Kümeleme Algoritmaları

Giriş

Bir küme temsil ettiği nesneleri en iyi şekilde ifade edecek biçimde düzenlenir. Kümeleme işleminin uygulandığı veri setindeki her bir veriye nesne adı verilir. Bu nesneler iki boyutlu düzlem üzerinde noktalarla gösterilir. Kümeleme analizi, veri indirgeme veya nesnelerin doğal sınıflarını bulma gibi çeşitli amaçlarla kullanılmaktadır. Bu alanlardan en çok gündemde olanlar örüntü tanıma, veri analizi, resim tanıma, pazarlama, metin madenciliği, doküman toplama, istatistik araştırmaları, makine öğrenimi, şehir planlama, coğrafik analizler (deprem, meteoroloji, yerleşim alanları), uzaysal veri tabanı uygulamaları, Web uygulamaları, müşteri ilişkileri yönetimi, sağlık ve biyoloji alanında yapılan araştırmalardır.

Kümelenme, bir "**denetimsiz öğrenme**" problemi olarak düşünülebilir; etiketlenmemiş verilerden oluşan bir koleksiyonda bir yapı bulmakla ilgilenir. **Kümelenme, "birbirine benzer üyeleri olan grupları, kümeler halinde düzenleme süreci**" olarak tanımlanabilir. Bu nedenle bir küme, aralarında "benzerlik" bulunan ve diğer kümelere ait nesnelere "benzemeyen" bir nesne koleksiyonudur.

Yukarıdaki şekilde verilerin 2 kümeye bölündüğü görülmektedir; benzerlik ölçütü "**mesafedir**": iki ya da daha fazla nesne eğer verili bir mesafeye göre birbirlerine yakınsa bunlar aynı kümeye aittir. Buna **mesafeye dayalı kümeleme** denir.

Bir başka kümeleme **kavramsal kümelemedir**. İki ya da daha fazla nesne -eğer nesnelerden biri tüm bu nesneler için ortak bir kavram tanımlıyorsa- aynı kümededir. Bir başka deyişle nesneler sadece benzerlik ölçümüne göre değil niteleyici kavrama uyuyorsa birlikte gruplanır.

Kümelemenin Amacı

Kümelenmenin amacı, bir grup etiketlenmemiş veride içsel gruplamayı belirlemektir.

Burada soru: neyin iyi bir kümelenme oluşturduğuna nasıl karar verileceğidir. Bir ölçüt belirlemek güçtür. Kriteri sağlaması gereken kullanıcıdır; kümelenin sonucu kullanıcının gereksinimine uymalıdır

Örneğin, iyi bir kümeleme homojen gruplar için **temsilciler** bularak sağlanabilir(veri azaltma/data reduction) ya da "doğal kümeler" bulma yoluna gidilebilir ve bilinmeyen özellikleri tanımlanabilir ("natural" data types), kullanışlı ve uygun gruplar bulma ("useful" data classes) veya olağandışı/istisnai veri nesneleri bulma (outlier detection) da olabilir.

Olası Uygulamaları

Kümeleme algoritmaları birçok alana uygulanabilir:

Pazarlamada: Müşterilerin özelliklerini ve geçmiş satın alma kayıtlarını içeren büyük bir veri tabanında benzer davranışa sahip müşterileri bulma;

Biyoloji: Bitkilerin ve hayvanların özelliklerine göre sınıflanması;

Kütüphanelerde: Kitap siparişinde;

Sigortacılıkta ortalama hasar maliyeti yüksek olan sigorta sahiplerinin belirlenmesi; dolandırıcılıkların belirlenmesi;

Şehir Planlamacılığında: konut gruplarının konut türlerine, değerlerine ve coğrafi konumlarına göre belirlenmesi;

Deprem çalışmalarında: tehlikeli bölgeleri tespit etmek için deprem merkez üslerinin gözlemlenmesi;

WWW'de doküman sınıflamak için: weblogların kümelenmesi için ve benzer erişim örüntüsü sergileyen grupları keşfetmek için.

Gereklilikler:

Bir kümeleme algoritmasının olmazsa olmazları:

Ölçeklenebilirlik;

Farklı niteliklerle/özelliklerle uğraşmak

Rasgele şekilli kümeleri keşfetmek

Girdi parametreleri belirlemek için bir alanın bilgisinin minimum gereklilikleri

Gürültü ve aykırı değerlerle başa çıkma yeteneği;

Girdi kayıtlarının sırasına duyarsızlık;

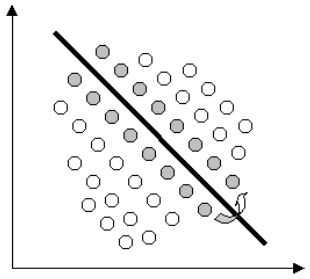
Yüksek çok boyutluluk;

Yorumlanabilirlik ve kullanılabilirlik.

