11.12.2020
Ödev No: 2

Veri Madenciliği

Review of Data Preprocessing Techniques in Data Mining

Journal of Engineering and Applied Sciences

Ergün Elvan Bilsel

Bilişim Enstitüsü – 198004709

1. Özet

Veri madenciliği, büyük bir veri kümesinden faydalı model ve modellerin çıkarılması sürecidir. Bu modeller ve kalıplar, bir karar verme görevinde etkili bir role sahiptir. Veri madenciliği temelde verilerin kalitesine bağlıdır. Ham veriler genellikle eksik değerlere, işlenmemiş verilere, eksik verilere, tutarsız verilere ve aykırı değer verilerine karşı hassastır. Bu nedenle, bu verilerin çıkarılmadan önce işlenmesi önemlidir. Verileri ön işleme, veri verimliliğini artırmak için önemli bir adımdır. Veri ön işleme, veri setinin veri hazırlanması ve dönüştürülmesi ile ilgilenen ve aynı zamanda bilgi keşfini daha verimli hale getirmeyi amaçlayan en veri madenciliği adımlarından biridir. Ön işleme, temizleme, entegrasyon, dönüştürme ve azaltma gibi çeşitli teknikleri içerir.

2. Çalışmanın Tanımı/Amacı/Kapsamı

Veri tabanında bilgi keşfi (Knowledge Discovery in Databases-KDD) verideki:

- **üstü kapalı** (anlaşılması güç ham veri)
- **geçerli** (bulunan yeni veriler üzerinde de geçerli olacak şekilde)
- **özgün** (beklenen değerlere kıyasla)
- **potansiyel olarak kullanışlı** (başarılı sonuçlara neden olabilecek)
- **anlaşılabilir** (yorumlanması kolay)

pattern'lerin bulunması ve tanımlanması sağlayan KDD süreçinin metodolojilerinin açıklanmsı.

3. Çalışmada Önerilen/İncelenen/Uygulanan Yöntem Detayları

Hedef Veriler:

Veri Tabanlarında Veri Madenciliği ve Bilgi Keşfi (Knowledge Discovery in Database – KDD)büyük veri kaynaklarından değerli bilgilerin çıkarılması sürecidir. Veri madenciliği, sınıflandırma, kümeleme, ilişkilendirme kuralları ve diğer birçok tekniği kullanarak büyük veri kümeleri için analiz ve modellergerçekleştiren bir KDD adımıdır.

Veriler önceden işlendikten sonra, veri madenciliği sürecine uygun forma dönüştürülmelidir. Daha sonra son

aşamada kesintiye uğrayan ve değerlendirilen örüntüleri çıkarmak için kümeleme, sınıflandırma, regresyon vb. Madencilik prosedürü

uygulanacaktır.

Veri temizleme:

Satır verilerinde eksik kayıtlar, gürültü değerleri, aykırı değerler ve tutarsız veriler olabilir. Veri temizleme, eksik değerleri bulmak, düzgün gürültü verilerini bulmak, aykırı değerleri tanımak ve tutarsızlığı düzeltmek için kullanılan veri ön işleme tekniklerinde ilk adımdır.

Ön işleme teknikleri:

Veri ön işleme, verilerin madencilik prosedürüne uygun bir forma hazırlanmasını ve dönüştürülmesini içeren en çok veri madenciliği görevlerinden biridir. Veri ön işleme, veri boyutunu küçültmeyi, veriler arasındaki ilişkileri bulmayı, verileri normalleştirmeyi hedefler.

Eksik değerler:

Kayıtları için kaydedilmemiş değerleri olan kayıtlar varsa, bu değerler aşağıdaki yollarla doldurulabilir:

1. **Eksik değeri doldurmak için ortalama öznitelik özelliğini kullanılması:** Bu yöntem, belirli bir özniteliğin eksik değerini o özniteliğin
2. **Veri demetinin yok sayılması:** Bu seçim, sınıf etiketinin değeri olmadığında seçilir (sınıflandırma madenciliği göreviyle birlikte kullanılır). Bu yöntem etkili değildir, ancak demet, boş değerlere sahip çeşitli özniteliklere sahip olduğunda kullanılır.
3. **Eksik değeri manuel olarak doldurmak:** Bu yaklaşım genel olarak insan çabası ve zaman gerektirir. Büyük boyutlu veri kümesiyle kullanılamaz.
4. **Verilen grupla aynı sınıfa ait tüm örnekler için ortalama öznitelik kullanılması:** Örneğin, kullanıcıları kredi riskine göre sınıflandırırsak, eksik değer, benzer kredi riski sınıfına ait kullanıcılar için ortalama gelir değeri ile değiştirilebilir. Bu yöntem, belirli bir özniteliğin eksik değerini o özniteliğin ortalama değeriyle değiştirerek çalışır.
5. **Eksik değeri doldurmak için genel bir sabit kullanın:** Bu yöntem, özniteliğin eksik değerlerini, tüm kayıtlar için benzer olan belirli bir sabitle değiştirerek çalışır, örneğin etiket olarak "Bilinmeyen" kullanmak.

