Anomali Data

Data set terdiri dari sejumlah data yang setiap data mempunyai sejumlah fitur yang mendeskripsikan karakter data tersebut. Dalam data mining data set tersebut dapat diolah untuk menghasilkan informasi yang berguna, seperti klasifikasi, clustering, dan sebagainya. Tidak jarang dalam data-data yang akan diolah ditemukan adanya data yang karakteristiknya secara signifikan menyimpang/berbeda dengan karakteristik data pada umumnya.

Menurut Prasetyo (2012) data-data yang mempunyai karakter seperti ini sering disebut sebagai penyimpangan (anomaly/outlier/noise). Jika data yang menyimpang disebut sebagai outlier, maka data yang berada dalam daerah yang wajar/normal maka disebut sebagai inlier. Pada kasus clustering, kehadiran outlier dapat memberikan hasil clustering tidak maksimal. Umumnya penghilangan outlier (outlier removal) menjadi pemrosesan awal (pre-processing) pada data set agar memberikan hasil yang baik.

Sedangkan pada kasus-kasus tertentu berguna untuk mendeteksi keadaan yang tidak biasa pada data yang didapat. Pada bab ini, istilah anomali data, outlier, dan noise yang digunakan mempunyai maksud yang sama.

Outlier biasanya dianggap sebagai obyek/data yang jumlahnya sangat kecil jika dibandingkan dengan data normal lainnya, misalnya probabilitas kemunculannya satu dari seribu data, tetapi bisa menjadi seribu jika data sudah berjumlah satu juta. Dengan demikian, pekerjaan deteksi outlier pada data yang menyimpang menjadi pekerjaan yang penting untuk berbagai keperluan dalam data mining.

Dalam dunia nyata, penyimpangan data dapat dilihat pada sebuah contoh kasus penyimpanan data usia manusia, umumnya usia manusia mulai 0 sampai 90 tahun, hal ini tidak menutup kemungkinan bahwa ada juga manusia yang berusia diatas 90 tahun, misalnya 100 tahun atau 110 tahun. Data manusia yang usianya menyimpang dari data usia pada umumnya disebut anomali/outlier/noise. Data seperti ini kadang disebut juga pengecualian data.

Deteksi anomali adalah proses menemukan outlier dalam dataset yang diberikan. Pencilan adalah objek data yang menonjol di antara objek data lainnya dan tidak sesuai dengan perilaku yang diharapkan dalam dataset. Algoritma pendeteksian anomali memiliki aplikasi luas dalam domain bisnis, ilmiah, dan keamanan di mana mengisolasi dan bertindak atas hasil deteksi outlier sangat penting. Untuk identifikasi anomali, algoritma yang dibahas dalam bab-bab sebelumnya seperti klasifikasi, regresi, dan clustering dapat digunakan. Jika dataset pelatihan memiliki objek dengan hasil anomali yang diketahui, maka salah satu algoritma sains data yang diawasi dapat digunakan untuk deteksi anomali. Selain algoritma yang diawasi, ada algoritma khusus (tanpa pengawasan) yang seluruh tujuannya adalah untuk mendeteksi outlier tanpa menggunakan dataset pelatihan berlabel.

Dalam konteks deteksi anomali tanpa pengawasan, algoritma dapat mengukur jarak dari titik data lain atau kepadatan di sekitar lingkungan titik data. Bahkan teknik pengelompokan dapat dimanfaatkan untuk deteksi anomali. Pencilan biasanya membentuk kluster terpisah dari kluster lain karena mereka jauh dari titik data lainnya. Beberapa teknik yang dibahas dalam bab-bab sebelumnya akan ditinjau kembali dalam konteks deteksi outlier. Sebelum membahas algoritme, istilah outlier atau anomali harus didefinisikan dan alasan titik data tersebut terjadi dalam dataset perlu dipahami.

