Pendahuluan

Data mining adalah suatu istilah yang digunakan untuk menguraikan penemuan pengetahuandi dalam database. Data mining adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan machine learning untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai database besar.

Definisi umum dari data mining itu sendiri adalah proses pencarian pola-pola yang tersembunyi (hidden patern) berupa pengetahuan (knowledge) yang tidak diketahui sebelumnya dari suatu sekumpulan data yang mana data tersebut dapat berada di dalam database, data werehouse, atau media penyimpanan informasi yang lain.

Pendahuluan

Hal penting yang terkait di dalam data mining adalah:

1. Data mining merupakan suatu proses otomatis terhadap data yang sudah ada.

2. Data yang akan diproses berupa data yang sangat besar.

3. Tujuan data mining adalah mendapatkan hubungan atau pola yang mungkin memberikan indikasi yang bermanfaat.

Data mining dilakukan dengan tool khusus, yang mengeksekusi operasi data mining yang telah didefinisikan berdasarkan model analisis. Data mining merupakan proses analisis terhadap data dengan penekanan menemukan informasi yang tersembunyi pada sejumlah data besar yang disimpan ketika menjalankan bisnis perusahaan.

Pendahuluan

Kemajuan luar biasa yang terus berlanjut dalam bidang data mining didorong oleh beberapa faktor antara lain:

1. Pertumbuhan yang cepat dalam kumpulan data.
2. Penyimpanan data dalam data warehouse, sehingga seluruh perusahaan memiliki akses ke dalam database yang andal.
3. Adanya peningkataan akses data melalui navigasi web dan internet.
4. Tekanan kompetisi bisnis untuk meningkatkan penguasaan pasar dalam globalisasi ekonomi.
5. Perkembangan teknologi perangkat lunak untuk data mining (ketersediaan teknologi.
6. Perkembangan yang hebat dalam kemampuan komputasi dan pengembangan kapasitas media penyimpanan.

Pendahuluan

Istilah data mining dan knowledge discovery in databases (KDD) sering kali digunakan secara bergantian untuk menjelaskan proses penggalian informasi tersembunyi dalam suatu basis data yang besar. Sebenarnya kedua istilah tersebut memiliki konsep yang berbeda, tetapi berkaitan satu sama lainnya. Salah satu tahapan dalam keseluruhan proses KDD adalah data mining.

Pendahuluan

Proses KDD itu ada 5 tahapan yang dilakukan secara terurut, yaitu:

1. Data selection

Pemilihan (seleksi) data dari sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi dalam KDD dimulai. Data hasil seleksi yang digunakan untuk proses data mining, disimpan dalam suatu berkas, terpisah dari basis data operasional.

2. Pre-processing / cleaning

Sebelum proses data mining dapat dilaksanakan, perlu dilakukan proses cleaning pada data yang menjadi fokus KDD. Proses cleaning mencakup antara lain membuang duplikasi data, memeriksa data yang inkonsisten, dan memperbaiki kesalahan pada data.

Pendahuluan

3. Transformation

Coding adalah proses transformasi pada data yang telah dipilih, sehingga data tersebut sesuai untuk proses data mining. Proses coding dalam KDD merupakan proses kreatif dan sangat tergantung pada jenis atau pola informasi yang akan dicari dalam basis data.

4. Data mining

Data mining adalah proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu. Teknik, metode, atau algoritma dalam data mining sangat bervariasi. Pemilihan metode atau algoritma yang tepat sangat bergantung pada tujuan dan proses KDD secara keselurahan.

Pendahuluan

5. Interpretation / evalution

Pola informasi yang dihasilkan dari proses data mining perlu ditampilkan dalam bentuk yang mudah dimengerti oleh pihak yang berkepentingan. Tahap ini merupakan bagaian dari proses KDD yang disebut interpretation. Tahap ini mencakup pemeriksaan apakah pola atau informasi yang ditemukanbertentangan dengan fakta atau hipotesis yang ada sebelumnya.

