

DATA MINING SEBAGAI SOLUSI BISNIS

Abstraksi

Dunia bisnis yang penuh persaingan membuat para pelakunya harus selalu memikirkan strategi-strategi terobosan yang dapat menjamin kelangsungan bisnis mereka. Salah satu aset utama yang dimiliki oleh perusahaan masa kini adalah data bisnis dalam jumlah yang luar biasa banyak. Ini melahirkan kebutuhan akan adanya teknologi yang dapat memanfaatkannya untuk membangkitkan “pengetahuan-pengetahuan” baru, yang dapat membantu dalam pengaturan strategi bisnis. Teknologi data mining hadir sebagai solusi. Makalah ini akan mengulas permasalahan bisnis yang ada dan dasar-dasar data mining melalui bahasan kegunaan, cara kerja dan metodologi-metodologi populer pada teknologi ini (pohon keputusan, klasifikasi, regresi non-linier, berbasis sampel, kebergantungan grafik, dll.).

Abstract

The world of business has always been full of competitions. The executors think relentlessly of the way to get survived. Fortunately, in the modern business world, there is valuable data warehouse that could be utilized to generate new knowledge to help the executives in arranging their business strategies. The knowledge generator, which is data mining technology, would be introduced to the readers. This paper presents the business problems to be solved and the foundations of data mining: the usage, how data mining works, the tasks, and the popular methods (decision rule, classification, non-linear regression, sample based, graphical dependency, etc.).

Diterima : 7 Maret 2002

Disetujui untuk dipublikasikan : 16 Maret 2002

1. Pendahuluan

Tahun 90-an telah melahirkan “gunungan” data di bidang ilmu pengetahuan, bisnis dan pemerintah. Kemampuan teknologi informasi untuk mengumpulkan dan menyimpan berbagai tipe data jauh meninggalkan kemampuan untuk menganalisis, meringkas dan mengekstraksi “pengetahuan” dari data. Metodologi tradisional untuk menganalisis data yang ada, tidak dapat menangani data dalam jumlah besar. Sementara para pelaku bisnis memiliki

kebutuhan-kebutuhan untuk memanfaatkan *gudang data* yang sudah dimiliki, para peneliti melihat peluang untuk melahirkan sebuah teknologi baru yang menjawab kebutuhan ini, yaitu *data mining*. Teknologi ini sekarang sudah ada dan diaplikasikan oleh perusahaan-perusahaan untuk memecahkan berbagai permasalahan bisnis.

Makalah ini akan membahas kebutuhan bisnis, solusi yang dipikirkan para pelaku bisnis, pemanfaatan, cara kerja tugas dan

metodologi-metodologi populer pada *data mining*. Bahasan akan diberikan dari sudut pandang pelaku bisnis dan peneliti. Hal ini dimaksudkan agar para pembaca memperoleh gambaran yang kongkret mengenai *data mining* di dunia bisnis, sekaligus juga mengenal konsep-konsep teoretis yang melandasi teknologi *data mining*.

2. Kebutuhan Bisnis

Dalam dunia bisnis yang selalu dinamis dan penuh persaingan, para pelakunya harus senantiasa memikirkan cara-cara untuk terus *survive* dan jika mungkin mengembangkan skala bisnis mereka. Untuk mencapai hal itu, dapat diringkaskan tiga kebutuhan bisnis, yaitu¹:

- a) Penambahan jenis maupun peningkatan kapasitas produk.
- b) Pengurangan biaya operasi perusahaan.
- c) Peningkatan efektifitas pemasaran dan keuntungan.

Untuk memenuhi kebutuhan-kebutuhan di atas, banyak cara yang dapat ditempuh. Salah satunya adalah dengan memikirkan teknik-teknik pemasaran yang efektif dengan biaya yang minimal. Berikut ini akan dibahas mengenai hal-hal yang berkaitan dengan kegiatan bisnis di bidang pemasaran, seperti identifikasi dan pemahaman permasalahan, analisis pencarian solusi dan pembahasan teknik-teknik untuk mengimplementasikan solusi.

3. Pemahaman Permasalahan dan Konsep Dasar Solusi Bisnis

Langkah pertama untuk menyelesaikan permasalahan bisnis adalah mendefinisikan permasalahan dengan sejelas-jelasnya. Sebagai contoh, permasalahan umum yang dihadapi oleh perusahaan-perusahaan *dot-com* adalah: (1) Bagaimana menyajikan advertensi kepada target yang tepat sasaran. (2) Menyajikan halaman Web yang khusus untuk setiap kustomer (mempersonalisasi

halaman Web) agar kustomer merasa diperlakukan secara khusus dan karenanya akan tetap setia dengan perusahaan itu. (3) Menampilkan informasi produk-produk lain yang biasa dibeli bersamaan dengan produk tertentu. (4) Mengklasifikasi artikel-artikel secara otomatis. (5) Mengelompokkan pengunjung Web yang memiliki kesamaan karakteristik tertentu. (6) Mengestimasi data yang hilang. (7) Memprediksi kelakukan di masa yang akan datang². Pencarian solusi dari masalah-masalah ini akan berkaitan dengan penemuan dan pemanfaatan dari berbagai jenis pola-pola yang tersembunyi dari *gudang data* yang kemungkinan sudah dimiliki oleh perusahaan.

