BINUS UNIVERSITY ONLINE LEARNING

People
Innovation
Excellence

LECTURE NOTES

COMP8014

Knowledge Data Discovery

Cluster Analysis: Advanced Methods

LEARNING OUTCOMES

LO4: Mahasiswa diharapkan mampu untuk menentukan metode data mining yang sesuai dengan sifat kealamiahan permasalahan yang dihadapi.

LO5: Mahasiswa diharapkan mampu untuk mengimplementasikan metode data mining.

LO6: Mahasiswa mampu mengevaluasi aplikasi dan trend Data Mining.

OUTLINE MATERI :

1. Model-based clustering.
2. Fuzzy Clustering
3. Probability model-based clustering
4. Mixture Models and EM Algorithm
5. Self-Organizing Map (SOM)
6. Clustering high-dimensional data
7. Topik Tambahan

ISI MATERI

Konsep dasar fuzzy:

- Fungsi fuzzy untuk model kesamaan ukuran
- Pada pengelompokan non-fuzzy atau hard yang dibahas sejauh ini, data dibagi menjadi cluster yang tajam, dimana setiap objek data ditugaskan ke satu cluster dengan tepat.
- Teori himpunan fuzzy yang diajukan oleh Lotfi Zadeh pada tahun 1965 memberi gambaran ketidakpastian kepemilikan yang digambarkan oleh fungsi keanggotaan.
- Dalam pengelompokan fuzzy, sebuah titik data dapat dimiliki lebih dari satu cluster sesuai dengan nilai keanggotaan fuzzy dimana titik data termasuk dalam kelompok yang berbeda.

Konsep dasar Fuzzy Clustering:

- Asumsikan setiap objek data milik cluster K menurut keanggotaan fuzzy sebagai model
- Versi soft K-means clustering
- Tentukan fungsi keanggotaan fuzzy
- Perlu mencari matriks keanggotaan fuzzy cluster
- Proses pembelajaran adalah meminimalkan nilai fungsi objektif
- Prosedur dan contoh dari Fuzzy Clustering dapat dilihat di slide ppt.

Algorithm Fuzzy Clustering:

- Algoritma Fuzzy c-Means (FCM)
    - FCM lebih baik daripada algoritma k-Means untuk menghindari minima lokal, FCM masih bisa bertemu dengan minimum lokal dari kriteria kesalahan kuadrat.
- Perpanjangan algoritma FCM:
    - Algoritma menggunakan ukuran jarak adaptif:
    - Algoritma Gustafson-Kessel (Gustafson dan Kessel, 1979), algoritma estimasi likelihood maksimum fuzzy (Gath and Geva, 1989).
    - Algoritma berdasarkan prototipe hyperplanar atau fungsional, atau prototip yang didefinisikan oleh fungsi:
        - Fuzzy c-varietas (Bezdek, 1981), fuzzy c-elliptotypes (Bezdek, et al., 1981), model regresi fuzzy (Hathaway & Bezdek, 1993).
- Algorithm dan contoh aplkasi dapat dilihat di slide ppt.

Probabilitic Model-Based Clustering:

- Kerangka kerja kerangka berbasis model probabilistik mengasumsikan:
    - Kategori tersembunyi adalah distribusi di atas ruang data, yang dapat diwakili secara matematis dengan menggunakan fungsi kepadatan probabilitas.
    - Data telah dihasilkan dari distribusi probabilitas K
    - Data yang bisa kita gambarkan dengan mencari model statistik yang paling sesuai dengan data.
- Proses statistik melibatkan
    - menentukan model statistik, mis. Model Campuran Gaussian, dan memperkirakan parameter distribusi model dari data
- Algoritma EM

Mixture Model:

- Model campuran melihat data sebagai seperangkat observasi (objek yang diamati) dari gabungan distribusi probabilitas secara independen.
- Setiap distribusi sesuai dengan cluster dan parameter masing-masing distribusi memberikan deskripsi cluster yang sesuai, biasanya: pusat dan penyebaran cluster.
- Tugas keluar: menyimpulkan satu set distribusi M yang kemungkinan besar menghasilkan dataset D dengan menggunakan proses pembangkitan data di atas
- Prosedur dari Gaussian mixture model dapat dilihat di slide ppt.

