Pertemuan 8, 9, 10

Teknik-teknik Data Mining

Outline

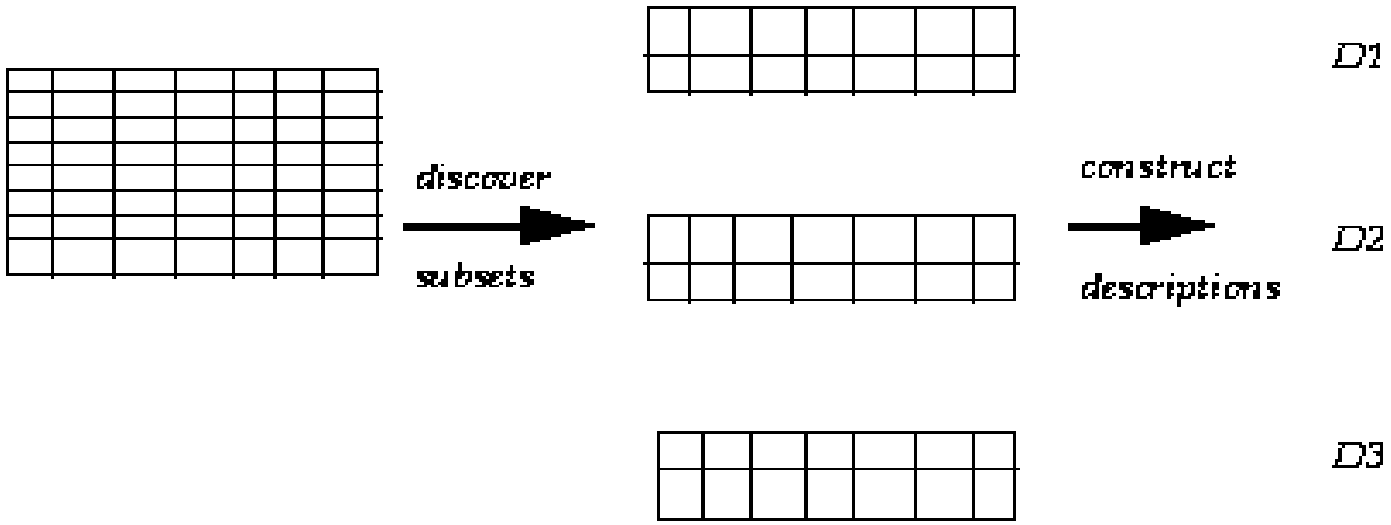
Teknik-teknik data mining terdiri dari :

- Analisis cluster
- Induksi (pohon keputusan dan aturan induksi)
- Jaringan syaraf buatan (Neural Network)
- Online Analytical Processing (OLAP)
- Visualisasi data

Analisis Cluster

- Dalam lingkungan 'unsupervised learning', sistem harus mendapatkan klas2nya sendiri dan ini dilakukan dengan meng-cluster data dalam database seperti tergambar pada gambar 1.
- Langkah pertama adalah dengan mendapatkan subset2 dari objek2 yang terhubung, kemudian mencari deskripsinya cth, D1, D2, D3, dst., yang menggambarkan masing2 subset.

Gambar 1. Perolehan cluster dan deskripsi pada database



Analisis Cluster (Lanjutan)

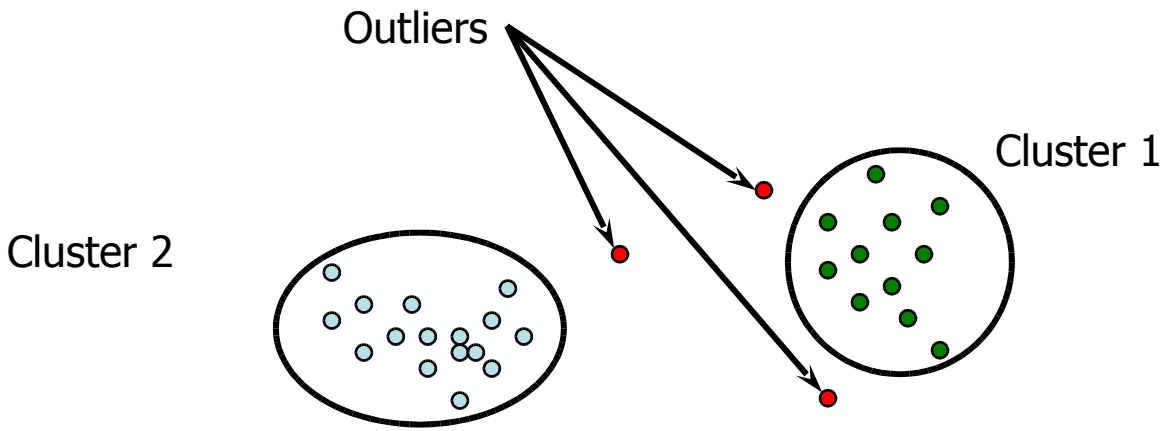
- Clustering dan segmentasi sebenarnya mempartisi database, karena itu setiap partisi atau group adalah sama menurut kriteria atau metrik tertentu. Jika pengukuran kesamaan tersedia, maka terdapat sejumlah teknik untuk membentuk cluster.

