Veri Madenciliği Projelerinin Yaşam Döngüsü - 1

Özet : Bu makalemizde Veri Madenciliği projelerinin yaşam döngüsünü inceleyeceğiz.Veri Madenciliği projelerinde takip edilmesi gereken başlıca adımları ve bu adımlarda yapılması gereken temel işlemlere ışık tutacağız.

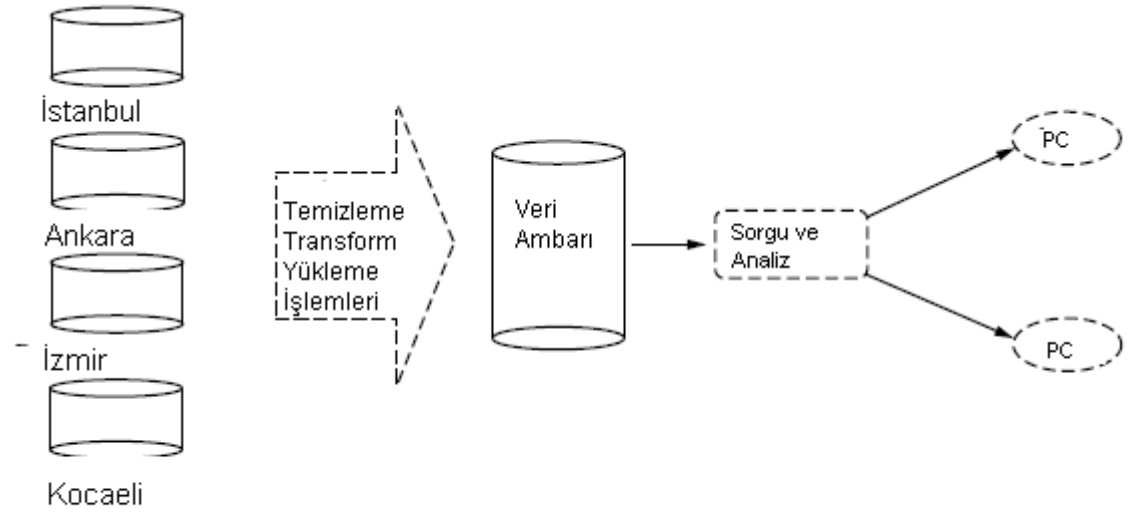
Makale : Veri Madenciliği Projelerinin yaşam döngüsü sırasıyla şu aşamalardan oluşmaktadır.

1) Data'nın Toplanması : Veri Madenciliği projeleri için ilk adımımız genellikle data'nın toplanmasıdır.Veri Madenciliğinde kullanacağımız dataları, Database ,Datamart yada Datawarehouse 'umuzdan data analizi için toplamımız gerekmektedir.

Data'yı toplayacağımız kaynaklar yada kullanabileceğimiz data çeşitleri ise şunlardan oluşabilir.(bunlara bağlı olarak data'yı alacağımız yere göre de Veri Madenciliği imkanları değişiklik göstermektedir.)

a) İlişkisel Veri Tabanları :RDMS (Relational Database Management System) olarak adlandırılan Veri Tabanı ,temel veri işlemlerinin yapıldığı (DML) yerdir.Bu tip veritabanlarında Normal Formların kurallarına göre tasarlanmıştır. İlişkisel Veri Tabanları için Entity-Relationship (ER) veri modeli tasarlanmıştır ve bu model, tablolar arasındaki ilişkileri gösterir.(İlişkisel Veri Tabanları hepimizin günlük hayatımızda kullandığımız MS-SQL Server , Oracle ,DB2 ,MySql gibi sistemlerdir)

b) Veri Ambarları : Veri Ambarları; RDMS gibi farklı kaynaklardan bilgilerin toplandığı ortak bir alandır.