Pencilan adalah objek data yang sangat berbeda dari objek lain dalam dataset. Oleh karena itu, pencilan selalu didefinisikan dalam konteks objek lain dalam dataset. Seorang individu berpenghasilan tinggi mungkin merupakan pencilan dalam dataset lingkungan kelas menengah, tetapi tidak dalam keanggotaan dataset kepemilikan kendaraan mewah. Secara alami kejadian tersebut, outlier juga jarang dan, karenanya, mereka menonjol di antara titik data lainnya. Misalnya, sebagian besar lalu lintas jaringan komputer adalah sah, dan satu serangan jaringan jahat akan menjadi outlier.

Penyebab Pencilan/outlier

Outliers dalam dataset dapat berasal dari kesalahan dalam data atau dari variabilitas inheren yang valid dalam data. Penting untuk memahami asal dari pencilan karena akan memandu tindakan apa, jika ada, yang harus dilakukan pada pencilan yang diidentifikasi. Namun, menentukan dengan tepat apa yang menyebabkan pencilan adalah tugas yang membosankan dan mungkin tidak mungkin untuk menemukan penyebab pencilan dalam dataset. Berikut adalah beberapa alasan paling umum mengapa pencilan terjadi dalam dataset:

Kesalahan data: Pencilan dapat menjadi bagian dari dataset karena kesalahan pengukuran, kesalahan manusia, atau kesalahan pengumpulan data. Misalnya, dalam kumpulan data ketinggian manusia, pembacaan seperti 1,70 cm jelas merupakan kesalahan dan kemungkinan besar secara keliru dimasukkan ke dalam sistem. Poin data ini sering diabaikan karena mereka mempengaruhi kesimpulan dari tugas ilmu data. Deteksi outlier di sini digunakan sebagai langkah preprocessing dalam algoritma seperti regresi dan jaringan saraf. Kesalahan data karena kesalahan manusia dapat berupa kesalahan yang disengaja atau kesalahan yang tidak disengaja karena kesalahan entri data atau bias yang signifikan. Varians normal dalam data: Dalam distribusi normal, 99,7% titik data berada dalam tiga standar deviasi dari rata-rata. Dengan kata lain, 0,26% atau 1 dalam 370 titik data berada di luar tiga standar deviasi dari rata-rata. Menurut definisi, mereka tidak sering muncul namun merupakan bagian dari data yang sah. Seseorang yang menghasilkan satu miliar dolar dalam setahun atau seseorang yang tingginya lebih dari 7 kaki berada di bawah kategori pencilan dalam dataset

pendapatan atau dataset tinggi manusia. Pencilan ini cenderung beberapa statistik deskriptif seperti rata-rata dataset. Apapun, mereka adalah poin data yang sah dalam dataset. Data dari kelas distribusi lain: Jumlah tampilan halaman harian untuk situs web yang menghadap pelanggan dari alamat IP pengguna biasanya berkisar dari satu hingga beberapa lusin. Namun, tidak jarang menemukan beberapa alamat IP yang mencapai ratusan ribu tampilan halaman dalam sehari. Pencilan ini dapat berupa program otomatis dari komputer (juga disebut bot) melakukan panggilan untuk mengikis konten situs atau mengakses salah satu utilitas situs, baik secara sah atau jahat. Meskipun mereka adalah pencilan, cukup "normal" bagi bot untuk mendaftarkan ribuan tampilan halaman ke situs web. Semua lalu lintas bot berada di bawah distribusi "traffic dari program" kelas yang berbeda dari lalu lintas dari browser biasa yang termasuk dalam kelas pengguna manusia. Asumsi distribusi: Poin data outlier dapat berasal dari asumsi yang salah dibuat pada data atau distribusi. Misalnya, jika data yang diukur adalah penggunaan perpustakaan di sekolah, maka selama ujian semester akan ada pencilan karena lonjakan penggunaan perpustakaan. Demikian pula, akan ada lonjakan penjualan ritel pada hari setelah Thanksgiving di Amerika Serikat. Pencilan dalam hal ini diharapkan dan tidak mewakili titik data dari ukuran tipikal.