- Kebanyakan aplikasi2 data mining menggunakan clusteing menurut similarity (kesamaan), contohnya segmentasi basis klien. Clustering menurut optimasi dari sekumpulan fungsi-fungsi digunakan pada analisis data, misalnya ketika mensetting tarif asuransi klien dapat disegmentasi menurut sejumlah parameter.

- Contoh aplikasi :
  - Perangkat 'stand-alone' : explore data distribution
  - Langkah preprocessing untuk algoritma lain
  - Pengenalan pola, analisis data spasial, pengenalan citra, market research, WWW, …
    - clustering dokumen2
    - clustering data log web untuk mendapatkan group dengan pola akses yang sama

Apa itu Clustering ?

- Penggelompokkan data ke cluster2
  - Data yang sama satu sama lain berada pada cluster yang sama
  - Yang tidak sama berada pada cluster lain
  - 'Unsupervised learning': klas2 yang belum ditentukan



Gambar 2. Clustering

Clustering Yang Baik

- Intraclass similarity (Kesamaan di dalam klas) yang tinggi dan interclass similarity (kesamaan antar klas) yang rendah
  - Bergantung pada pengukuran kesamaan
- Kemampuan untuk memdapatkan beberapa atau semua pola yang tersembunyi

Kebutuhan Clustering

- Scalability
- Kemampuan mengerjakan atribut2 dari berbagai tipe
- Penemuan clusters dengan bentuk yang tidak tentu
- Kebutuhan minimal untuk pengetahuan domain untuk menentukan parameter input
- Dapat menerima noise dan outlier
- Tidak mengindahkan susunan record dari input
- Dimensi yang tinggi
- Menyatu dengan batasan yang dispesifikasikan oleh user
- Interpretability and usability

Tipe-tipe Data pada Clustering

- Variabel2 berskala interval
- Variabel biner
- Variabel nominal, ordinal dan rasio
- Variable2 dari berbagai tipe variabel

Kategori Pendekatan Clustering

- Algoritma Partisi
  - Mempartisi objek2 ke dalam k cluster
  - Realokasi objek2 secara iteratif untuk memperbaiki clustering
- Algoritma Hirarkis
  - Agglomerative: setiap objek merupakan cluster, gabungan dari cluster2 membentuk cluster yang besar
  - Divisive: semua objek berada dalam suatu cluster, pembagian cluster tsb membentuk cluster2 yang kecil
- Metode berbasis densitas
  - Berbasis koneksitas dan fungsi densitas
  - Noise disaring, kemudian temukan cluster2 dalam bentuk sembarang
- Metode berbasis grid
  - Kuantisasi ruang objek ke dalam struktur grid
- Berbasis Model
  - Gunakan model untuk menemukan keadaan data yang baik

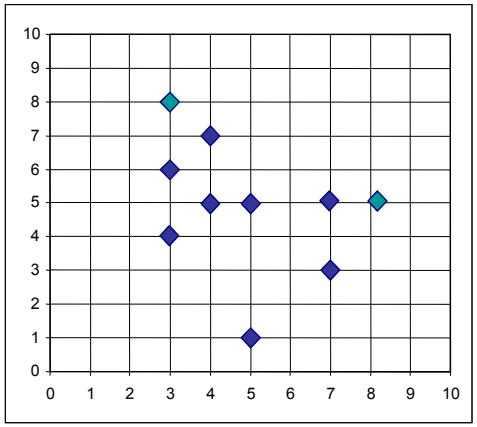
Algoritma Partisi : Konsep Dasar

- Partisi n objek ke dalam k cluster
  - Optimasi kriteria partisi yang dipilih
- Global optimal: dicoba semua partisi
  - (k^n-$(k-1)^n$-…-1) partisi yang mungkin
- Metode heuristik : k-means dan k-medoids
  - K-means: cluster direpresentasikan oleh pusat
  - K-medoids or PAM (partition around medoids): setiap cluster direpresentasikan oleh salah satu objek pada cluster