4 Farklı Şehirde Veri Tabanı Olan Bir Şirket İçin Örnek Uygulama

Veri Ambarlarında,RDMS sistemlerin tersi bir şekilde sadece işimize yaracak özet bilgiler tutulur.(örneğin satışların bölgelere yada aylara göre özeti gibi) Veri Ambarları genellikle çok boyutlu database yapıları gibi modellenir ve her boyut bir attribute dür , her hücre ise bir toplamdır (toplam satış miktarı gibi).Gerçekte Veri Ambarları'nın fiziksel yapısı ilişkisel data alanları yada çok boyutlu veri küpleri (Data Cube) şeklinde olabilir.

c) İşlemsel Veri Tabanları (Transactional Databases) : Herbir kayıdın bir hareketi(transaction) gösterdiği dosyalardır.Transactional dosyalarda herbir kayıt bir ID ile ifade edilir ve aynı ID değerine sahip transactionda yapılan işlemler sıralanır.

IslemID 'leri	UrunID 'leri
T100	I1, I3, I8, I16
T200	I2, I8
...	...

Yapılan İşlemler ve Alınan Urunlerin ID'leri

Transactional Databaselerde "T100 numaralı işlemde satın alınan ürünler hangisidir ? " gibi soruların cevabı bulunabilir.Birliktelik (Association) ilişkilerinde sıkça kullanılacak dataları içermektedir.

d) Uzaysal Veritabanları (Spatial Database) : Uzaysal veritabanlarına harita veritabanları ve uydu görüntüleri örnek olarak verilebilir.Orman ve Ekolojik planlamada,telefon ve elektrik kablolarının döşenmesi gibi kamu hizmetlerinin kullanımda bu tip data kullanılmaktadır.

e) Metin Veritabanları ve Multimedya Veritabanları : Metin veritabanları uzun cümleler ve paragraflardan oluşan ,içerisinde uyarı mesajları ,buglar raporları gibi metinsel ifadeler içeren veritabanlarıdır.Metin verileri üzerine yapay zeka algoritmaları da kullanılarak müşterilerden gelen talep ve istekler üzerine CRM projeleri geliştirilebilir.

Multimedya veritabanları ise görüntü,ses,video verisi gibi verilerden oluşmaktadır.Ses ve görüntü tanıma temelli projelerde kullanılacak verilerdir.

f) İnternet (The World Wide Web) : Kullanıcıların internetde bırakmış oldukları verilerdir diyebiliriz.Şöyle ki kullanıcıların bir alışveriş sitesinde tıkladığı linkleri analiz ederek doğru reklam politikaları izlenebilir.

2) Data'nın Temizlenmesi ve Yeniden Yapılandırılması (Cleaning and Transformation) :

Veri Madenciliği projelerinin 2.aşaması olan Data'nın Temizlenmesi ve yeniden yapılandırılması (data cleaning and transformation) aşaması yoğun bir şekilde,veri kaynağıyla ilgili işlemleri içermektedir.

Data'nın temizlenmesinden kasıt; gürültülerin (*yanlış yada aşırı uç değerlere sahip verilere gürültülü veri denir.örneğin doğum tarihinin 1200 olması gibi*).

Data'nın temizlenmesi ve yeniden yapılandırılmasında uygulanan yöntemler ise şunlardır.

a) Data Tipinin Transformasyonu:Basit olarak veri tipinin türünün yeniden yapılandırılmasıdır.Bazı Veri Madenciliği algoritmaları sadece integer (sayısal) tiplerdeki verilerle hızlı bir şekilde çalışırken,kimisi de mantıksal verilerle(boolean) hızlı bir şekilde çalışmaktadır.

b) Sürekli Kolonların Transformasyonu : Bu yeniden yapılandırma türünde; sürekli veriler Normalizasyon işleminden geçirilmektedir.Örneğin 500 TL ile 20000 TL arasında değişen maaş verilerini 4 gruba bölmüş olalım (500-1000,1000-5000,5000-10000,10000-20000 gibi). Yapay Sinir Ağları benzeri algoritmalar bu verileri kabul etmeyecektir.İşte eldeki bu gibi verileri 0.0 - 1.0 gibi aralıklara indirme işlemine *Normalizasyon* denmektedir.(*Normalizasyon işlemi için çeşitli yöntemler vardır.Bunlar;min-maks normalizasyonu,sıfır ortalama normalizasyonu,ondalıklı normalizasyondur*.)