6. **Eksik değeri doldurmak için en olası değeri kullanmak:** Bu yaklaşım, bir karar ağacı indüksiyonu veya Bayes biçimciliği kullanan çıkarıma dayalı regresyon gibi tekniklerle kullanılır.

Noise verileri:

Madencilik sürecini en çok etkileyen sorunlardan biri noise'dir. Noise, ölçülen bir değişkendeki rastgele bir hatanveya varyanstır. Noise verileri, verilerde veya aykırı değerlerde normalden sapan bir hata olduğu anlamına gelir.

1. Bölme: Bu yöntem, depolanan verileri, etrafındaki değerler olan "komşuluğuna" dayalı olarak yumuşatmak için çalışır. Sıralanan değerler, "kümeler" veya bölmelerden oluşan bir kümeye bölünür. Bu yöntemler komşunun verilerine bağlı olduğundan, yerel yumuşatma gerçekleştirirler.
2. Kümeleme: Kümeleme, bir mesafe ölçüsüne göre noktaların kümeler halinde gruplanması olarak tanımlanır. Kümelemenin sonucu, her bir kümenin birbirinden küçük mesafede ve diğer kümelerden büyük uzaklıkta olan bir dizi noktaya sahip olacağı kümeler kümesidir. Bu teknik, benzer noktaları bir kümede gruplandırırken, kümelerin dışında kalan noktalar aykırı noktalar olarak kabul edildiğinden, aykırı değerleri tespit edebilir.
3. Regresyon: Bu, moothie'nin verilerini bir işleve uydurarak yöntemler. Doğrusal regresyonörnek olarak, iki değişkene (veya özniteliklere) uyacak en iyi çizginin belirlenmesini içerir, böylece her bir öznitelik diğerini tahmin etmek için kullanılabilir.

Veri entegrasyonu: Birden çok veriyi tek bir yerde birleştirme işlemidir. Birleştirme sırasında yapıların birleştirilmesi, yapıların eşleşmesi, azaltma göz önünde bulundurulması gereken önemli noktalardır.

Veri dönüşümü:

1. Yumuşatma: Gürültülü(noise) verileri yok etmek için kullanılır.
2. Toplama: Medyan, varyans gibi istatiksel değerleri özetlemek için kullanılır.
3. Genelleştirme: Alt seviyedeki bilgileri daha üst seviyeye yuvarlamak için kullanılır.
4. Normalleştirme: Bu metot veriler 0-1 gibi belirli bir tam sayı değerine yuvarlamak için kullanılır.
5. Veri azaltma: Veri setinin boyutunu azaltmak için kullanılır.

4. Model Başarısını Değerlendirme Yöntemi

Çalışmada veri setini anlamlandırmak ve işleme hazırlamak için yöntemlerden bahsedilmiştir.

5. Deneysel Çalışma Detayları

Gerçek dünya verileri eksik, tutarsız, gürültülü ve eksik olma eğilimindedir; veri ön işleme, hem veri ambarlama hem de veri madenciliği için önemli konulardan biridir; veri ön işleme, veri temizleme, veri entegrasyonu, verileri içerir. Dönüştürme ve veri azaltma; Bu çalışma, veri ön işleme yöntemine genel bir bakış ve örneklerini açıklamaktadır. Veri ön işlemenin amacı, veri madenciliği, metin madenciliği ve veri madenciliği gibi her türden araştırma için nitelik verilerine verilir.

Veri temizleme yöntemi, gürültülü verilerin giderilmesi, tamamlanmamış veriler üzerinde tamamlanması ve gereksiz verilerin kaldırılması için kullanılır.

Veri ön işleme tekniklerinin, büyük veri ölçeğinde hazırlanmasında, analizinde ve işlemesinde verimli, etkili ve önemli bir role sahip olduğu sonucuna varılmıştır.

6. Çalışmanın Güçlü ve Zayıf Yönleri

Veri ön işleme, veri boyutunu küçültmeyi, veriler arasındaki ilişkileri bulmak için kullanılan tekniler örneklendirilerek incelenmiştir. Bu şekilde daha anlaşılabilir bir hale getirilmiştir.

Çalışmanın incelediği yöntemlerin kullanım istatislikleri ve kullanım sıralamalarına detaylıca yer verilmemiştir.