Pengelompokan Data Mining

Data mining dibagi menjadi beberapa kelompok berdasarkan tugas yang dapat dilakukan, yaitu:

1. Deskripsi

Terkadang peneliti dan analis secara sederhana ingin mencoba mencari cara untuk menggambarkan pola dan kecendrungan yang terdapat dalam data. Sebagai contoh, petugas pengumpul suara mungkin tidak menemukan keterangan atau fakta bahwa siapa yang tidak cukup profesional akan sedikit didukung dalam pemilihan presiden. Deskripsi dari pola dan kecendrungan sering memberikan kemungkinan penjelasan untuk suatu pola atau kecendrungan.

Pengelompokan Data Mining

2. Estimasi

Estimasi hampir sama dengan klasifikasi, kecuali variabel target estimasi lebih ke arah numerik dari pada ke arah kategori. Model dibangun dengan record lengkap menyediakan nilai dari variabel target sebagai nilai prediksi. Selanjutnya, pada peninjauan berikutnya estimasi nilai dari variabel target dibuat berdasarkan nilai variabel prediksi.

3. Prediksi

Prediksi hampir sama dengan klasifikasi dan estimasi, kecuali bahwa dalam prediksi nilai dari hasil akan ada di masa mendatang. Beberapa metode dan teknik yang digunakan dalam klasifikasi dan estimasi dapat pula digunakan (untuk keadaan yang tepat) untuk prediksi.

Pengelompokan Data Mining

4. Klasifikasi

Dalam klasifikasi, terdapat target variabel kategori. Sebagai contoh, penggolongan pendapatan dapat dipisahkan dalam tiga kategori, yaitu pendapatan tinggi, pendapatan sedang, dan pendapatan rendah.

5. Pengklusteran

Pengklusteran merupakan pengelompokan record, pengamatan, atau memperhatikan dan membentuk kelas objek-objek yang memiliki kemiripan. Kluster adalah kumpulan record yang memiliki kemiripan satu dengan yang lainnya dan memiliki ketidakmiripan dengan record-record dalam kluster lain.

Pengelompokan Data Mining

Pengklusteran berbeda dengan klasifikasi yaitu tidak adanya variabel target dalam pengklusteran. Pengklusteran tidak mencoba untuk melakukan klasifikasi, mengestimasi, atau memprediksi nilai dari variabel target. Akan tetapi, algoritma pengklusteran mencoba untuk melakukan pembagian terhadap keseluruhan data menjadi kelompok-kelompok yang memiliki kemiripan record dalam satu kelompok akan bernilai maksimal, sedangkan kemiripan dengan record dalam kelompok lain akan bernilai minimal.

6. Asosiasi

Tugas asosiasi dalam data mining adalah menemukan atribut yang muncul dalam satu waktu. Dalam dunia bisnis lebih umum disebut analisis keranjang belanja (market basket analysis).

Arsitektur Sistem Data Mining

Arsitektur utama dari sistem data mining, pada umumnya terdiri dari beberapa komponen sebagai berikut:

1. Database, data warehouse, atau media penyimpanan informasi, terdiri dari satu atau beberapa database, data warehouse, atau data dalam bentuk lain. Pembersihan data dan integrasi data dilakukan terhadap data tersebut.
2. Database, data warehose, bertanggung jawab terhadap pencarian data yang relevan sesuai dengan yang diinginkan pengguna atau user.
3. Basis pengetahuan (Knowledge Base), merupakan basis pengetahuan yang digunakan sebagai panduan dalam pencarian pola.