Penjelasan lebih lanjut dari masalah-masalah di atas dan konsep dasar yang dipikirkan oleh para pelaku dan penganalisis bisnis sebagai solusinya diberikan di bawah ini.

a. Perumusan target. Para ahli pemasaran menggunakan teknik-teknik tertentu untuk memilih orang-orang yang menjadi target pemasaran untuk disuguhi advertensi tertentu. Tujuannya antar lain adalah untuk meningkatkan profit perusahaan, pengenalan produk secara luas, atau hasil-hasil terukur lainnya.

b. Personalisasi. Para ahli pemasaran memanfaatkan personalisasi untuk memilih advertensi yang paling sesuai untuk (atau memberikan rekomendasi tertentu kepada) orang tertentu. Personalisasi dapat dipandang sebagai kontradiksi dari “perumusan target”. Pada perumusan target, yang disasar adalah sebanyak mungkin orang yang memiliki potensi untuk membeli produk-produk tertentu, sedangkan pada personalisasi, tujuannya adalah agar kustomer yang sudah menjadi pelanggan membeli sebanyak mungkin produk-produk yang dijual oleh perusahaan.

c. Asosiasi (juga dinamakan analisis keranjang-pasar). Asosiasi ini mengidentifikasi item-item produk yang mungkin dibeli bersamaan dengan produk lain, atau “dilihat” secara bersamaan pada saat mencari informasi mengenai produk tertentu. Pada halaman Web, kustomer diingatkan untuk melihat atau membeli produk-produk yang berkaitan dengan produk yang menjadi minat kustomer.

d. Manajemen pengetahuan. Sistem ini mengidentifikasi dan memanfaatkan pola-pola di dalam dokumen yang berbahasa alami, atau berformat text. Di sini didefinisikan asosiasi antara kata-kata dan konteksnya dalam konsep tingkat-atas. Hal ini dapat dilakukan dengan “melatih” sistem dengan dokumen-dokumen yang sudah ditandai dengan konsep-konsep yang relevan. Sistem kemudian membangun sebuah pencocok pola untuk tiap konsep. Ketika dihadapkan pada dokumen baru, pencocok pola akan memutuskan tingkat relevansi dari dokumen ini terhadap konsep. Pendekatan ini dapat digunakan untuk menyortir dokumen-dokumen baru yang masuk ke dalam kategori-kategori yang sudah ada. Juga dapat digunakan untuk mempersonalisasi publikasi online. Selain itu, dapat juga dimanfaatkan untuk menciptakan atau membangkitkan dokumen jawaban-jawaban secara otomatis terhadap pertanyaan-pertanyaan yang masuk.

e. Pengelompokan (*Clustering*). Pengelompokan mengidentifikasi orang-orang yang memiliki kesamaan karakteristik tertentu, dan kemudian menggunakan karakteristik tersebut sebagai “vektor karakteristik” atau “centroid”. Pengelompokan ini digunakan oleh perusahaan untuk membuat laporan mengenai karakteristik umum dari grup-grup pengunjung (kustomer) yang berbeda.

f. Estimasi dan Prediksi. Estimasi menerka sebuah nilai yang belum diketahui, misalnya penghasilan seseorang, ketika informasi lain mengenai orang tersebut diketahui. Prediksi memperkirakan nilai untuk masa mendatang, misalnya probabilitas orang untuk membeli sebuah mobil baru tahun depan, ketika orang itu belum melakukannya. Atau nilai saham yang akan dibeli tahun depan.

g. Pohon keputusan. Pohon keputusan ini dapat dipandang sebagai diagram alir dari titik-titik pertanyaan yang menuju pada sebuah keputusan. Pohon keputusan ini diterapkan pada sistem pemilihan produk-produk yang dijual perusahaan.

4. Kebutuhan dan Kesempatan untuk *Data Mining*

Ketersediaan data yang melimpah, kebutuhan akan informasi (atau pengetahuan) sebagai pendukung pengambilan keputusan untuk membuat solusi bisnis, dan dukungan infrastruktur di bidang teknologi informasi merupakan cikal-bakal dari lahirnya teknologi *data mining*.

Ketersediaan data transaksi dalam volume yang besar: Bidang-bidang industri yang memiliki data transaksi dalam volume besar ini misalnya jaringan ritel, telekomunikasi, perbankan, kartu kredit, dll. Sistem manajemen transaksi pada industri tersebut merekord informasi-informasi rinci yang diperlukan dalam bisnis mereka.

Informasi sebagai aset perusahaan yang penting: Kebutuhan terhadap informasi telah melahirkan *gudang data* yang mengintegrasikan informasi dari sistem-sistem yang tersebar untuk mendukung pengambilan keputusan. Seringkali *gudang data* ini juga dilengkapi dengan data demografis kustomer dan informasi mengenai rumah-tangga.

Ketersediaan teknologi informasi dalam skala yang terjangkau: Saat ini teknologi informasi berbasis sistem yang terbuka sudah dapat diadopsi secara luas. Ini termasuk sistem manajemen basis data, kaskas penganalisis, dan yang terkini adalah pertukaran informasi dan publikasi melalui jaringan Intranet.

