

# Eksplorasi Data (Data Mining)

Kontribusi: Taufik Abidin  
Thursday, 14 September 2006

Perkembangan teknologi komputer, jaringan (network), media penyimpanan, dan multimedia akhir-akhir ini telah mengakibatkan jumlah data yang disimpan meningkat dengan sangat pesat, seperti bioinformatik dengan proyek genome-nya, internet dengan situs-situsnya (world wide web), perbankan dengan data transaksi dan nasabahnya, dan bisnis retail dengan data inventori dan transaksi penjualannya. Perkembangan data yang cukup pesat ini membuka peluang akan kebutuhan teknik-teknik data mining yang dapat mengekstrak informasi dari data berskala besar. Data mining atau juga dikenal dengan sebutan knowledge discovery in database lahir karena data yang terkumpul sekarang ini sudah mencapai terrabyte (1000 gigabytes). Data mining merupakan proses mencari pola-pola menarik dalam data [1]. Secara garis besar, data mining teknik dapat dibagi menjadi 3 kelompok: association rules mining (ARM), clustering, and classification. ARM adalah teknik mencari hubungan dan korelasi menarik diantara objek dalam database yang memenuhi nilai minimum support dan confidence. Aplikasi ARM yang paling umum adalah market basket research (MBR) yang menganalisa korelasi antara pola beli pelanggan dengan data item yang dibeli oleh pelanggan. Hasil analisa ini dapat membantu pengambil keputusan dalam mendesain katalog barang, mengatur letak dan susunan rak barang, dan menentukan kebijakan pemasaran secara tepat. Clustering dapat didefinisikan sebagai proses mengelompokkan sekumpulan objek sedemikian hingga objek dalam satu grup lebih serupa karakteristiknya dibandingkan dengan objek-objek di grup-grup yang lain. Clustering juga dikenal dengan unsupervised learning karena objek-objek dalam database tidak memiliki klas (tipe) yang membedakan antara satu objek dengan objek yang lain. Analisa grup sangat bermanfaat untuk mengetahui dan memahami distribusi data dan sering sekali digunakan sebagai proses awal sebelum teknik-teknik data mining lain digunakan. Berbeda dengan clustering, classification (klasifikasi) merupakan proses menentukan klas (label) dari suatu objek yang tidak memiliki label. Pelabelan objek dilakukan berdasarkan kesamaan karakteristik antara sekumpulan objek (training set) dengan objek baru tersebut. Classification juga dikenal sebagai supervised learning karena training objek digunakan sebagai acuan dalam melakukan klasifikasi. Salah satu contoh aplikasi teknik data mining yang satu ini dibidang perbankan adalah dalam menentukan apakah aplikasi kartu kredit dari seorang nasabah dapat disetujui atau ditolak. Dalam hal ini, terdapat dua klas yaitu: disetujui dan ditolak. Sejumlah data nasabah kartu kredit digunakan sebagai training set dengan peubah-peubah (variable) seperti: umur nasabah, jumlah penghasilan, pekerjaan, klas, dan peubah-peubah lainnya yang berkaitan. Khusus untuk peubah klas, domainnya adalah disetujui atau ditolak. Contoh lain aplikasi classification dalam bidang bioinformatika adalah menentukan fungsi dari gen-gen yang baru ditemukan. Sama halnya dengan contoh sebelumnya, sekumpulan data gen yang fungsinya sudah diketahui digunakan sebagai training objek. Bila jumlah sampel dan data tidak terlalu besar, mungkin proses pengklasifikasian dapat dilakukan secara manual. Namun, di era informasi sekarang ini, jumlah sampel dan data sudah sangat besar, sehingga pengklasifikasian secara manual tidak mungkin lagi digunakan. Ekspedisi kelautan, Sorcerer II, yang dipimpin oleh Dr. Venter [2], membuktikan bahwa algoritma klasifikasi yang efisien dan mampu menangani data berskala besar sangat dibutuhkan. Dalam ekspedisi ini, para peneliti menemukan lebih kurang 1800 spesies bakteri baru dan 1.2 juta gen baru dari sekitar 200 liter air laut yang diambil di laut Sargasso dekat Bermuda. Teknik-teknik Clustering Secara garis besar teknik-teknik clustering dapat dikategorikan dalam 3 kelompok. Teknik clustering berdasarkan jarak (distance-based), berdasarkan kepadatan (density-based), and teknik clustering berdasarkan hirarki (hierarchy-based). Hierarchy-based clustering terbagi menjadi 2 jenis yaitu agglomerative dan divisive. Pendekatan secara agglomerative memulai clustering dengan mengambil setiap objek sebagai objek yang terpisah satu sama lainnya dan menggabungkannya satu persatu berdasarkan suatu metric (measurement). Sebaliknya, divisive memulai clustering dengan menganggap bahwa semua objek berada dalam satu cluster kemudian memecahkannya satu persatu sehingga pada akhirnya setiap objek merupakan suatu cluster tersendiri. Contoh teknik clustering berdasarkan jarak adalah k-mean dan k-median. Contoh teknik clustering berdasarkan kepadatan yang sangat terkenal adalah DBSCAN dan OPTICS. Tutorial tambahan tentang clustering dapat diperoleh di website Andrew Moore, salah seorang professor bidang computer science di Carnegie Mellon University, Andrew Moore website. Teknik-teknik Classification Teknik klasifikasi yang paling sederhana tapi handal adalah KNN (k-nearest neighbor). KNN terkenal karena kesederhanaannya dan kemampuannya memodelkan beragam masalah klasifikasi diberbagai bidang. Teknik ini mencari k objek dalam training set yang memiliki kesamaan terdekat dengan sampel baru yang ingin diberi label, kemudian membiarkan k objek tersebut melakukan voting. Klas yang dominan dalam k objek tersebut akan menjadi klas dari sampel baru. Kelemahan dari teknik ini adalah sulitnya menentukan nilai k itu sendiri. Selain itu, tingkat kompleksitas teknik ini adalah linier terhadap jumlah training set  $O(n)$ . Semakin besar training set, semakin besar pula waktu yang dibutuhkan untuk melakukan klasifikasi. Teknik klasifikasi yang terkenal lainnya adalah SVM (support vector machine), yang dikembangkan oleh Vapnik [3]. Algoritma ini mentransform input ke dimensi yang lebih tinggi (higher dimensional feature space) dengan menggunakan nonlinear mapping (fungsi kernel). Dengan kernel yang sesuai, SVM menentukan hyperplane yang maksimum atau pembatas (decision boundary) sedemikian hingga jarak antara hyperplane dan objek yang terdekat dalam setiap klas adalah maksimum.

[1] J. Han and M. Kamber, Data Mining Concepts and Techniques, 2nd edition, Morgan Kaufmann Publishers, San Francisco, 2006. [2] Sorcerer Expedition, <http://www.sorcerer2expedition.org/version1/HTML/main.htm>, February 6, 2006. [3] V. Vapnik, The Nature of Statistical Learning Theory, Springer-Verlag Publisher, NY, 1995.