c) Gruplama :Gruplama işlemiyle,aslında ayrı gibi görünen bölümlerin ortak bir paydada birleştirilmesi söz konusudur.Örneğin;BilgisayarMühendisliği ,Elektrik Mühendisliği,Endüstri Mühendisliği,Eczacılık,Doktorluk gibi ayrılmış meslek gruplarımız olsun.Bu meslek grupları yeniden yapılandırılarak Mühendislik,Eczacılık,Doktorluk gibi daha düzgün bir sınıflandırmaya hazır hale getirilebilir.Bu uygulama,bize zamandan kazanç olarak geri dönecektir.

d) Kümeleme : Kümeleme ise ; bir başka verinin yeniden yapılandırma sürecidir.Örneğin bir GSM operatörü müşterilerini aylık konuşma verilerine göre segmente etmeye çalışıyor olsun.Çözüm olarak çok fazla detaylı bilgiden sıyrılabilmek amacıyla,toplam görüşme sayılarına göre kümeleme yapılmalıdır.(Kümeleme aslında bir veri madenciliği modelidir.)

e) Kayıp Verilerin İşlenmesi : Verilerin yeniden yapılandırılması aşamasında bir diğer önemli konu ise kayıp yada Null değerlerin ne olacağı sorusudur.İki farklı OLTP sisteminin birleştirilmesi sonucunda kayıp değerler ortaya çıkabileceği gibi bilgi giriş elemanları yada müşteriler tarafından bilerek yada bilmeyerek yanlış veya boş değerler(Null Values) oluşabilmektedir.

Gerçekleştirilecek projenin ve kayıp,yanlış olan verilerin durumuna göre farklı çözümler bulunabilir.

i)*Kayıp verilerin bulunduğu kayıdı, veri kümesinden çıkarmak yada bu gibi kayıtları iptal etmek*.(Eğer kayıp verinin miktarı toplam verinin içinde küçük bir değerse)

ii) *Kayıp verileri elle teker teker doldurmak* (Kullanılan Veritabanı küçükse ve gerçek hayatta kayıp verilere ulaşmak kolay ve zaman problemimiz yoksa)

iii) *Tüm kayıp verilere aynı bilgiyi vermek*.Örneğin doğum tarihi bilgisini vermemiş müşterilerimiz varsa bunlar için DTY(doğum tarihi yok) şeklinde bir veri girişi yapılabilir.Ama buradan çok farklı bir sonuç ortaya çıkıp;doğum tarihini vermemiş olan kişilerin bir ortak özelliği olduğu ve aynı davranışı sergiledikleri , tahmin edilemeyen bir satış fırsatını ortaya çıktığı durumlar da olabilir.(Örneğin doğum tarihini yazmayan kişilerin bakım ürünlerini daha çok satın alması gibi)

iv)*Kayıp olan verilere tüm verilerin ortalama değerinin verilmesi.*

v)*Regresyon yöntemi kullanılarak, diğer değişkenlerin yardımıyla kayıp olan verilerin tahmin edilmesi.*

f) Uç Verilerin Ortadan Kaldırılması : Bazı durumlarda aşırı uç veriler(ortalama değerlere göre çok düşük yada çok yüksek değerlere sahip veriler) projenin başarı oranının düşmesine neden olabilir.Eğer bu veriler oran olarak kayıda değer bir sayıda değilse,yok sayılabilir.(Hassasiyeti etkileyecek seviyedelerse faydadan çok zarar da oluşabilir.)

Bunlar dışında da birçok veri temizleme ve verileri yeniden yapılandırma tekniği mevcutdur.SQL Server Integration Services(SSIS)yardımıyla,bu makalede saymış olduğum teknikler uygulanabilmektedir.

Bu makalemizde Veri Madenciliği yaşam döngüsünün ilk 2 aşaması hakkında (Data'nın toplanması ve Temizlenmesi ,Yeniden Düzenlenmesi) bilgiler vermeye çalıştım.Bir sonra ki bağlantılı makalemde ,sonraki aşamalar hakkında detaylı bilgiler sunmaya devam edeceğim.

Başka bir İş Zekası makalesinde görüşmek dileğiyle...

Bilgehan GÜRÜNLÜ

www.gurunlu.com
bilgehan@gurunlu.com

Kaynaklar :

Data Mining with Sql Server 2005.

Data Mining:Concepts and Techniques.