Arsitektur Sistem Data Mining

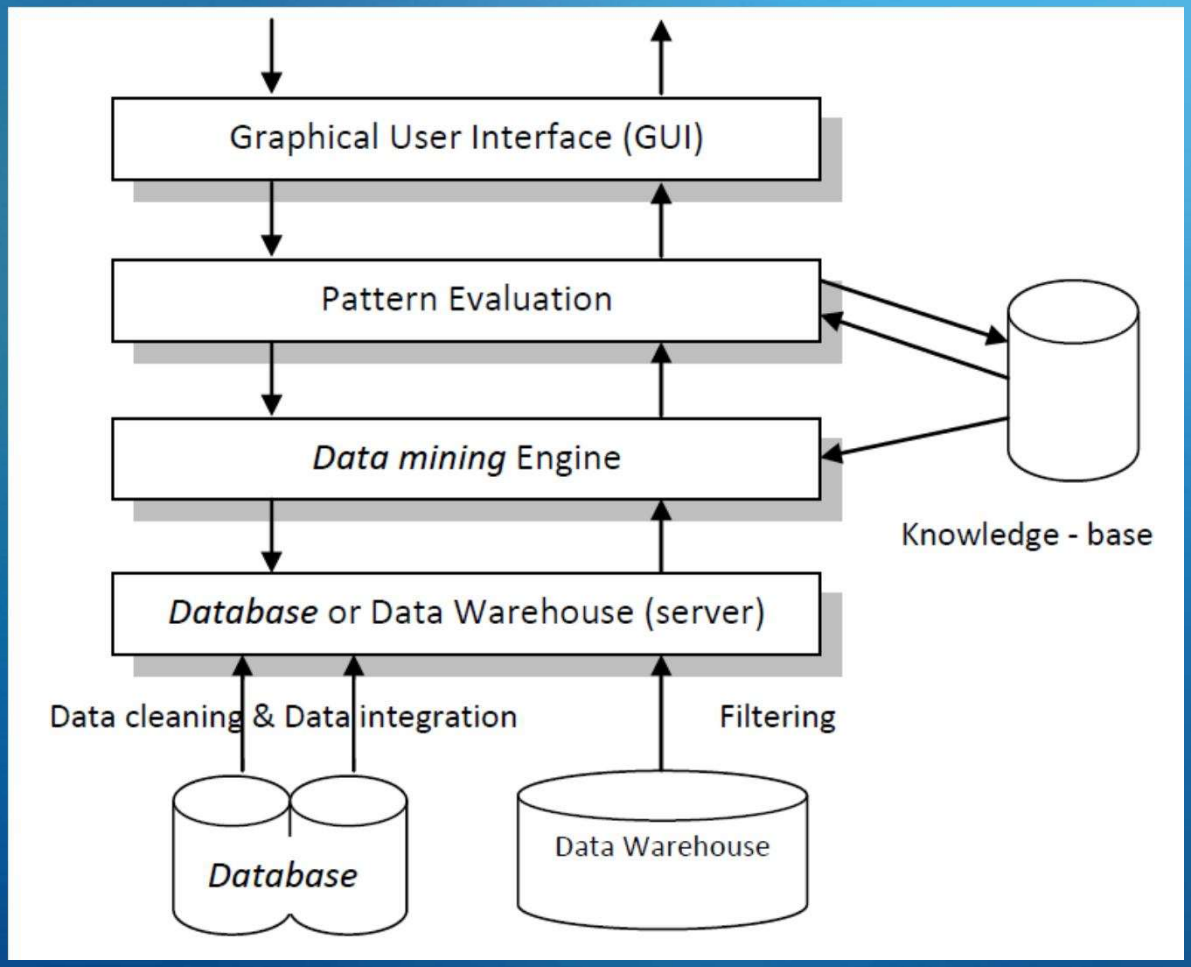
4. Data mining engine, merupakan bagaian penting dari sistem dan idealnya terdiri dari kumpulan modul-modul fungsi yang digunakan dalam proses karakteristik (characterization), klasifikasi (clasiffication), dan analisis kluster (cluster analysis). Dan merupakan bagian dari software yang menjalankan program berdasarkan algoritma yang ada.

5. Evaluasi pola (pattern evaluation), komponen ini pada umumnya berinteraksi dengan modul-modul data mining. Dan bagian dari software yang berfungsi untuk menemukan pattern atau pola-pola yang terdapat dalam database yang diolah sehingga nantinya proses data mining dapat menemukan knowledge yang sesuai.

Arsitektur Sistem Data Mining

6. Antar muka (Graphical user interface), merupakan modul komunikasi antara pengguna atau user dengan sistem yang memungkinkan pengguna berinteraksi dengan sistem untuk menentukan proses data mining itu sendiri.