K-means

- Pilih k objek sembarang sebagai inisial pusat cluster
- Sampai tidak ada perubahan, kerjakan
  - Tunjukkan setiap objek pada cluster dimana objeknya hampir sama, berdasarkan nilai tengah dari objek2 pada cluster
  - Update the cluster means, i.e., calculate the mean value of the objects for each cluster

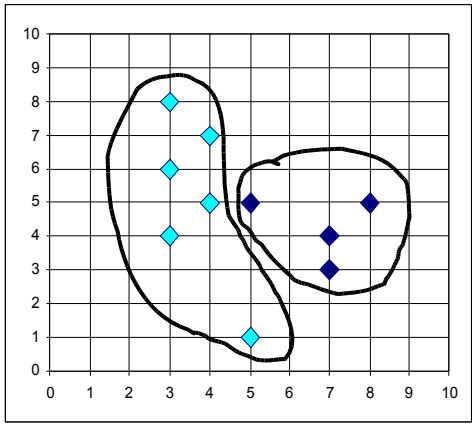
Gambar 3. Contoh : K-Means



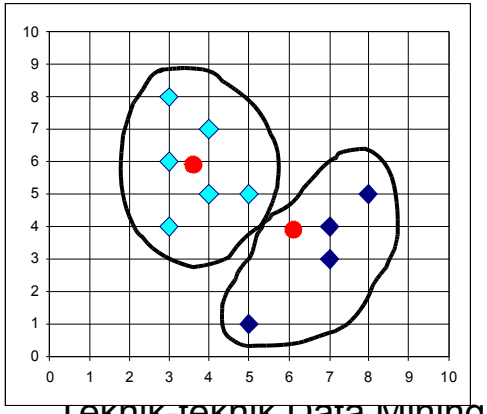
K=2

Arbitrarily choose K object as initial cluster center

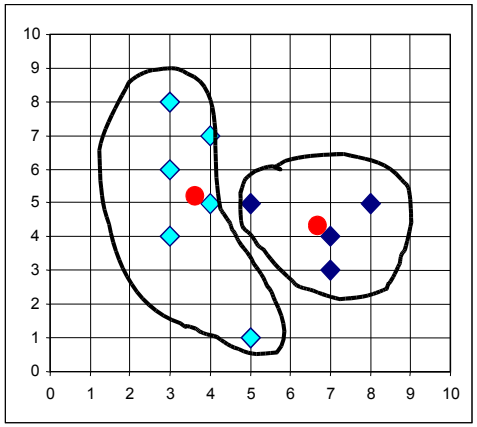
Assign each objects to most similar center



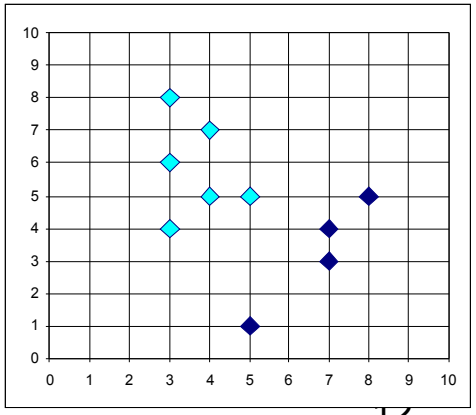
reassign



Update the cluster means



reassign



Update the cluster means

Induksi

- Induksi merupakan salah satu teknik inferensi informasi pada database.

- Ada dua teknik inferensi yakni
  - Induksi merupakan teknik inferensi informasi yang digeneralisasi dari database, contohnya setiap pegawai mempunyai manajer.