Faktor-faktor tersebut di atas dikombinasikan dengan konsep solusi bisnis yang telah diuraikan sebelumnya, telah melahirkan teknologi *data mining*. *Data mining* dimaksudkan untuk memberikan solusi nyata bagi para pengambil keputusan di dunia bisnis, untuk mengembangkan bisnis mereka.

5. Bahasan Umum *Data Mining*

Data Mining merupakan teknologi baru yang sangat berguna untuk membantu perusahaan-perusahaan menemukan informasi yang sangat penting dari *gudang data* mereka. Kaskas *data mining* meramalkan tren dan sifat-sifat perilaku bisnis yang sangat berguna untuk mendukung pengambilan keputusan penting. Analisis yang diotomatisasi yang dilakukan oleh *data mining* melebihi yang dilakukan oleh sistem pendukung keputusan tradisional yang sudah banyak digunakan. *Data Mining* dapat menjawab pertanyaan-pertanyaan bisnis yang dengan cara tradisional memerlukan banyak waktu untuk menjawabnya. *Data Mining* mengeksplorasi basis data untuk menemukan pola-pola yang tersembunyi, mencari informasi pemrediksi yang mungkin saja terlupakan oleh para pelaku bisnis karena terletak di luar ekspektasi mereka.

Definisi *Data Mining*

Data mining didefinisikan sebagai satu set teknik yang digunakan secara otomatis untuk mengeksplorasi secara menyeluruh dan membawa ke permukaan relasi-relasi yang kompleks pada set data yang sangat besar. Set data yang

dimaksud di sini adalah set data yang berbentuk tabulasi, seperti yang banyak diimplementasikan dalam teknologi manajemen basis data relasional. Akan tetapi, teknik-teknik *data mining* dapat juga diaplikasikan pada representasi data yang lain, seperti domain data *spatial*, berbasis text, dan multimedia (citra). *Data mining* dapat juga didefinisikan sebagai “pemodelan dan penemuan pola-pola yang tersembunyi dengan memanfaatkan data dalam volume yang besar”¹.

Data mining menggunakan pendekatan *discovery-based* dimana pencocokan pola (*pattern-matching*) dan algoritma-algoritma yang lain digunakan untuk menentukan relasi-relasi kunci di dalam data yang dieksplorasi. *Data mining* merupakan komponen baru pada arsitektur sistem pendukung keputusan (DSS) di perusahaan-perusahaan.

Ruang Lingkup *Data Mining*

Data mining (penambangan data), sesuai dengan namanya, berkonotasi sebagai pencarian informasi bisnis yang berharga dari basis data yang sangat besar. Usaha pencarian yang dilakukan dapat dianalogikan dengan penambangan logam mulia dari lahan sumbernya.

Dengan tersedianya basis data dalam kualitas dan ukuran yang memadai, teknologi *data mining* memiliki kemampuan-kemampuan sebagai berikut¹:

- Mengotomatisasi prediksi tren dan sifat-sifat bisnis. *Data mining* mengotomatisasi proses pencarian informasi pemrediksi di dalam basis data yang besar. Pertanyaan-pertanyaan yang berkaitan dengan prediksi ini dapat cepat dijawab langsung dari data yang tersedia. Contoh dari masalah prediksi ini misalnya target pemasaran, peramalan kebangkrutan dan bentuk-bentuk kerugian lainnya.

- Mengotomatisasi penemuan pola-pola yang tidak diketahui sebelumnya. Kakas *data mining* “menyapu” basis data, kemudian mengidentifikasi pola-pola yang sebelumnya tersembunyi dalam satu sapan. Contoh dari penemuan pola ini adalah analisis pada data penjualan ritel untuk mengidentifikasi produk-produk, yang kelihatannya tidak berkaitan, yang seringkali dibeli secara bersamaan oleh kustomer. Contoh lain adalah pendeteksian transaksi palsu dengan kartu kredit dan identifikasi adanya data anomali yang dapat diartikan sebagai data salah ketik (karena kesalahan operator).

Cara Kerja *Data Mining*

Bagaimana tepatnya *data mining* “menggali” hal-hal penting yang belum diketahui sebelumnya atau memprediksi apa yang akan terjadi? Teknik yang digunakan untuk melaksanakan tugas ini disebut pemodelan. Pemodelan di sini dimaksudkan sebagai kegiatan untuk membangun sebuah model pada situasi yang telah diketahui “jawabannya” dan kemudian menerapkannya pada situasi lain yang akan dicari jawabannya.

Sebagai contoh di sini diambil pencarian solusi bisnis di bidang telekomunikasi³. Ada beberapa perusahaan telekomunikasi yang beroperasi di sebuah negara dan

dimisalkan pihak manajemen sebuah perusahaan bermaksud untuk menjaring kustomer baru untuk jasa layanan sambungan langsung jarak jauh (SLJJ). Pihak manajemen dapat “menghubungi” calon-calon kustomer dengan memilih secara acak kemudian menawari mereka dengan diskon khusus, dengan hasil yang kemungkinan besar kurang menggembarakan, atau dengan memanfaatkan pengalaman-pengalaman bisnis yang saat ini sudah tersimpan di basis data perusahaan untuk membangun sebuah model. Perusahaan ini telah memiliki banyak informasi mengenai kustomer perusahaan tersebut: umur, jenis kelamin, sejarah penggunaan fasilitas kredit dan penggunaan SLJJ. Juga sudah diketahui informasi mengenai calon-calon kustomer: umur, jenis kelamin, sejarah penggunaan fasilitas kredit, dll. Masalahnya adalah penggunaan SLJJ untuk para calon kustomer ini belum diketahui, karena mereka saat ini menjadi kustomer dari perusahaan lain. Yang dipikirkan pihak manajemen adalah mencari calon kustomer yang akan menggunakan banyak jasa SLJJ. Usaha untuk mencari jawaban masalah ini dilakukan dengan membangun sebuah model. Tabel 1 memberikan ilustrasi mengenai pembangunan model untuk menentukan calon kustomer (prospek) di sebuah *gudang data*.