Memahami mengapa outlier terjadi akan membantu menentukan tindakan apa yang harus dilakukan setelah deteksi outlier. Dalam beberapa aplikasi, tujuannya adalah untuk mengisolasi dan bertindak pada pencilan seperti yang dapat dilihat dalam pemantauan penipuan transaksi kartu kredit. Dalam hal ini, transaksi kartu kredit menunjukkan perilaku yang berbeda dari transaksi yang paling normal (seperti frekuensi tinggi, jumlah tinggi, atau pemisahan geografis yang sangat besar antara titik transaksi berurutan) harus diisolasi, diperingatkan, dan pemilik kartu kredit harus dihubungi segera untuk memverifikasi keaslian transaksi. Dalam kasus lain, outlier harus disaring karena mereka mungkin condong pada hasil akhir. Deteksi outlier di sini digunakan sebagai teknik preprocessing untuk ilmu data lain atau tugas analitis. Misalnya, para pencari nafkah ultra-tinggi mungkin perlu dihilangkan untuk menggeneralisasi pola pendapatan suatu negara. Di sini outlier adalah poin data yang sah tetapi sengaja diabaikan untuk menggeneralisasi kesimpulan.

DETEKSI PENIPUAN KLIK DALAM IKLAN ONLINE

Peningkatan dalam iklan online telah menjamin model bisnis dan perusahaan Internet yang sukses. Iklan online membuat layanan Internet gratis, seperti pencarian web, konten berita, jejaring sosial, aplikasi seluler, dan layanan lainnya, layak. Salah satu tantangan utama dalam iklan online adalah memitigasi penipuan klik. Penipuan klik adalah proses di mana program otomatis atau seseorang meniru tindakan pengguna normal mengklik iklan online, dengan maksud jahat menipu pengiklan,

penerbit, atau jaringan iklan. Penipuan klik dapat dilakukan oleh pihak kontraktor atau pihak ketiga, seperti pesaing yang mencoba menghabiskan anggaran iklan atau untuk menodai reputasi situs. Penipuan klik mendistorsi ekonomi periklanan dan menimbulkan tantangan besar bagi semua pihak yang terlibat dalam periklanan online (Haddadi, 2010). Mendeteksi, menghilangkan, atau mendiskontokan penipuan klik membuat seluruh pasar dapat dipercaya dan bahkan memberikan keunggulan kompetitif bagi semua pihak.Mendeteksi penipuan klik mengambil keuntungan dari fakta bahwa lalu lintas yang curang menunjukkan pola penelusuran web yang tidak lazim bila dibandingkan dengan data clickstream pada umumnya.Lalu lintas yang curang seringkali tidak mengikuti urutan tindakan yang logis dan berisi tindakan berulang yang akan berbeda dari lalu lintas reguler lainnya (Sadagopan & Li, 2008). Misalnya, sebagian besar traffic palsu menunjukkan salah satu atau banyak dari karakteristik ini:

mereka memiliki kedalaman klik sangat tinggi (jumlah halaman web yang diakses jauh di dalam situs web); waktu antara setiap klik akan sangat singkat; satu sesi akan memiliki jumlah klik yang tinggi pada iklan dibandingkan dengan pengguna normal; alamat IP asal akan berbeda dari target pasar iklan; akan ada sangat sedikit waktu yang dihabiskan untuk situs web target pengiklan; dll. Ini bukan satu sifat yang membedakan lalu lintas palsu dari lalu lintas biasa, tetapi kombinasi dari sifat-sifat tersebut. Mendeteksi penipuan klik adalah proses yang berkelanjutan dan berkembang. Semakin banyak, pelaku penipuan klik semakin canggih dalam meniru karakteristik pengguna browsing web biasa. Karenanya, penipuan klik tidak dapat dihilangkan sepenuhnya, namun, hal itu dapat diatasi dengan terus mengembangkan algoritma baru untuk mengidentifikasi lalu lintas yang curang.Untuk mendeteksi outlier penipuan klik, data streamstream pertama harus disiapkan sedemikian rupa sehingga deteksi menggunakan ilmu data lebih mudah. Dataset baris-kolom relasional dapat disiapkan dengan setiap kunjungan yang menempati setiap baris dan kolom-kolom yang menjadi ciri seperti kedalaman klik, waktu antara setiap klik, klik iklan, total waktu yang dihabiskan di situs web target, dll. Dataset multidimensi ini dapat digunakan untuk deteksi outlier menggunakan ilmu data. Ciri atau atribut Clickstream harus dipertimbangkan, dievaluasi, ditransformasikan, dan ditambahkan ke dalam dataset. Dalam ruang data multidimensi, lalu lintas palsu (titik data) adalah jauh dari catatan kunjungan lain karena atributnya, seperti jumlah klik iklan dalam satu sesi. Kunjungan biasa biasanya memiliki satu atau dua klik iklan dalam satu sesi, sedangkan kunjungan yang curang akan menghasilkan puluhan klik iklan. Demikian pula, atribut lain dapat membantu mengidentifikasi pencilan lebih tepat. Algoritma pendeteksian outlier yang diulas dalam bab ini memberikan skor outlier (skor kecurangan) untuk semua poin data clickstream dan catatan dengan skor yang lebih tinggi diprediksi akan outlier.