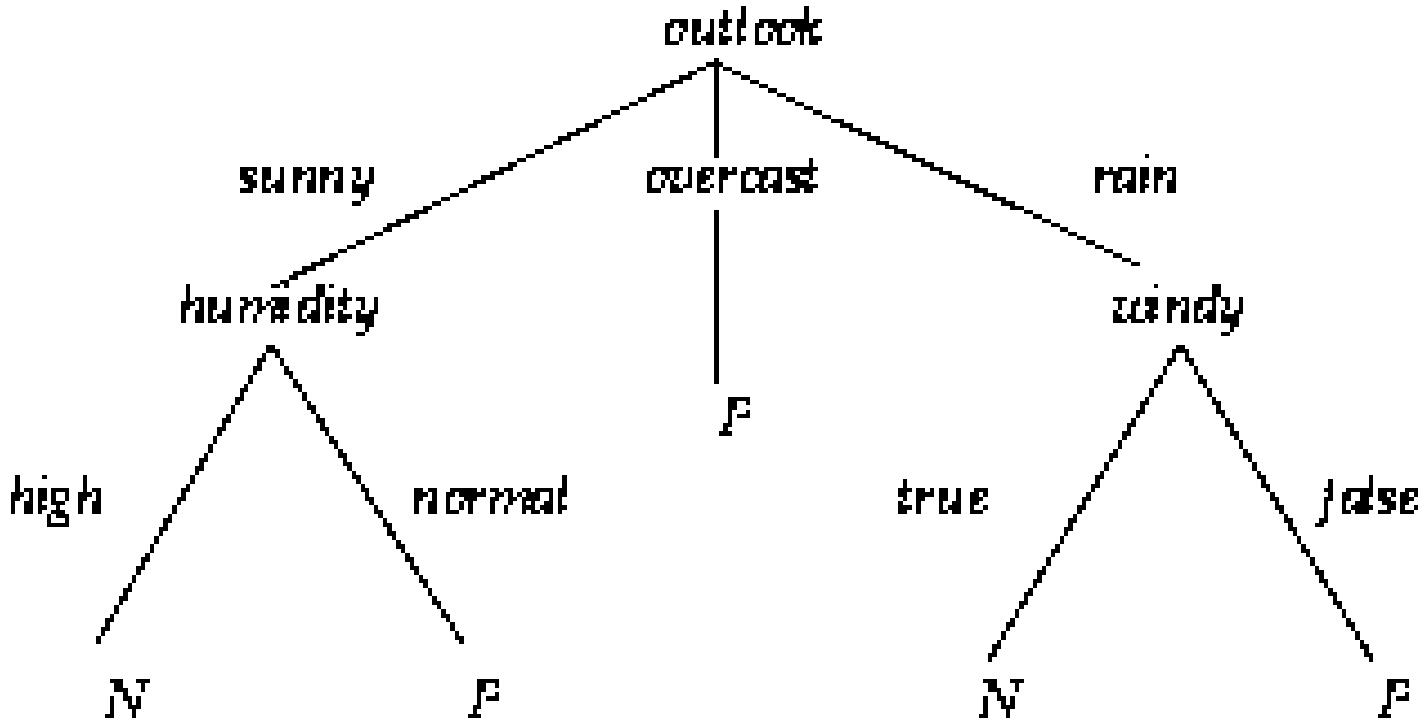
  - Deduksi merupakan teknik inferensi informasi dari konsekuensi logis informasi pada database, contohnya operasi join pada dua tabel; dimana yang pertama mengenai pegawai dan departemen sedangkan yang kedua mengenai departemen dan manajer, menghasilkan relasi antara pegawai dan manajer.

Pohon Keputusan

- Pohon keputusan merupakan representasi pengetahuan yang simpel. Pohon keputusan ini mengklasifikasikan contoh2 pada klas2 dengan angka finit, node diberi nama atribut, edge di beri nilai atribut sedangkan leave diberi nama klas. Objek2 diklasifikasikan dengan struktur pohon, dengan menggunakan dahan2nya sebagai nilai atribut dari objek.

- Gambar berikut mengenai keadaan cuaca. Objek2 berisikan informasi mengenai suasana cuaca, kelembaban dll. Beberapa objek merupakan contoh positif dinotasikan dengan P sedangkan yang lain negatif atau N.

Gambar 4. Struktur Pohon Keputusan



Induksi Aturan

- Sistem data mining harus dapat menyimpulkan suatu model dari database dimana model ini mendefinisikan klas2 seperti halnya database yang terdiri atas satu atau lebih atribut yang menunjukkan klas dari tupel. Klas dapat didefinisikan oleh kondisi atribut.
- Aturan produksi dipergunakan untuk merepresentasikan pengetahuan sistem pakar dan keuntungannya mudah diinterpretasikan oleh kepakaran manusia dikarenakan modularitas yakni aturan yang tunggal dapat dipahami dengan sendirinya dan tidak perlu referensi aturan lain.