Tabel 1. *Data Mining* untuk Menentukan Prospek

	kustomer	prospek
informasi umum (contoh: data demografis)	diketahui	diketahui
informasi khusus (contoh: trasaksi kustomer)	diketahui	target

Gol dari pemodelan ini adalah untuk membuat perkiraan yang didasari kalkulasi untuk mengisi informasi di kuadran kanan bawah pada Tabel 1, berdasar pada informasi umum dan khusus yang sudah ada (dimiliki oleh perusahaan itu). Misalnya, sebuah model

sederhana untuk perusahaan telekomunikasi itu adalah: 98% kustomer “milik” perusahaan itu yang berpenghasilan \$60.000/tahun membelanjakan lebih dari \$80/bulan untuk penggunaan SLJJ. Model ini kemudian dapat diterapkan untuk menarik

kesimpulan dari informasi khusus (sebagai data prospek), dimana saat ini informasi khusus tersebut tidak dimiliki oleh perusahaan. Dengan model ini, calon-calon kustomer baru dapat ditarget secara selektif.

Skenario lain dalam membangun model adalah: memprediksi apa yang akan terjadi di masa mendatang. Model ini ditunjukkan oleh Tabel 2.

Tabel 2. *Data Mining* untuk Prediksi

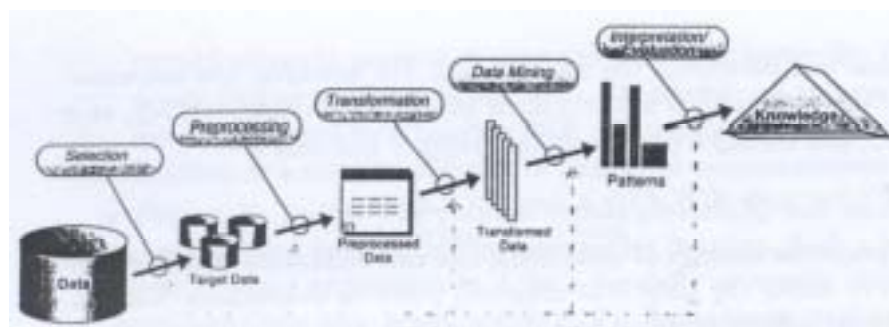
	kemarin	sekarang	besok
informasi statis dan rencana terkini (contoh: data demografis, rencana pemasaran, dll.)	diketahui	diketahui	diketahui
informasi dinamik (contoh: transaksi kustomer)	diketahui	diketahui	target

6. Bahasan Teknis *Data Mining*

Hubungan *Data Mining* dan *Knowledge Data Discovery* (KDD)

Penjelasan umum yang diberikan di atas memberikan pengertian bahwa seolah-olah teknologi *data mining* adalah teknologi utuh dan berdiri sendiri. Dibandingkan dengan *knowledge data*

discovery (KDD), istilah *data mining* lebih dikenal para pelaku bisnis. Pada aplikasinya, sebenarnya *data mining* merupakan bagian dari proses KDD. Sebagai komponen dalam KDD, *data mining* terutama berkaitan dengan ekstraksi dan penghitungan pola-pola dari data yang ditelaah.



Gambar 1. Langkah-langkah dalam proses KDD⁴.

Secara garis besar, langkah-langkah utama dalam proses KDD adalah (lihat Gambar 1):

1. Pemahaman terhadap domain dari aplikasi, relevansinya terhadap pengetahuan yang ada dan *goal* dari *end-user*.
2. Menciptakan himpunan data target: pemilihan himpunan data, atau memfokuskan pada subset variabel atau sampel data, dimana penemuan (*discovery*) akan dilakukan.
3. Pemrosesan pendahuluan dan pembersihan data: operasi dasar seperti penghapusan *noise* dilakukan.
4. Proyeksi dan pengurangan data: pencarian fitur-fitur yang berguna untuk mempresentasikan data bergantung kepada goal yang ingin dicapai.
5. Pemilihan tugas *data mining*: pemilihan *goal* dari proses KDD misalnya klasifikasi, regresi, *clustering*, dll.
6. Pemilihan algoritma *data mining* untuk pencarian (*searching*).

7. *Data mining*: pencarian pola-pola yang diinginkan di himpunan representasi.
8. Penterjemahan pola-pola yang dihasilkan dari *data mining* (langkah 7), kemungkinan dapat kembali langkah 1-7 untuk iterasi lebih lanjut.
9. Konsolidasi pengetahuan yang ditemukan: pendokumentasian hasil, pencarian penyelesaian apabila ada konflik dengan pengetahuan yang telah dipercaya sebelumnya.