EM Algorithma:

- Pada situasi umum kita tidak tahu titik mana yang dihasilkan oleh distribusi mana.
- Temuan probabilitas maksimum untuk memperkirakan parameter distribusi adalah masalah optimasi kendala non linier yang sulit dipecahkan.
- Algoritma EM adalah kerangka kerja untuk mendekati kemungkinan maksimum untuk memperkirakan parameter distribusi.
- Algoritma EM adalah metode berulang, setiap iterasi EM terdiri dari dua tahap: Expectation (E-step) dan Maximization (M-step).
- Algoritma EM dapat dilihat di slide ppt.

Keunggulan dan kekurangan metode Gaussian Mixture:

<u>Kekuatan</u>:

- Model campuran lebih umum daripada partisi dan fuzzy clustering
- Cluster dapat dicirikan oleh sejumlah kecil parameter
- Hasilnya dapat memenuhi asumsi statistik model generative

Kelemahan:

- Konvergen ke optimal lokal
- Komputasional mahal jika jumlah distribusinya besar, atau kumpulan data hanya berisi sedikit sekali titik data yang teramati
- Perlu kumpulan data yang besar
- Sulit memperkirakan jumlah cluster

SOM:

- Peta fitur pengorganisasian sendiri (SOM) atau self-organizing feature map (SOFM) adalah teknik clustering dan teknik visualisasi data berdasarkan sudut pandang jaringan syaraf tiruan yang dilatih menggunakan pembelajaran kompetitif.
- SOM mampu memetakan data berdimensi tinggi dalam dimensi yang lebih rendah, biasanya dua dimensi, yang membuat SOM berguna untuk memvisualisasikan tampilan berdimensi rendah dari data berdimensi tinggi.
- Peta 2D berguna dalam
    - Mengidentifikasi kelompok
    - Menganalisis dan menemukan pola pada ruang input
    - Mendeteksi korelasi non linier antar fitur
- SOM beroperasi dalam dua mode: pelatihan dan pemetaan:
    - Pelatihan membangun peta dengan menggunakan contoh masukan (proses kompetitif, yang disebut juga kuantisasi vektor),
    - Pemetaan mengklasifikasikan secara otomatis vektor masukan baru.
- SOM terdiri dari node atau neuron yang terkait dengan vektor bobot dengan dimensi yang sama dengan data masukan, dan posisi di ruang peta.
- Simpul biasanya disusun dalam grid empat dimensi persegi panjang atau kisi grid heksagonal.
- Prosedur untuk menempatkan vektor dari ruang data ke peta adalah untuk menemukan simpul dengan vektor bobot terdekat ke vektor ruang data.
- Algortima SOM dapat dilihat di slide ppt.

Topik Tambahan:

- Semi Supervised Learning
- Klustering di dimensi tinggi

SIMPULAN

1. Kami sudah mempelajari beberapa algoritma clustering kemajuan Pengelompokan fuzzy:
    a. FCM
    b. Pengelompokan berbasis probabilitas: Model Campuran Gaussian dan Algoritma EM
    c. Self-Organizing Map (SOM)
2. Clustering data berdimensi tinggi memiliki banyak aplikasi di dunia nyata.
3. Beberapa algoritma clustering rumit dan ada banyak parameter sehingga parameter tuningnya memakan waktu lama.

DAFTAR PUSTAKA

1. Han, J., Kamber, M., & Pei, Y. (2006). “Data Mining: Concepts and Technique”. Edisi 3. Morgan Kaufman. San Francisco
2. Tan, P.N., Steinbach, M., & Kumar, V. (2006). “Introduction to Data Mining”. Addison-Wesley. Michigan
3. Witten, I. H., & Frank, E. (2005). “Data Mining : Practical Machine Learning Tools and Techniques”. Second edition. Morgan Kaufmann. San Francisco