Jaringan Syaraf Buatan

- Merupakan pendekatan perhitungan yang melibatkan pengembangan struktur secara matematis dengan kemampuan untuk ‘belajar’.
- Mampu menurunkan pengertian dari data yang kompleks dan tidak jelas dan dapat digunakan pula untuk mengekstrak pola dan mendeteksi tren2 yang sangat kompleks untuk dibicarakan baik oleh manusia maupun teknik komputer lainnya.
- Jaringan syaraf buatan yang terlatih dapat dianggap sebagai ‘pakar’ dalam kategori informasi yang akan dianalisis. Pakar ini dapat digunakan untuk memproyeksi situasi baru dari ketertarikan dan jawaban dari pertanyaan ‘what if’

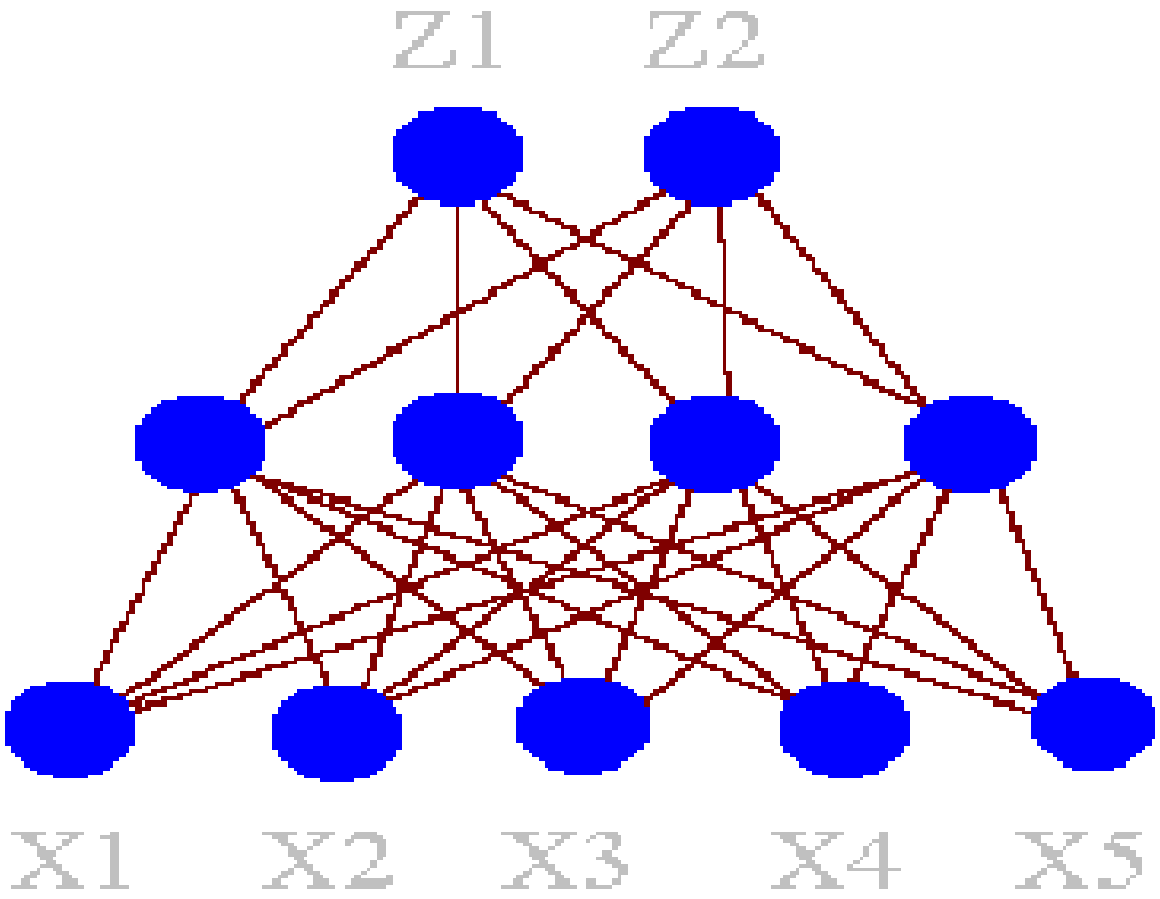
Jaringan Syaraf Buatan (Lanjutan)