Metodologi *Data Mining*

Komponen *data mining* pada proses KDD seringkali merupakan aplikasi iteratif yang berulang dari metodologi *data mining* tertentu. Pada pembahasan di sini akan digunakan istilah *pola* dan *model*. Pola dapat diartikan sebagai instansiasi dari model. Sebagai contoh $f(x) = 3x^2 + x$ adalah pola dari model $f(x) = ax^2 + bx$.

Data mining melakukan “pengepasan” atau pencocokan model ke, atau menentukan pola dari data yang diobservasi. Ada dua pendekatan matematis yang digunakan dalam pencocokan model: *statistik* yang memberikan efek non-deterministik dan *logik* yang murni deterministik. Yang lebih banyak digunakan adalah pendekatan statistik, mengingat ketidakpastian yang ada dalam proses pembangkitan data di dunia nyata.

Kebanyakan metodologi *data mining* didasarkan pada konsep mesin belajar, pengenalan atau pencocokan pola dan statistik: klasifikasi, pengelompokan (*clustering*), pemodelan grafis, dll.⁴

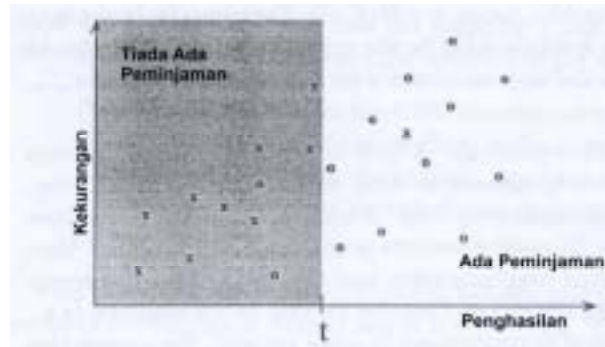
Tugas Utama *Data Mining*

Telah disebutkan di ruang lingkup *data mining* bahwa pada kebanyakan aplikasinya, gol utama dari *data mining* adalah untuk membuat prediksi dan deskripsi. Prediksi menggunakan beberapa variabel atau field-field basis data untuk memprediksi nilai-nilai variabel masa mendatang yang diperlukan, yang belum diketahui saat ini. Deskripsi berfokus pada penemuan pola-pola tersembunyi dari data yang ditelaah. Dalam konteks KDD, deskripsi dipandang lebih penting daripada prediksi⁴. Ini berlawanan dengan aplikasi pengenalan pola dan mesin belajar.

Prediksi dan deskripsi pada *data mining* dilakukan dengan tugas-tugas utama yang akan dijelaskan di bawah ini. Pada setiap tugas akan diberikan *pointer* ke masalah bisnis yang dapat diselesaikan (yang telah dibahas pada butir 3). Gambar-gambar yang ada dimisalkan menunjukkan hubungan antara penghasilan pengecer dan kekurangan pembayaran yang ditanggung oleh distributor (pemasok barang).

a) Klasifikasi adalah fungsi pembelajaran yang memetakan (mengklasifikasi) sebuah unsur (item) data ke dalam salah satu dari beberapa kelas yang sudah didefinisikan. Gambar 2 menunjukkan pembagian sederhana pada data peminjaman menjadi dua

ruang kelas (punya dan tidak punya peminjaman). Pada gambar tersebut x merepresentasikan peminjaman yang bermasalah dan o peminjaman yang pengembaliannya lancar. (Sebagai solusi 3.e, 3.d dan 3.g).

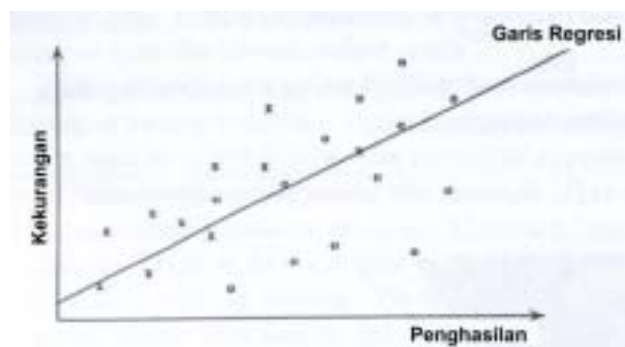


Gambar 2.

Batas klasifikasi linier sederhana pada himpunan data peminjaman⁴.

b) Regresi adalah fungsi pembelajaran yang memetakan sebuah unsur data ke sebuah variabel prediksi bernilai nyata. Aplikasi dari regresi ini misalnya adalah pada prediksi volume biomasa di hutan dengan didasari pada pengukuran gelombang mikro penginderaan jarak jauh (*remotely-sensed*), prediksi kebutuhan kustomer terhadap sebuah produk baru sebagai fungsi dari

pembiayaan advertensi, dll. Gambar 3 menunjukkan regresi linier sederhana dimana “total peminjaman” (*total debt*) diplot sebagai fungsi linier dari penghasilan (*income*): pengeplotan ini menghasilkan kesalahan besar karena hanya ada korelasi sedikit antara kedua variabel ini. (Solusi 3.a dan 3.f)

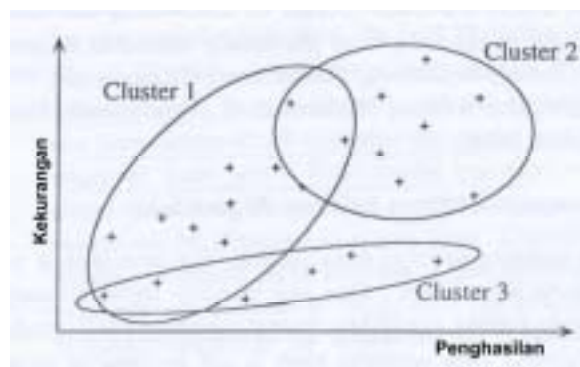


Gambar 3.