Teknik Deteksi Anomali

Manusia memiliki perlengkapan bawaan untuk fokus pada pencilan. Siklus berita yang dialami setiap hari terutama bergantung pada acara-acara outlier. Ketertarikan untuk mengetahui siapa yang tercepat, siapa yang menghasilkan paling banyak, dan siapa yang memenangkan medali atau skor gol terbanyak sebagian karena peningkatan perhatian terhadap outlier. Jika data dalam satu dimensi seperti penghasilan kena pajak untuk individu, pencilan dapat diidentifikasi dengan fungsi penyortiran sederhana. Visualisasi data berdasarkan sebar, histogram, dan bagan kotak-kumis dapat membantu mengidentifikasi pencilan dalam kasus dataset atribut tunggal juga. Teknik yang lebih maju akan menyesuaikan data dengan model distribusi dan menggunakan teknik sains data untuk mendeteksi outlier.

Deteksi Outlier Menggunakan Metode Statistik

Outlier dalam data dapat diidentifikasi dengan membuat model distribusi statistik dari data dan mengidentifikasi titik data yang tidak sesuai dengan model atau titik data yang menempati ujung ekor distribusi. Distribusi yang mendasari banyak set data praktis termasuk dalam distribusi Gaussian (normal). Parameter untuk membangun distribusi normal (mis., Mean dan standar deviasi) dapat diperkirakan dari dataset dan kurva distribusi normal dapat dibuat seperti yang ditunjukkan pada Gambar. 13.1.

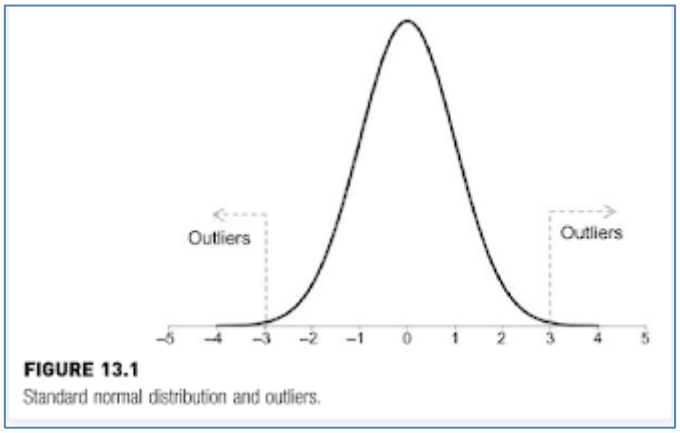

FIGURE 13.1
Standard normal distribution and outliers.

