarı oranlarını detaylı bir şekilde göstermektedir. Ayrıca uyguladığı yöntemi detaylı bir şekilde açıklayarak, bu yazıyı okuyan insanların, bu metodu uygulamasına yardımcı olmaktadır. Isolated forest algoritmasının sahip olduğu dezavantajları ayrıntılı bir şekilde açıklamaktadır. Ve yazıda bahsedilen bazı sorunların iyileştirilmesi için anlattığı yöntemler oldukça anlaşılır bir şekilde yazılmıştır. Bu yazının, bu konuda sorun yaşayan araştırmacılar için iyi bir bilgi kaynağı olduğunu düşünmekteyim.