- Dikarenakan jaringan syaraf buatan adalah terbaik dalam mengidentifikasikan pola atau tren dalam data, maka cocok pula digunakan untuk kebutuhan memprediksi antara lain:
  - Prediksi penjualan
  - Pengontrolan proses industri
  - Riset Pelanggan
  - Validasi data
  - Manajemen resiko
  - Pemasaran target
  - dll

Jaringan Syaraf Buatan (Lanjutan)

- Jaringan ini menggunakan sekumpulan elemen2 pemrosesan (node) analog pada syaraf otak manusia. Elemen2 pemrosesan ini terhubung dalam jaringan dimana dapat mengidentifikasikan pola2 dalam data sewaktu dipertunjukkan pada data, artinya jaringan belajar dari pengalaman seperti halnya manusia.
- Pada gambar 5, layer bawah adalah lapisan input dengan x1 – x5. Layer tengah disebut juga layer tersembunyi dengan sejumlah variabel node. Layer atas merupakan layer output dengan node z1 – z2 yang diperoleh dari input yang dicobakan.
- Contoh, prediksi penjualan (output) berdasarkan penjualan lama, harga dan cuaca (input).

Gambar 5. Struktur Jaringan Syaraf Buatan

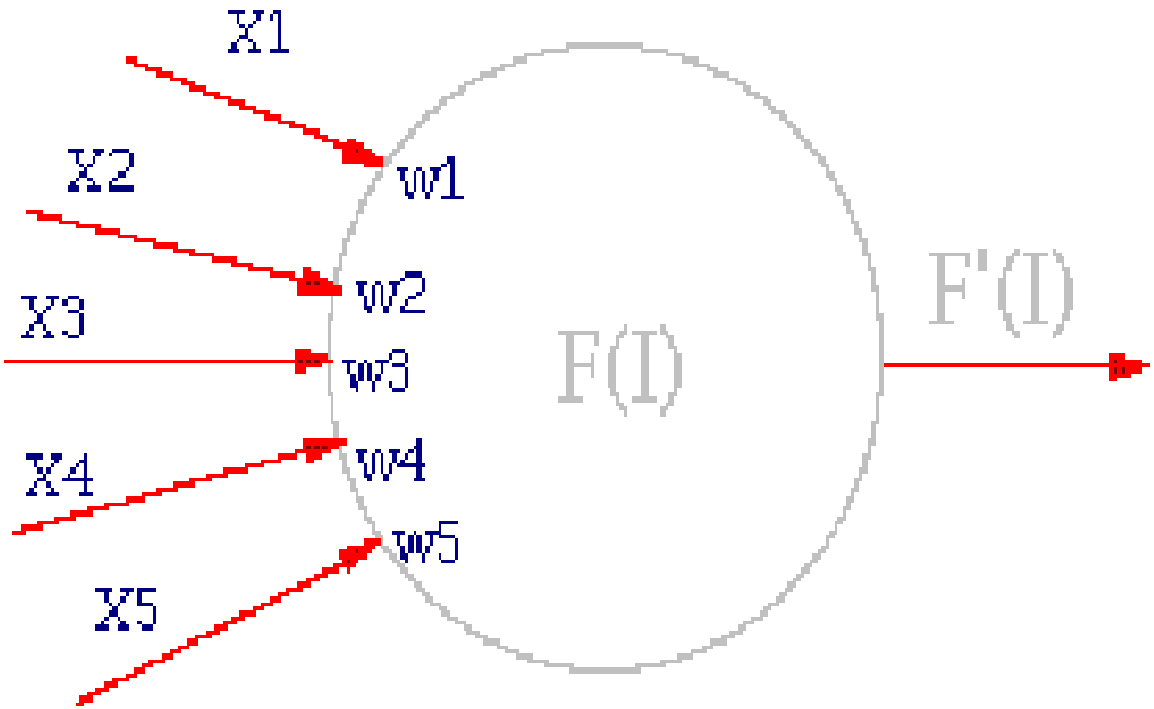


Jaringan Syaraf Buatan (Lanjutan)

- Setiap node yang ada pada layer tersembunyi, secara keseluruhan terhubung dengan input, berarti setiap yg dipelajari didasarkan pada semua input yg diambil bersamaan. Hal ini terlihat pada gambar 6.