Regresi linier sederhana untuk himpunan data peminjaman⁴.

c) Pengelompokan (*clustering*) merupakan tugas deskripsi yang banyak digunakan dalam mengidentifikasi sebuah himpunan terbatas pada kategori atau *cluster* untuk mendeskripsikan data yang ditelaah. Kategori-kategori ini dapat bersifat eksklusif dan ekshaustif mutual, atau mengandung representasi yang lebih kaya seperti kategori yang hirarkis atau saling menumpu

(*overlapping*). Gambar 4 menunjukkan pembagian himpunan data peminjaman menjadi 3 *cluster*. Di sini, *cluster - cluster* dapat saling menumpu, sehingga titik-titik data dapat menjadi anggota lebih dari satu *cluster*. (Label x dan o pada gambar sebelumnya diubah menjadi + untuk mengindikasikan bahwa keanggotaan kelas diasumsikan belum diketahui.) (Solusi 3.e).



Gambar 4.
Pengelompokan himpunan data peminjaman menjadi 3 *cluster*⁴.

d) Peringkasan melibatkan metodologi untuk menemukan deskripsi yang ringkas dari sebuah himpunan data. Satu contoh yang sederhana adalah mentabulasikan mean dan deviasi standar untuk semua field-field tabel. (Solusi 3.f).

f) Pendeteksian Perubahan dan Deviasi berfokus pada penemuan perubahan yang paling signifikan di dalam data dari nilai-nilai yang telah diukur sebelumnya. (Solusi 3.f)

e) Pemodelan Kebergantungan adalah penemuan sebuah model yang mendeskripsikan kebergantungan yang signifikan antara variabel-variabel. Model kebergantungan ini ada di 2 tingkat: tingkat struktural yang menspesifikasikan variabel-variabel yang secara local bergantung satu sama lain, dan tingkat kuantitatif yang menspesifikasikan tingkat kebergantungan dengan menggunakan skala numerik. (Solusi 3.c).

Komponen Algoritma *Data Mining*

Setelah tugas-tugas utama dari data mining didefinisikan seperti di atas, maka perlu dirumuskan algoritma-algoritma untuk mencari solusi dari tugas-tugas tersebut di atas. Dalam setiap algoritma data mining ada tiga komponen utama yaitu representasi model, evaluasi model dan metodologi pencarian.

a) Representasi Model adalah bahasa untuk mendeskripsikan pola-pola yang dapat ditemukan. Jika representasi terlalu terbatas, maka tidak akan ada jumlah waktu

pelatihan maupun sampel yang mencukupi, yang akan menghasilkan model yang akurat untuk data.

- b) Evaluasi Model mengestimasi tingkat kecocokan sebuah pola tertentu untuk memenuhi kriteria pada proses KDD. Evaluasi pada keakuratan prediksi (validasi) didasarkan pada validasi silang. Evaluasi kualitas deskriptif berkaitan dengan akurasi, kebaruan, utilitas dan kemampuan untuk dipahami dari model yang diterapkan. Kriteria logika dan statistik dapat digunakan untuk evaluasi model.
- c) Metodologi Pencarian terdiri dari dua komponen: *pencarian parameter* dan *pencarian model*. Pada pencarian parameter, algoritma harus mencari parameter-parameter yang mengoptimisasi kriteria evaluasi model dengan tersedianya data yang diobservasi dan representasi model yang tetap. Pencarian model terjadi sebagai sebuah *loop* di atas metodologi pencarian parameter: representasi model diubah sehingga dibentuk satu keluarga model-model. Untuk setiap representasi model, metodologi pencarian parameter diinstansiasi untuk mengevaluasi kualitas dari model itu. Implementasi metodologi pencarian model cenderung untuk menggunakan teknik pencarian *heuristic*.

7. Metodologi Data Mining yang Populer

Ada banyak metodologi *data mining*, tapi di sini hanya akan dibahas yang populer saja. Bahasan metodologi akan meliputi segi representasi model, evaluasi model dan metodologi pencarian.

