Kümeleme Algoritmaları

Tahir Emre KALAYCI

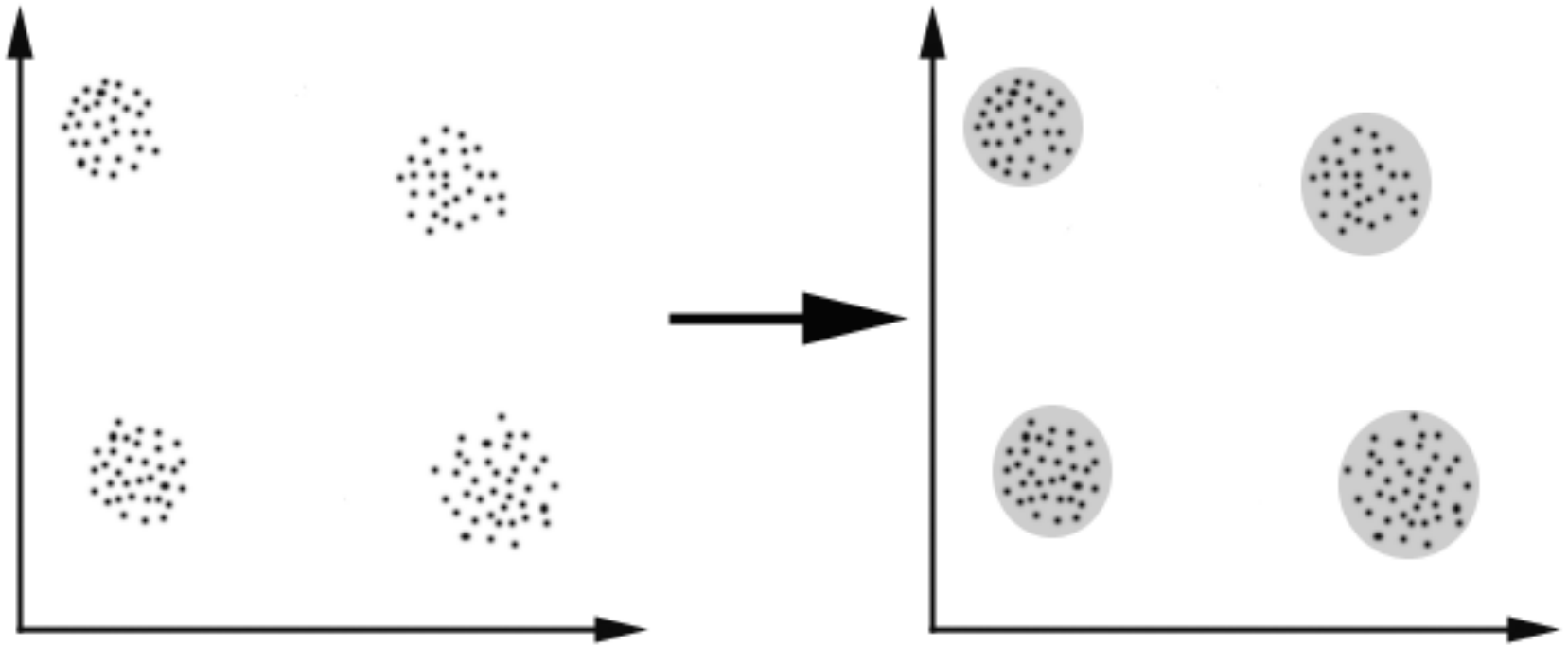
2010

Gündem

Kümeleme nedir?

- En önemli gözetimsiz öğrenme (unsupervised learning) problemi olarak değerlendirilmektedir
- Bu türdeki diğer problemler gibi etiketsiz veri kolleksiyonları için bir yapı bulmakla ilgilenmektedir
- Kümelemeyi basitçe "nesneleri benzerliklerine göre belli gruplar halinde düzenleme işi" olarak tanımlayabiliriz
- Bilgisayarlar yardımıyla görüntüler oluşturulmasıdır
- Küme bu tanımdan yola çıkarsak aynı kümedeki nesnelerle "benzer", diğer kümelere ait nesnelerle "benzer olmayan" nesnelerin kolleksiyonu anlamına gelmiş olur

Kümeleme nedir?



Kümeleme nedir?