Arsitektur Sistem Data Mining



Contoh Program

Untuk menentukan bermain tenis atau tidak, kriteria yang diperlukan meliputi:

-Cuaca

-Angin

-Kelembaban

-Temperatur udara

•Salah satu atribut merupakan data solusi per item data yang disebut target atribut -> misalnya atribut “play” degan nilai “main” atau “tidak main”

•Atribut memiliki nilai-nilai yang dinamakan “instance”

•Misalkan atribut “Cuaca” memiliki instance -> cerah, berawan, dan hujan.

Contoh Program

No	Outlook	Temperature	Humidity	Windy	Play
1	Sunny	Hot	High	FALSE	No
2	Sunny	Hot	High	TRUE	No
3	Cloudy	Hot	High	FALSE	Yes
4	Rainy	Mild	High	FALSE	Yes
5	Rainy	Cool	Normal	FALSE	Yes
6	Rainy	Cool	Normal	TRUE	Yes
7	Cloudy	Cool	Normal	TRUE	Yes
8	Sunny	Mild	High	FALSE	No
9	Sunny	Cool	Normal	FALSE	Yes
10	Rainy	Mild	Normal	FALSE	Yes
11	Sunny	Mild	Normal	TRUE	Yes
12	Cloudy	Mild	High	TRUE	Yes
13	Cloudy	Hot	Normal	FALSE	Yes
14	Rainy	Mild	High	TRUE	No

Contoh Program

Berdasakan tabel diatas akan dibuat tabel keputusan untuk menentukan main tenis atau tidak dengan melihat keadaan Outlook (cuaca), Temperatur, Humidity (kelembaban), dan windy (keadaan angin).

Contoh Program

```
<?php
class C45{
protected $data;
protected $attributes;
protected $target;
protected $rules;
protected $finalRules;
protected $hasilPrediksi;
public function setData(array $data)
{$this->data = $data;
return $this;
}
```

Contoh Program

```
$this->attributes = $attributes;
return $this;
}
protected function getTarget()
{
$target = [];
foreach($this->data as $item) {
$target[] = $item[count($item) -1];
}
return $target;
}
```

Contoh Program

```
public function _hitung(array $data, array $attributes, $base = null, $kasus = null)
{
// HITUNG JUMLAH DATA
$jumlah_data = count($data);
// MENGAMBIL DATA KOLOM TARGET
$kolom_target = [];
foreach($data as $item) {
$kolom_target[] = $item[count($item)-1];
}
```

Contoh Program

```
foreach(array_count_values($kolom_target) as $t) {
$entropy_total = $entropy_total - $t/$jumlah_data * log($t/$jumlah_data, 2);
}
/**
* UNTUK TIAP ATRIBUT:
* - MENGHITUNG ENTROPY TIAP KASUS
* - MENGHITUNG GAIN
*/
```

Contoh Program

```
foreach(array_count_values($kolom_target) as $t) {
$entropy_total = $entropy_total - $t/$jumlah_data * log($t/$jumlah_data, 2);
}
/**
* UNTUK TIAP ATRIBUT:
* - MENGHITUNG ENTROPY TIAP KASUS
* - MENGHITUNG GAIN
*/
```

Contoh Program

Untuk Penggunaan pertama anda harus memiliki data yang akan di-training dan data yang akan di-testing. Data set yang digunakan ialah mengenai apakah akan bermasil tenis atau tidak berdasarkan parameter dan data yang telah ada.

Hasil:

YES

- if HUMADITY == HIGH
  - if OUTLOOK == SUNNY
    return NO
  - if OUTLOOK == CLOUDY
    return YES
  - if OUTLOOK == RAINY
    - if WINDY == FALSE
      return YES
    - if WINDY == TRUE
      return NO
- if HUMADITY == NORMAL
  return YES

Daftar Pustaka

- Sitorus Lamhot, Algoritma dan Pemrograman, Andi, 2010
- Febriana Henny, Perdana Agus, Sulistianingsih Indri, Belajar algoritma dan pemrograman C++, Deepublish, 2010.