a. Aturan dan Pohon Keputusan

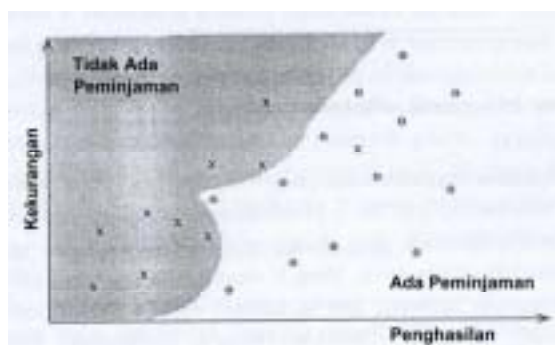
Metodologi ini, yang menggunakan pemisahan (*split univariate*), mudah dipahami oleh pemakai karena bentuk representasinya yang sederhana.. Akan tetapi, batasan-batasan yang diterapkan

pada representasi aturan dan pohon tertentu dapat secara signifikan membatasi bentuk fungsional dari model. Sebagai contoh, Gambar 2 memberikan ilustrasi mengenai efek penerapan pemisahan, yang didasarkan pada nilai ambang tertentu, pada variabel penghasilan (*income*) di himpunan data peminjaman: sangat jelas terlihat bahwa penerapan pemisahan nilai ambang sederhana sangat membatasi tipe batas (*boundary*) klasifikasi yang dapat dihasilkan. Jika ruang model dilebarkan untuk memfasilitasi ekspresi-ekspresi yang lebih umum (misalnya *multivariate hyperplanes* pada berbagai sudut), maka model ini menjadi lebih canggih untuk prediksi. Hanya saja, mungkin akan lebih sulit untuk dipahami pemakai.

Metodologi ini terutama digunakan untuk pemodelan prediksi, keduanya untuk klasifikasi dan regresi⁴. Selain itu, dapat digunakan juga untuk pemodelan deskripsi ringkasan.

b. Metodologi Klasifikasi dan Regresi Non-linier

Kedua metodologi ini terdiri dari sekumpulan teknik-teknik untuk memprediksi kombinasi variabel-variabel masukan yang pas dengan kombinasi linier dan non-linier pada fungsi-fungsi dasar (sigmoid, *splines*, polinomial). Contohnya antara lain adalah jaringan saraf *feedforward*, metodologi *spline* adaptif, dan proyeksi regresi *pursuit*. Gambar 5 menunjukkan tipe boundary keputusan non-linier yang mungkin dihasilkan oleh jaringan saraf. Metodologi regresi non-linier, walaupun canggih dalam representasinya, mungkin sulit untuk diinterpretasikan. Gambar 5 bisa jadi lebih akurat dibandingkan dengan Gambar 2, tapi Gambar 2 lebih mudah untuk diinterpretasikan (jika penghasilan lebih dari *t*, maka peminjaman akan memiliki status yang bagus).



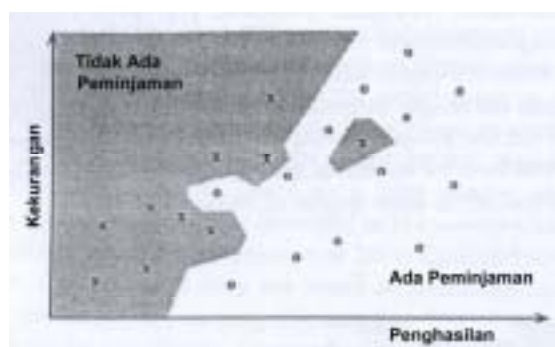
Gambar 5.

Contoh *boundary* klasifikasi yang “dipelajari” pengklasifikasi non-linier⁴.

c. Metodologi Berbasis-sampel

Representasi dari metodologi ini cukup sederhana: gunakan sampel dari basisdata untuk mengaproksimasi sebuah model, misalnya, prediksi sampel-sampel baru diturunkan dari properti sampel-sampel yang “mirip” di dalam model yang prediksinya sudah diketahui. Teknik ini misalnya adalah klasifikasi tetangga-

terdekat, algoritma regresi dan sistem *reasoning* berbasis-kasus. Gambar 6 menunjukkan hasil dari klasifikasi tetangga terdekat pada himpunan data peminjaman: kelas pada setiap titik di dalam ruang 2-dimensi sama dengan kelas dari titik terdekat di dalam himpunan data yang ditelaah dan orisinal.



Gambar 6.

Boundary klasifikasi untuk pengklasifikasi tetangga-terdekat pada himpunan data peminjaman⁴.

Kekurangan pada metodologi berbasis-sampel (misalnya jika dibandingkan dengan berbasis-pohon) adalah dibutuhkan metrik jarak yang akurat untuk mengevaluasi jarak antara titik-titik data.

d. Model Kebergantungan Grafik Probabilistik

Model grafik menspesifikasikan kebergantungan probabilistik yang mendasari sebuah model dalam menggunakan struktur grafik. Dalam bentuknya yang paling sederhana, model ini menspesifikasikan variabel-variabel mana yang bergantung satu sama lain. Pada umumnya, model ini digunakan dengan variabel kategorial atau bernilai diskret, tapi pengembangan untuk kasus khusus, seperti densitas Gaussian, untuk variabel yang bernilai *real* (pecahan) juga dimungkinkan. Baru-baru ini riset di bidang inteligensia buatan dan statistik dilakukan untuk mencari teknik dimana struktur dan parameter-parameter pada model grafik “dipelajari” secara langsung dari basisdata.

e. Model Belajar Relasional

Jika aturan dan pohon-keputusan memiliki sebuah representasi yang terbatas pada logika proporsional, pembelajaran relasional (yang juga dikenal sebagai pemrograman logika induksi) menggunakan bahasa pola yang lebih sederhana dengan logika tingkat-satu. Pembelajar relasional dengan mudah dapat menemukan formula seperti $X=Y$. Kebanyakan riset pada metodologi evaluasi model untuk pembelajaran relasional bersifat logik.