- İki türdür : uzaklık tabanlı kümeleme, kavramsal kümeleme
- Kümelemenin hedefi: Kümelemenin hedefi elimizdeki etiketsiz verilerin içsel gruplarını belirlemektir
- Peki iyi kümelemeyi oluşturanlara nasıl karar veririz

Uygulama alanları

- Pazarlama: müşteri özelliklerini ve geçmiş alışverişlerini barındıran geniş bir veritabanından benzer davranışları sergileyen müşterilerin gruplandırılması
- Biyoloji: bitkilerin ve havyanların verilen özelliklerine göre sınıflandırılması
- Kütüphaneler: kitapların düzenlenmesi
- Sigortacılık: motor sigorta poliçe sahiplerinin ortalama maliyetlerinin hesaplanması için, sahtekarlıkların belirlenmesi için gruplandırılması
- Şehir plancılığı: tiplerine, değerlerine ve coğrafik konumlarında göre evlerin gruplandırılması
- Deprem araştırmaları: gözlenen deprem merkezlerinin gruplanarak tehlikeli bölgelerin belirlenmesi
- WWW: belge sınıflandırma, benzer erişim desen gruplarının tespiti için web kaydı verilerinin kümelenmesi

Gereksinimler

Bir kümeleme algoritmasının barındırması gereken özellikler şu şekildedir:

- ölçeklenebilirlik
- farklı tipteki özellikleri destekleme
- farklı büyüklüklerdeki kümeleri belirleyebilme
- girdi parametrelerini belirlemeye yönelik alan bilgisi için minimal gereksinimler
- sapkınlık ve gürültüyle başa çıkabilme
- girdi kayıtlarını düzenlemek için duyarsızlık
- yüksek boyutluluk
- yorumlanabilirlik ve kullanışlılık

Problemler

Kümelemenin bir çok problemi vardır:

- var olan kümeleme teknikleri tüm gereksinimleri yeterince karşılamıyor
- çok sayıda boyutla ve çok sayıda veri öğesiyle uğraşmak zaman karmaşıklığından dolayı problematiktir
- uzaklık tabanlı kümelemeler için yöntemin verimliliği "uzaklık"ın tanımına bağlıdır
- eğer açık bir uzaklık tanımı yoksa bizim "tanım"lamamız gerekiyor, ki bu genellikle kolay değildir, özellikle çok boyutlu uzaylarda
- kümeleme algoritmasının sonucu farklı şekillerde yorumlanabilir