- Pada gambar 7. dijelaskan mengenai jaringan syaraf buatan The Clementine User Guide untuk mengidentifikasikan resiko kanker dari berbagai faktor input.

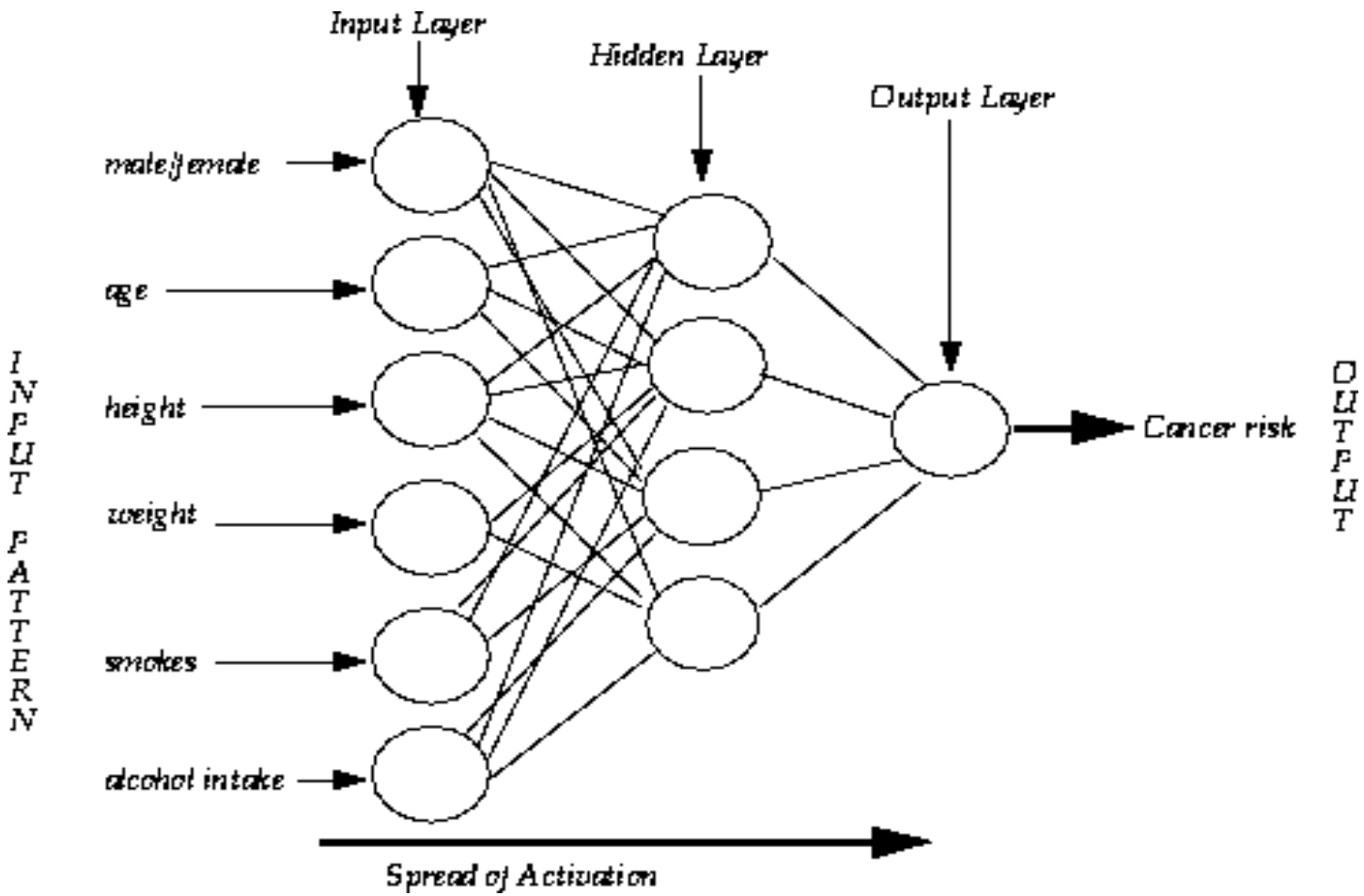
Gambar 6. Di dalam Node



$$F(I) = X1 * w1 + X2 * w2 + X3 * w3 + X4 * w4 + X5 * w5$$

F'(I)=some non-linear transformation of F(I)