8. Teknologi yang Mendatangkan Profit

Banyak perusahaan yang sudah meluncurkan aplikasi *data mining* (KDD) dan telah mendapatkan keuntungan. Teknologi ini tidak hanya cocok untuk digunakan oleh industri-industri yang

mengelola informasi secara intensif seperti perbankan, tetapi juga perusahaan apa saja yang ingin memanfaatkan *gudang data* untuk manajemen kustomer dengan lebih baik. Dua faktor penting yang menentukan keberhasilan penggunaan dari *data mining* adalah: *gudang data* yang berukuran besar dan terintegrasi dengan baik, dan pemahaman atau identifikasi yang baik terhadap proses bisnis dimana *data mining* akan diaplikasikan⁵.

Beberapa contoh bidang-bidang bisnis yang telah berhasil menerapkan aplikasi *data mining* adalah:

- a) Perusahaan farmasi dapat menganalisis aktivitas penjualan terkininya dan menggunakan hasilnya untuk mentargetkan dokter-dokter yang berpotensi menggunakan produknya dan menentukan aktifitas pemasaran yang paling efektif untuk beberapa bulan mendatang.
- b) Perusahaan kartu kredit dapat memanfaatkan data transaksi kustomer-kustomernya untuk merancang produk kredit baru yang akan menarik minat para kustomer tersebut.
- c) Perusahaan transportasi yang menyediakan berbagai jenis pelayanan. Data mining dapat digunakan untuk mengidentifikasi prospek-prospek pelayanan yang menjanjikan keuntungan.
- d) Perusahaan produk makanan atau kebutuhan sehari-hari. Data mining dapat dimanfaatkan untuk meningkatkan penjualan produk ke para pengecer (*retailer*). Data kustomer, pengiriman, aktivitas kompetitor dapat digunakan untuk menganalisis sebab-sebab kustomer berpindah ke produk merek lain. Kemudian, hasilnya dapat digunakan untuk menyusun strategi pemasaran yang lebih efektif.

9. Pengembangan KDD dan *Data Mining*

Walaupun telah banyak diaplikasikan di dunia bisnis dan mendatangkan profit, teknologi KDD dan *Data Mining* masih memiliki tantangan-tantangan yang harus diatasi. Riset untuk menyempurnakan KDD diperlukan antar lain untuk mengatasi⁴:

- a) Basisdata yang berukuran besar, dengan ratusan tabel, jutaan rekord dan berukuran sampai dengan multi-gigabyte.
- b) Dimensi yang besar, basisdata tidak hanya memiliki jutaan rekord tetapi juga jumlah field (atribut, variabel) yang besar.
- c) Data dan pengetahuan yang berubah terus sehingga pola-pola yang telah ditemukan sebelumnya menjadi tidak berlaku lagi.
- d) Data yang hilang dan banyak salah, hal ini banyak terjadi pada basisdata.
- e) Relasi antar-field basisdata yang kompleks. Saat ini *data mining* masih dirancang untuk relasi yang cukup sederhana.
- f) Integrasi dengan sistem lain. Sistem KDD *standalone* bisa jadi agak kurang bermanfaat. Integrasi yang dimaksud bisa terjadi dengan DBMS, kakas-kakas spreadsheet dan visualisasi, serta pencatat sensor waktu-nyata.

10. Kesimpulan

Data mining, yang hadir sebagai teknologi untuk memanfaatkan ketersediaan data bisnis yang melimpah, telah membantu para pelaku bisnis untuk mempertahankan dan mengembangkan bisnis mereka. Akan tetapi, agar teknologi *data mining* dan KDD ini dapat dimanfaatkan terus dengan baik,

teknologi ini harus terus dapat “bekerja” berdampingan dengan bidang lain di dunia teknologi informasi yang berkembang dengan sangat cepat. Penyempurnaan di sana-sini masih terus diperlukan, karena itu peluang riset di bidang ini masih terbuka lebar.

Pustaka

- [1] Seiner R., “*Digging Up \$\$\$ with Data Mining – An Executive’s Guide*”, The Data Administration Newsletter, 1999, <http://www.tdan.com/i010ht01.htm>.
- [2] Greening D., “*Data Mining on the Web: There’s Gold in that Mountain of Data*”, Web Techniques, Januari 2000, <http://www.webtechniques.com/archives/2000/01/greening/>.
- [3] Therling K., “*An Introduction to Data Mining: Discovering hidden value in your data warehouse*”, <http://www.thearling.com>.
- [4] Fayyad U.M., Piatetsky-Shapiro G., Smyth P., Uthurusamy R., “*Advance in Knowledge Discovery and Data Mining*”, MIT Press, Cambridge MA, 1996.
- [5] Moxon B., “*Defining Data Mining*”, DBMS Online, 1996, <http://www.dbmsmag.com/9608d53.html>.
- [6] Michalski R.S., Bratko I., Kubat M., “*Machine Learning and Data Mining, Methods and Applications*”, John Wiley & Sons Ltd., New York, 1999.

Penulis

Veronica S. Moertini adalah staf pengajar Jurusan Ilmu Komputer, Universitas Katolik Parahyangan, Bandung.