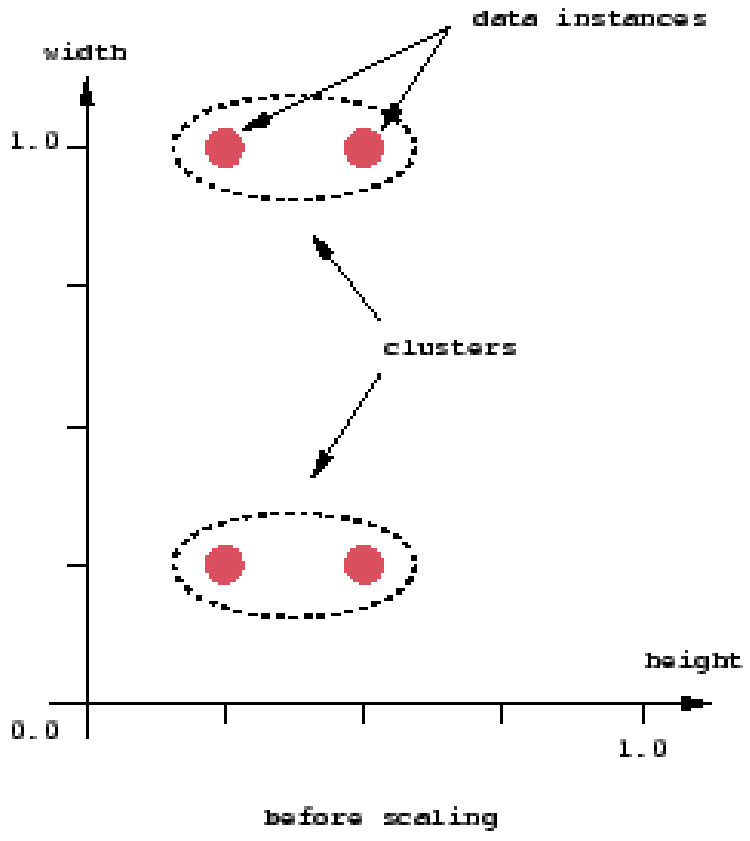
Sınıflandırma

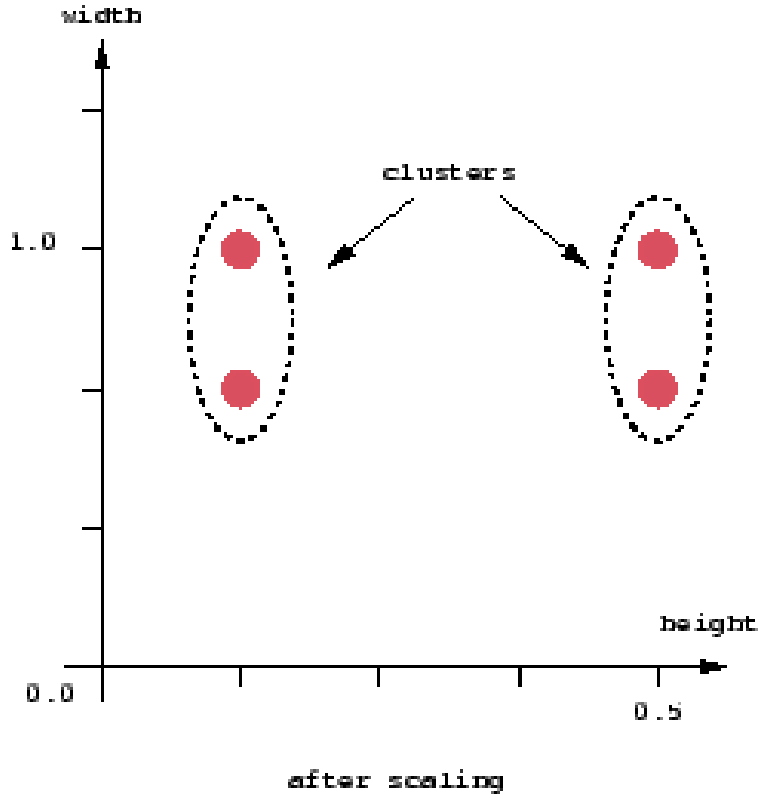
- Dışlayan (exlusive) kümeleme: eğer bir veri bir kümeye aitse başka bir kümeye ait olamaz (K-means)
- Örtüşüm (overlapping) kümeleme: verinin kümelenmesi için bulanık kümeler kullanılıyor, böylece her bir nokta iki veya daha fazla kümeye farklı derecelerde üye olabilirler (C-means)
- Sıradüzensel kümeleme: en yakın iki kümenin birleşimine dayanmaktadır (sıradüzensel - hierarhical)
- Olasılıksal kümeleme: tamamen olasılıksal yaklaşımı kullanır (Gaussianların karışımı - mix of gaussians)

Uzaklık ölçütü

- Kümeleme algoritmalarının önemli bir bileşeni veri noktaları arasındaki uzaklık ölçütüdür
- Eğer veri örneği vektör bileşenlerinin hepsi de aynı fiziksel birimde ise basit Öklit uzaklığı ölçütü benzer veri örneklerini gruplamak için yeterli olacaktır
- Ancak, bu durumda bile Öklit uzaklığı yanıltıcı olabilir
- Her iki ölçümde aynı fiziksel birimde alınmış olsa da, karar ölçeklemeye uygun olarak yapılmalıdır. Farklı ölçeklemeler farklı kümelemelere neden olacaktır

Uzaklık ölçütü





Minkowski ölçütü

- Yüksek boyutlu veri için popüler bir ölçüt Minkowski ölçütüdür
- $d_p(x_i, x_j) = \left(\sum_{K=1}^{d} |x_{ik} - x_{jk}|^p \right)^{\frac{1}{p}}$
- d verinin boyutluluğudur. Öklit uzaklığı p=2 olan özel bir durumdur, Manhattan ölçütü için p=1'dir.
- Herşeye rağmen verilen bir uygulama için ölçüt seçmeye yönelik herhangi bir teorik rehber yoktur.
- Sıklıkla karşılaşılan başka bir durum veri özellik vektörlerinin karşılaştırılabilir olmamasıdır.
- Bileşenler sürekli değişkenler (uzunluk gibi) değil de kategoriler olabilir (haftanın günleri gibi).
- Bu durumda uygun bir ölçüt için formül üretmede alan bilgisi mutlaka kullanılmalıdır.

Kaynaklar

- Matteo Matteucci, “A Tutorial on Clustering Algorithms”, http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/
- Tariq Rashid, “Clustering” http://www.cs.bris.ac.uk/home/tr1690/documentation/fuzzy_cluster
- Osmar R. Zaïane, “Principles of Knowledge Discovery in Databases - Chapter 8: Data Clustering” http://www.cs.ualberta.ca/ za-iane/courses/cmput690/slides/Chapter8/index.html