Kavram ve Algoritmalarıyla Veri Madenciliği (G.Silahtaroğlu)

Veri Madenciliği Projelerinin Yaşam Döngüsü - 2

Özet : Bir önceki makalemizde Veri Madenciliği Projelerinin yaşam döngüsünün ilk 2 adımı olan Data'nın toplanması ve Data'nın temizlenip,yeniden yapılandırılması aşamalarını detaylıca incelemiştik.Bu makalemizde ise;Model Oluşturma,Modelin Keşifi (Doğrulanması,değerlendirilmesi) ,Raporlama,Tahmin(Skorlama) ,Uygulamalarla Entegrasyonu ve en son olarak da Modelin yönetilmesi aşamalarını inceleyeceğiz.

Makale : İlk 2 aşamamızı bir önceki makalemizde anlatmıştık ,şimdi kaldığımız yeren itibaren (3.aşamadan itibaren) incelemeye devam ediyoruz.

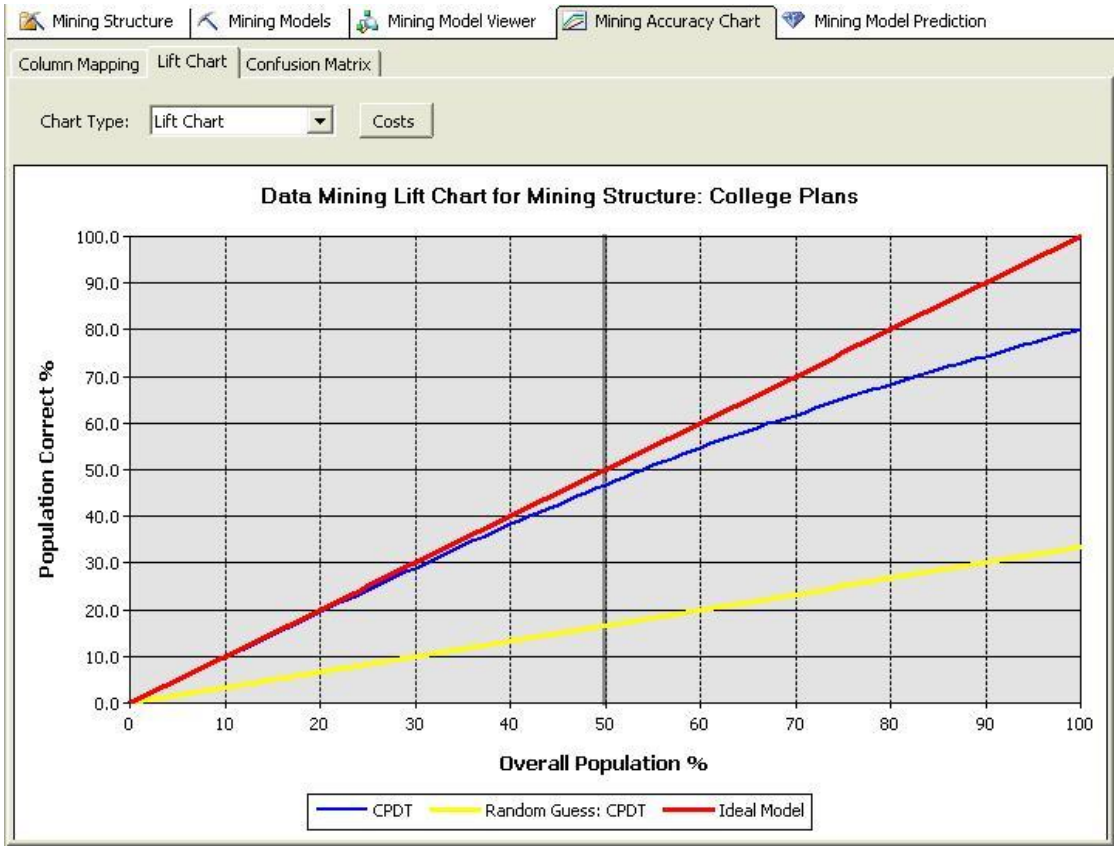
3) Model Oluşturma (Model Building) : Veriler temizlendikten ve değişkenler yeniden düzenlendikten sonra sıra geldi Veri Madenciliği Modelimizi oluşturmaya.Veri Madenciliği projemizde modelimizi oluştururken hedeflerimizin neler olduğunu ve hangi tip verilerle nasıl bir veri madenciliği görevini gerçekleştireceğimizi asla unutmamalıyız.Projemizin;bir sınıflandırma(classification) mı ? Birliktelik(Association,Market Basket Analyse) mi ? yoksa bir segmentasyon projesi mi olacağı gerçeğinden yola çıkarak modelimizi oluşturmalıyız.Model oluştururken iş analistlerimizle beraber ortak kararlar almak zorundayız.Örneğin telekom sektörü için bir proje gerçekleştiriyorsak;ilgili uzmanlık yetkinliklerine sahip ,bölüm yöneticilerini de proje ekibine dahil etmek akıllıca olacaktır.