Kümeleme Algoritmaları

Sınıflama:

Özel Kümeleme / Exclusive Clustering: Eğer belirli veriler, kesin bir kümeye ait ise o zaman başka bir kümeye dahil edilemez. Aşağıdaki şekilde 2 boyutlu bir düzlemde bir doğrunun ayırdığı noktalar görülmektedir. Örn: K-means algoritması.



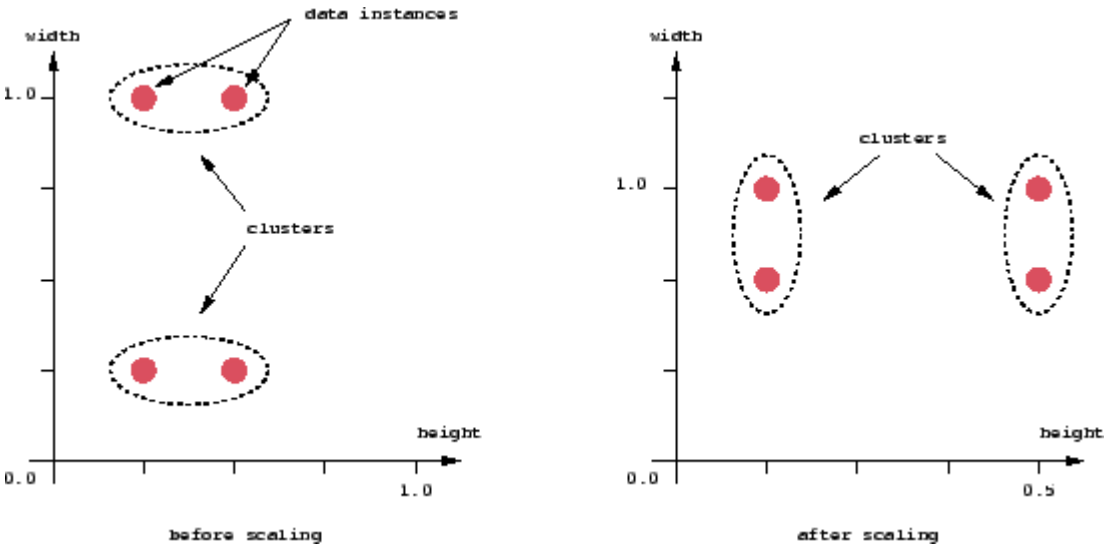
Örtüşen/ Çakışan Kümeleme / Overlapping Clustering: Burada bulanık veri setleri kullanılır. Böylelikle her nokta farklı üyelik dereceleri ile birden fazla kümeye dahil edilebilir. Bu durumda veri, uygun bir üyelik değerine sahiptir. Örn: Fuzzy C-means algoritması.

Hiyerarşik Kümeleme / Hierarchical Clustering: Hiyerarşik bir kümeleme algoritması, en yakın iki küme arasındaki birleşmeye dayanır. Başlangıç koşulu, her verinin bir küme olarak ayarlanmasıyla gerçekleşir. Birkaç tekrardan sonra, istenen son kümelere ulaşılır.

Olasılıklı Kümeleme / Probabilistic Clustering: Tamamen olasılık yaklaşımını kullanır. Örn: Mixture of Gaussian algoritması.

Mesafe Ölçümü

Kümeleme algoritmasının önemli bir bileşeni veri noktaları arasındaki mesafenin ölçülmesidir. Veri vektörlerinin bileşenlerinin tümü aynı fiziksel birimlerdeyse, basit Öklid uzaklık metriğinin benzer veri örneklerini başarıyla gruplamak için yeterli olması mümkündür. Bununla beraber bazen Öklid mesafesi yanıltıcı olabilir.



Yukarıdaki şekilde bir nesnenin genişlik ve yükseklik ölçümleri gösterilmektedir. Her iki ölçüm aynı fiziksel birimlerde(nesnelerde) yapılmasına rağmen, **farklı ölçeklendirmeler** farklı kümelenmelere yol açabilir.

Bununla birlikte, bunun sadece grafik bir sorun olmadığına dikkat edilmesi gerekir: sorun, kümeleme amacıyla veri vektörlerinin arasındaki mesafeleri ölçmek için kullanılan **matematiksel formülden** kaynaklanmaktadır: Uygun bir mesafe ölçüsü formüle edebilmek için kullanılan **farklı formüller farklı kümelenmemeler** üretebilmektedir.