Gambar 7. Jaringan Syaraf Buatan
The Clementine User Guide



OLAP (On-line Analytical Processing)
Definisi Menurut E.F. Codd

- OLAP atau *On line Analytical Processing* merupakan salah satu aplikasi database untuk memproses database yang sangat besar dengan data yang kompleks.
- OLAP didefinisikan oleh E.F. Codd (1993) sebagai sintesis dinamik, analisis dan konsolidasi dari data multidimensional yang sangat besar.
- Aturan atau kebutuhan dari sistem OLAP :
  - View konseptual multidimensional
  - Transparansi
  - Aksesibilitas
  - Kinerja reporting yang konsisten
  - Arsitektur client/server
  - Dimensionalitas generik
  - Penanganan *dynamic sparse matrix*
  - Pendukung multi-user
  - Operasi *unrestricted cross dimensional*
  - Manipulasi data intuitif
  - Reporting yang fleksibel
  - Level agregasi dan dimensi yang tidak terbatas

Definisi OLAP Menurut Nigel Pendse

- OLAP didefinisikan oleh Nigel Pendse sebagai *Fast Analysis of Shared Multidimensional Information*, artinya
  - *Fast* dimana pemakai memperoleh respon dalam detik sehingga tidak terputus rantai pemikirannya
  - *Analysis* dimana sistem menyediakan fungsi2 analisis dan lingkup intuitif dan fungsi2 ini dapat mensuplai logika bisnis dan analisis statistikal yang relevan dengan aplikasi user
  - *Shared* dimana sistem mendukung user yang banyak secara konkurensi
  - *Multidimensional* merupakan kebutuhan utama sehingga sistem mensuplai view konseptual multidimensional dari data termasuk pendukung untuk hirarki multiple
  - *Information* merupakan data dan informasi yang diwariskan, dimana dibutuhkan oleh aplikasi user

Komponen OLAP Menurut Kirk Cruikshank

- Kirk Cruikshank dari Arbor Software mengidentifikasikan ada 3 komponen OLAP :
  - Database multidimensional harus dapat mengekspresikan kalkulasi bisnis yang kompleks dengan mudah. Data harus bereferensi dan didefinisikan matematis
  - Navigasi intuitatif dalam penyusunan data ‘roam around’ yang mana membutuhkan hirarki mining
  - Respons instan, yang artinya kebutuhan untuk memberi user informasi secepat mungkin

Contoh OLAP

- Contoh database OLAP misalnya data penjualan yang dikumpulkan dari region, tipe produk dan cabang penjualan.
- Queri OLAP harus mengakses database penjualan yang lebih dari satu tahun dan multi-gigabyte untuk menemukan penjualan produk di setiap region per-tipe produk.
- Queri OLAP dapat dikarakterisasikan sebagai transaksi online yang
  - Mengakses data dalam jumlah besar, mis: data penjualan beberapa tahun
  - Menganalisis relationship antara tipe elemen bisnis mis: penjualan, wilayah, produk dan cabang
  - Melibatkan data yang terkumpul mis: volume penjualan, dollar yang dianggarkan dan dollar yang dihabiskan

Contoh OLAP (lanjutan)

- Menyajikan data dalam berbagai perspektif, mis: penjualan berdasarkan wilayah vs penjualan berdasarkan cabang dari produk dalam setiap wilayah
- Membandingkan data yang terkumpul dalam periode waktu secara hirarki, mis: bulanan, tahunan
- Melibatkan kalkulasi kompleks antara elemen data , mis: keuntungan yang diharapkan sebagai fungsi dari pendapatan penjualan untuk setiap tipe dari cabang penjualan dalam suatu wilayah tertentu.
- Dapat merespon permohonan user secara cepat sehingga user dapat mengikuti proses pemikiran yang analitik tanpa masuk pada sistem

Visualisasi Data

- Visualisasi data memungkinkan si analis menperoleh pemahaman yang dalam dan lebih intuitif mengenai data dan dapat bekerja sebaik mungkin pada data mining.
- Data mining memperbolehkan si analis memfokuskan pola2 dan trend2 tertentu dan menjelajahi ke dalam menggunakan visualisasi.

Selesai