Pencilan dapat dideteksi berdasarkan di mana titik data berada dalam kurva distribusi normal standar. Ambang batas untuk mengklasifikasikan pencilan dapat ditentukan, katakanlah, tiga standar deviasi dari rata-rata. Setiap titik data yang lebih dari tiga standar deviasi diidentifikasi sebagai pencilan. Mengidentifikasi pencilan menggunakan metode ini hanya mempertimbangkan satu atribut atau dimensi pada suatu waktu. Teknik statistik yang lebih maju mempertimbangkan berbagai dimensi dan

menghitung jarak Mahalanobis alih-alih simpangan baku dari rata-rata dalam distribusi univariat. Jarak Mahalanobis adalah generalisasi multivariat untuk menemukan berapa standar deviasi suatu titik dari rata-rata distribusi multivariat. Deteksi outlier menggunakan statistik menyediakan kerangka kerja sederhana untuk membangun model distribusi dan untuk deteksi berdasarkan varians dari titik data dari rata-rata. Salah satu batasan menggunakan model distribusi untuk menemukan outlier adalah bahwa distribusi dataset tidak diketahui sebelumnya. Bahkan jika distribusinya diketahui, data aktual tidak selalu sesuai dengan model.

Deteksi Outlier Menggunakan Data science

Outliers menunjukkan serangkaian karakteristik yang dapat dieksploitasi untuk menemukannya. Berikut ini adalah kelas teknik yang dikembangkan untuk mengidentifikasi outlier dengan menggunakan karakteristik unik mereka (Tan, Steinbach, & Kumar, 2005). Masing-masing teknik ini memiliki banyak parameter dan, karenanya, suatu titik data yang dilabeli sebagai pencilan dalam satu algoritma mungkin bukan pencilan bagi yang lain. Oleh karena itu, lebih baik mengandalkan beberapa algoritma sebelum memberi label outlier.

Berbasis jarak: Secara alami, outlier berbeda dari objek data lain dalam dataset. Dalam ruang Kartesius multidimensi mereka jauh dari titik data lain, seperti yang ditunjukkan pada Gambar. 13.2.

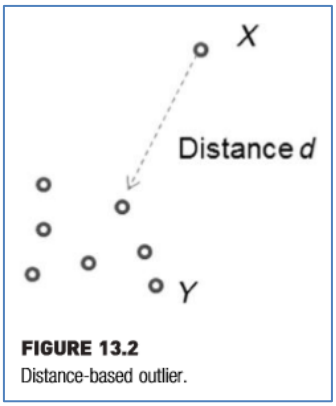

FIGURE 13.2
Distance-based outlier.

Jika jarak rata-rata tetangga N terdekat diukur, outlier akan memiliki nilai lebih tinggi dari titik data normal lainnya. Algoritma berbasis jarak memanfaatkan properti ini untuk mengidentifikasi outlier dalam data. Berbasis kepadatan: Kepadatan titik data di lingkungan berbanding terbalik dengan jarak ke tetangganya. Pencilan menempati area dengan kepadatan rendah sedangkan titik data reguler berkumpul di area dengan kepadatan tinggi. Ini berasal dari fakta bahwa kejadian relatif dari pencilan adalah rendah dibandingkan dengan frekuensi titik data normal.

Berbasis distribusi: Pencilan adalah titik data yang memiliki kemungkinan kejadian yang rendah dan mereka menempati ujung ekor dari kurva distribusi. Jadi, jika seseorang mencoba mencocokkan dataset dalam distribusi statistik, titik data anomali ini akan menonjol dan, karenanya, dapat diidentifikasi. Distribusi normal sederhana dapat digunakan untuk memodelkan dataset dengan menghitung mean dan standar deviasi.

Pengelompokan: Pencilan menurut definisi tidak serupa dengan titik data normal dalam suatu dataset. Mereka adalah titik data langka yang jauh dari titik data biasa dan umumnya tidak membentuk cluster ketat. Karena sebagian besar algoritma pengelompokan memiliki ambang minimum titik data untuk membentuk sebuah cluster, outlier adalah titik data tunggal yang tidak dikelompokkan. Bahkan jika outlier membentuk sebuah cluster, mereka jauh dari cluster lain.