K-Means

En eski kümeleme metotlarından biri olan k-means algoritmasının genel mantığı n adet veri nesnesinden oluşan bir veri setini, giriş parametresi olarak verilen k adet kümeye bölümlemektir. Amaç, gerçekleştirilen bölümleme işlemi sonunda elde edilen kümelerin, **küme içi benzerliklerinin maksimum** ve **kümeler arası benzerliklerinin minimum olmasını sağlamaktır.** Küme benzerliği, kümenin ağırlık merkezi olarak kabul edilen bir nesne(sentroid) ile kümedeki diğer nesneler arasındaki uzaklıkların ortalama değeri ile ölçülmektedir

K-means, iyi bilinen kümeleme problemini çözen en basit **denetimsiz öğrenme** algoritmalarından biridir.

Makine öğrenimi algoritmaları, insan müdahalesi olmadan verilerden öğrenebilen ve deneyimler ile geliştirebilen programlardır. Makine öğrenimi algoritmaları denetimli öğrenme ve denetimsiz öğrenme olarak kategorize edilir.

Denetimsiz Öğrenme (Unsupervised Learning): Sadece veriler vardır; onlar hakkında bilgi verilmez. Bu verilerden sonuçlar çıkarılmaya çalışılır. En baştan veriler hakkında herhangi bir bilgi verilmediği için çıkartılan sonuçların kesinlikle doğru olduğu söylenemez. Veriler değişkenler arasındaki ilişkilere dayalı olarak kümelenerek çeşitli modeller, yapılar oluşturulur. Kümeleme buna örnektir.

Sınıflandırma ve regresyon ise **denetimli öğrenme** örnekleridir**. Denetimli öğrenme**, tahmin modelleri geliştirmek için sınıflandırma ve regresyon tekniklerini kullanır.

Algoritmanın adımları:

Kümelenecek olan nesnelerin kaç gruba bölüneceğini ifade eden başlangıç sentroidleri (merkez noktalar) belirlenir. Örneğin nesneler 2 küme altında gruplanacaksa, iki sentroid var demektir; K=2 olur. Farklı konumlar farklı sonuçlara neden olduğu için bu sentroidler dikkatli bir şekilde seçilmelidir. En iyi seçenek olabildiğince birbirinden uzak olmalarına bakmaktır.

Bir sonraki adımda her bir nesne, seçilen merkez nokta ile yakınlıklarına göre ilgili kümelere atanır.

Bütün nesneler atandıktan sonra sentroidlerin pozisyonları yeniden hesaplanır. Bu adımlar sentroidde bir değişiklik olmayana kadar tekrarlanır. Bir başka deyişle Oluşan kümelerin yeni merkez noktaları o kümedeki tüm nesnelerin ortalama değeri ile değiştirilir. K-means algoritmasında her bir nesnenin merkez noktalara uzaklığını hesaplamak için bazı formüller kullanılır. Örneğin: Öklid mesafesi kümeleme analizine sıradışı olabilecek yeni nesnelerin eklenmesinden etkilenmez. Ancak boyutlar arasındaki ölçek farklılıkları Öklid mesafesini önemli ölçüde etkilemektedir.

$$distance(x, y)\ \{\ (x_i \text{-----} y_i\ )^2\ \}^{1/2}$$

K-means algoritması, hata parametresinin değerini minimum yapmak için büyük kümeleri bölerek mümkün olduğunca birbirinden ayrık ve kendi içinde sıkışık kümeler bulmaya çalışmaktadır.

Bkz Document Similarity and Clustering in RapidMiner

http://www.youtube.com/watch?v=ToxzfYECxOU

Kümeleme, bir örüntü dermesinin **vektör ölçümleriyle** ya da çok boyutlu uzayda noktalar şeklinde **kümelerin benzerliğine** göre düzenlenmesidir. Bir küme içindeki örüntüler, farklı bir küme içindeki örüntülere oranla birbirleriyle daha benzerdir. **Kümeleme**(denetimsiz sınıflama) ile **ayrım analizi**(denetimli sınıflama) arasındaki farkı bilmek gerekir.

Fuzzy c-means (FCM) bir kümeleme yöntemidir. Bir nesne birden fazla kümede yer alabilir. Bu yöntem sıklıkla örüntü tanımada kullanılır. Nesneler bir **üyelik fonksiyonu** aracılığıyla bir kümeye bağlanır, bu da bu algoritmanın bulanık davranışını temsil eder.

Sparck Jones, rastgele iki sözcüğün aynı belgede birlikte ortaya çıkma sıklığına bağlı olarak anahtar kelimeler arasındaki ilişki ölçülerini kullanarak bu çalışmayı sürdürmüştür....

Hiyerarşik kümeleme yöntemi: Birçok hiyerarşik kümeleme yöntemi vardır, bunlardan birkaçı: tam bağlantı, ortalama bağlantı vb'dir. Hiyerarşik yöntemlerin belge kümelemeye uygun olduğu göz önüne alındığında şu soru ortaya çıkar: Hangi yöntem? Cevap, Jardine ve Sibson'da belirli koşullar altında tek kabul edilebilir hiyerarşik küme yöntemi olan tek bağlantı(single-link)dir.