Model oluşturmak veri madenciliği için çok önemli bir aşamadır.Bu aşamada öncelikle hangi tip veriyle,hangi tip veri madenciliği görevini gerçekleneceği çok iyi anlaşılmalı ve buna uygun algoritmalar seçilmelidir.Bazı durumlarda,modelin eğitilmesi öncesinde,hangi algoritmanın elimizdeki data için uygun olduğu bilinmeyebilir.Bu gibi durumlarda attribute'ların ilişkileri incelenerek hangi alogitmanın kullanılacağına karar verilebilir.Örneğin input *attribute*lar ve tahmin edilecek(class attribute) arasında liner bir ilişki mevcutsa karar ağaçları (Decision Tree,classification algorithm) kullanılması yerinde olacaktır.Bir başka örnek için attribute'lar arasında ilişki karmaşık ise yapay sinir ağı algoritmaları kullanılmalıdır.

Modelimizin oluşturulmasında kullanacağınız algoritmanın projeniz için doğru algoritma olup olmadığını ;farklı algoritmaları "*lift chart*" gibi toollarda kullanarak görebilirsiniz.(Bir sonra ki adımda lift chart'dan bahsedeceğim).

4) Modelin Keşfi (Model Assessment) : Bir önceki adımımızda (Modelin Oluşturulması) farklı algoritmalara ve parametrelere göre modelimizi oluşturmaya çalıştık.Peki seçmiş olduğumuz algoritmanın bizim projemiz için en doğru algoritma olduğuna nasıl karar vereceğiz.İşte bu noktada karşılaştırma yapmak amacıyla bazı toollar karşımıza çıkıyor.

Bu toollar arasında en sık kullanılan *Lift Chart* adlı tooldur.Lift Chart ile değerlerin tahmin edilmesi için model eğitilmekte ve dataset test edilmektedir.Lift Chart değerlerin tahmin edilmesi ve olasılıklarının hesaplanması esasına dayanarak,grafiksel olarak modeli bize göstermektedir.



Modelin keşfi aşamasında sadece toolları kullanıp sonucun doğru olup olmadığını teknik insanların tek başlarına karar vermesi uygun değildir.Bu aşamada çıkan örnek sonuçlar projenin yapıldığı departmanın uzmanlarıyla tartışılıp,sonucun doğruluğuna karar verilmelidir.

Bazı durumlarda model yararlı desenler (patterns) içermeyebilir.Bunun temelde 2 tane nedeni vardır.Birincisi data ,tamamen rastgele seçilmiştir ki birçok durumda gerçek datasetler zengin bilgiler içerir.İkinci sebep ise kurulan modelde;değişkenlerin kullanım için en uygunlarından seçilmiş olmamasıdır.Bu durumla daha sık karşılaşılır,çözüm olarak data temizleme ve yeniden yapılandırma aşaması daha anlamlı değişkenler için tekrar edilir.

Veri Madenciliği birçok aşamadan oluşan bir döngü şeklinde yapıya sahip olduğundan dolayı bazı aşamalara geri dönüşler yapılabilir.

5) Raporlama: Raporlama; veri madenciliği sonuçlarını gösterebilmek için en etkili kanaldır.Bazı veri madenciliği projelerinin amacı pazarlama çalışmaları için raporlar sunmak olabilir.Hemen hemen bütün veri madenciliği toolları kullanıcısına metinsel ve grafiksel rapor çıktıları alabilme imkanını sunar.(Desenlerle ilgili yada tahminlerle ilgili olarak.)