Teknik klasifikasi: Hampir semua teknik klasifikasi dapat digunakan untuk mengidentifikasi pencilan, jika sebelumnya tersedia data rahasia. Dalam teknik klasifikasi untuk mendeteksi pencilan, diperlukan dataset uji yang dikenal di mana salah satu label kelas harus disebut "Pencilan." Model klasifikasi pendeteksian outlier yang dibangun berdasarkan pada dataset uji dapat memprediksi apakah data yang tidak diketahui adalah pencilan atau tidak. Tantangan dalam menggunakan model klasifikasi adalah ketersediaan data yang sebelumnya diberi label. Data outlier mungkin sulit didapat karena jarang. Ini sebagian dapat diselesaikan dengan pengambilan sampel bertingkat di mana catatan outlier dilampau melawan catatan normal.Metode klasifikasi yang diawasi telah dibahas dalam bab-bab sebelumnya dan metode pendeteksian outlier yang tidak diawasi akan dibahas pada bagian selanjutnya. Fokus utamanya akan ditempatkan pada teknik deteksi berbasis jarak dan kepadatan di bagian yang akan datang.

Penerapan Teknik Anomali Data

Penerapan deteksi anomali, dapat ditemukan pada sejumlah bidang-bidang kegunaan seperti: deteksi penggelapan (fraud detection), deteksi penyusupan (intrusion detection), gangguan ekosistem (ecosystem disturbances), kesehatan masyarakat (public health), dan kedokteran. Berikut penjelasan penggunaan deteksi anomali dalam bidang tersebut sebagai berikut :

a. Deteksi penggelapan (fraud detection)

Pembelian barang secara online dengan kartu kredit memicu munculnya para pencuri kartu kredit yang menggunakan nomor kartu kredit pelanggannya dengan membeli barang-barang yang

biasanya mempunyai jenis, jumlah, pola pembelian yang berbeda dengan pelanggan yang seharusnya. Tanpa ada usaha dari bank penyedia layanan kartu kredit untuk secara berkala melakukan deteksi penyimpangan pola transaksi kartu kredit setiap pelanggannya tentu penggelapan isi kartu kredit yang dilakukan pencuri akan merugikan pelanggannya.

b. Deteksi penyusupan (intrusion detection)

Akses ke sumber data dalam instansi, baik komputer maupun jaringan tidak dapat dibantah bahwa datangnya dari tempat umum (internet), tetapi untuk akses yang bisanya bertujuan misalnya untuk mematikan fungsi server atau merusak data yang tersimpan akan memiliki perilaku yang berbeda, misalnya dari isi header protokol atau isi pesan tertentu yang disisipkan didalamnya. Dengan mengintegrasikan metode deteksi anomali pada protokol jaringan seperti firewall atau router, maka diharapkan dapat meningkatkan sistem keamanan jaringan.

c. Gangguan ekosistem (ecosystem disturbances)

Dalam kehidupan di bumi, ada sejumlah pola-pola kehidupan seperti musim, pola kehidupan hewan, yang biasanya ada pola tertentu yang mengalami penyimpangan. Penyimpangan ini dapat mengakibatkan adanya gejala alam yang tidak biasa terjadi. Misalnya, pada masalah penyimpangan pola musim yang dapat mengakibatkan pola kehidupan hewan tertentu menjadi berubah. Deteksi penyimpangan kondisi alam sangat penting untuk dilakukan agar dapat mengetahui sejak dini bahaya-bahaya yang mungkin bisa terjadi, seperti : banjir, kebakaran, gempa, pencemaran, dan sebagainya.

d. Kesehatan masyarakat (public health)

Di Indonesia, ada puskesmas di setiap kecamatan atau kabupaten, data-data rekam medis yang tersimpan di setiap puskesmas mencatat kasus-kasus penyakit yang ditangani disana yang umumnya pasiennya dari daerah terdekat dari lokasi puskesmas. Pola-pola penyakit yang diderita pasien dari keseluruhan puskemas dalam regional tertentu dapat diamati untuk mengetahui pola penyakit yang berbeda yang ditangani pada puskesmas seperti kesalahan cara vaksinasi masyarakat (untuk anak-anak).