6)Tahminleme(Prediction,Scoring): Bazı veri madenciliği projelerinin süresinin neredeyse yarısı desenlerin bulunmasıyla geçmektedir.Daha sonrasında bulunan model kullanılarak tahminleme yapılır.(Tahminleme;prediction,veri madenciliği terminolojisinde scoring olarak da geçer) .Tahminleme yapabilmemiz için eğitilmiş bir model ve kurgulanmak için hazır bir senaryoya ihtiyaç vardır.

Bankaların müşterilerine kredi vermek için yaptığı inceleme senaryosunu düşündüğümüzde,kredi riski üzerine eğitilmiş bir model vardır.Bankaya hergün binlerce kredi talebi

gelmektedir ve bu talepler risk değerlendirme modeline göre tahminler yürütülerek,potansiyel risk oluşturan başvuruları belirlenmektedir.

7)Uygulamanın Entegrasyonu : İş uygulamalarında ki gömülü veri madenciliği entegrasyonları yapılan tüm işlerin ve çalışmaların zeka kısmını oluşturmakla birlikte analiz döngüsünün de son basamağıdır.*Gartner*'a göre ;önümüzdeki yıllarda daha fazla iş uygulamasının içerisinde,gömülü veri madenciliği bileşenlerini görebileceğiz ve bu tür iş uygulamaları bizim için ayrı bir değere sahip olacaktır.

Örneğin,CRM (Customer Relationship Management,Müşteri İlşikileri Yönetimi) müşterileri segmente etmek için Veri Madenciliği özelliklerinden faydalanmaktadır.Son zamanlarda işletmelere yeni bir soluk getiren ERP (Enterprise Resource Planning ,Kurumsal Kaynak Planlama) uygulamaları ise üretim tahminleri için Veri Madenciliği özelliklerinden yararlanmaktadır.

Bir kitap alışveriş sitesini düşündüğümüzde müşterilerine gerçek zamanlı olarak kitap tavsiyelerinde bulunabiliyorsa bu, Veri Madenciliğinin bir maharetidir.İş uygulamalarının bu tip gerçek zamanlı tahminlerde bulunması,Veri Madenciliği projelerinin önemli bir aşaması olan Entegrasyon aşamasının sonunda gerçekleşmektedir

8)Modelin Yönetimi: Buraya kadar ki aşamalarda modelimizi oluşturduk,tahminlerimizi yaptık,CRM ve ERP benzeri yapılarımızla entegrasyonu sağladık.Ama her madencilik modeli bir yaşam döngüsüne sahiptir ve bazen statik bir şekilde çalışabilir , ve sık aralıklarla tekrardan eğitilmeye ihtiyaç duymayabilir.Fakat veri'nin sıkça değiştiği durumlarda tekrar eğitilmeye ihitiyaç duymaktadır.Örneğin online kitap mağazasına hergün yeni kitaplar ürün listesine dahil edilmektedir.Gelen her kitap içinde,hergün yeniden bir ilişki kurulması gerekmektedir.Bu süreçte madencilik modelleri sınırlıdır ve yeni versiyon sıklıkla bir model oluşturmaktadır.Eninde sonunda bu modelin doğruluğu test edilmesi ve yeni versiyonun oluşturulması otomatik işlemlerle tamamlanmış olmalıdır.

Veri Tabanlarında(RDMS) olduğu gibi madencilik modelleri için de en önemli yönetim konularının başında güvenlik konuları gelmektedir.Madencilik modelleri desenler içermektedir ve bu değişik sayıdaki desenin okuma,yazma,tahmin gibi haklarını farklı kullanıcı profilleri için korunmak zorundadır.Madencilik modeline erişecek sınırlı sayıdaki kullanıcıların,hakları proje yöneticisi tarafından gerektiğinde verilmeli yada görevlendirmeler bittiğinde geri alınmalıdır.

Bu iki makalemizde;Veri Madenciliği Projelerinin yaşam döngüsü'nün basamakları ve bu basamakların içerikleri hakkında yeterince bilgi sahibi olduğumuzu düşünüyorum.

Başka bir İş Zekası makalesinde görüşmek dileğiyle...

Bilgehan GÜRÜNLÜ

www.gurunlu.com
bilgehan@gurunlu.com

Kaynaklar :

Data Mining with Sql Server 2005.