DETEKSI DATA PENCILAN MENGGUNAKAN K_MEANS CLUSTERING

Naniek Widyastuti
Teknik Informatika, Fakultas Teknologi Industri
Institut Sains & Teknologi AKPRIND Yogyakarta
e_mail: naniek_wid@yahoo.com

ABSTRACT

Outlier detection is an extremely important task in a wide variety of application e g frand detection, identifying computer network intrusions and bottleneck, credit card fraud, criminal activities in e-commerce. In this paper we are concerned with outlier detection using K_means clustering. In this case number of cluster, is regarded as parameter and incrementally added until we get small cluster and regarded as a collection of outlier. Finally it is illustrated how this method work on sets of data.

Key words : clustering, outlier, K_means

INTISARI

Deteksi data pencilan sangat penting dan mempunyai banyak aplikasi diantaranya adalah identifikasi adanya pengacauan dan sumbatan dalam jaringan komputer, aktivitas kriminal dalam *e-commerce*, deteksi pemalsuan kartu kredit dan aktivitas-aktivitas yang mencurigakan. Dalam tulisan ini dibicarakan deteksi data pencilan menggunakan metode clustering k_means, dengan jumlah cluster dianggap parameter dan secara *incremental* ditambah sampai didapat cluster kecil yang kemudian dianggap sebagai data pencilan. Akhirnya diberikan ilustrasi bagaimana metode tersebut diterapkan pada beberapa kelompok data.

Kata kunci: clustering, data pencilan, k_means

PENDAHULUAN

Data pencilan adalah kumpulan obyek-obyek yang dipandang sangat berbeda dibandingkan keseluruhan data (Han. dan Chamber, 2006). Deteksi data pencilan merupakan persoalan penting dan mempunyai banyak aplikasi diantaranya adalah identifikasi adanya pengacauan dan sumbatan dalam jaringan komputer, aktivitas criminal dalam *e-commerce*, deteksi pemalsuan kartu kredit dan aktivitas-aktivitas yang mencurigakan. Banyak pendekatan telah diusulkan untuk mendeteksi data pencilan dan survai tentang data pencilan diantaranya dapat dilihat dalam Hodge dan Austin (2004).

Clustering adalah teknik yang sangat popular untuk mengelompokkan data atau obyek sejenis ke dalam cluster (Johnson dan Wichren, 2004). Secara umum terdapat empat jenis teknik clustering yaitu pendekatan partisi, hirarki, densitas dan grid. Dalam teknik pendekatan partisi seperti K_means dan Self Organization Maps (SOM) jumlah cluster harus ditentukan terlebih dahulu. Jumlah cluster dapat ditambah atau dikurangi untuk mendapatkan keakuratan hasil clustering. Pernyataan terakhir memungkinkan clustering dapat digunakan untuk mendeteksi data pencilan dalam arti data biasa masuk dalam cluster ukuran besar sedang data pencilan masuk dalam cluster dengan ukuran kecil. Metode K_means berusaha mengelompokkan data, sedemikian sehingga jarak dalam satu klaster kecil dan jarak antar kluster besar. Dengan kata lain, metode ini berusaha untuk meminimalkan variasi antar data yang berada dalam satu cluster dan memaksimalkan variasi dengan data yang berada dalam cluster lain.

Dalam tulisan ini, ditunjukkan bagaimana metode clustering K_means digunakan untuk mendeteksi data pencilan, yaitu dengan cara kumpulan data dikelompokkan ke dalam cluster menggunakan metode K_means dengan jumlah cluster merupakan parameter dan cluster data pencilan adalah kelompok dengan ukuran cluster kecil.

Seperti didiskusikan dalam Niu(2007) tidak ada pendekatan umum yang tunggal dalam mendeteksi data pencilan. Karena

itu, banyak pendekatan telah diusulkan untuk mendeteksi data pencilan. Secara umum menurut Zang dan Wang (2007) ada empat katagori dalam mendeteksi data pencilan yaitu: berbasis distribusi, berbasis jarak, berbasis densitas, dan berbasis cluster.

Dalam pendekatan berbasis distribusi menurut Hawkins (1980), dari kumpulan data yang ada dibangun model statistik yang sesuai (biasanya berdistribusi normal), kemudian dilakukan uji statistik untuk menentukan apakah suatu data masuk dalam model ini atau tidak. Data dengan probabilitas kecil untuk masuk dalam distribusi ini dikatagorikan sebagai data pencilan. Tetapi pendekatan berbasis distribusi tidak dapat diterapkan untuk kasus multidimensi, karena alamiahnya metode tersebut univariat. Di samping itu, pengetahuan awal tentang bentuk distribusi harus diketahui, sehingga sukar diterapkan untuk masalah praktis.

Data pencilan, menurut pendekatan berbasis jarak (Knorr, 2000) dideteksi dengan cara sebagai berikut. Suatu titik q disebut data pencilan terhadap parameter M dan d, dengan M adalah ukuran cluster dan d adalah jarak, bila terdapat kurang dari M titik dengan jarak d dari q, dengan M dan d ditentukan oleh pemakai, masalahnya di sini adalah sukarnya menentukan nilai M dan d.

Dalam pendekatan berbasis densitas (Breunig, 2000), data dalam suatu daerah dihitung kepekatan densitasnya dan data dengan kepekatan densitas rendah dideklarasikan sebagai data pencilan. Setiap data pencilan diberi skor dan disebut Local Outlier Factor (LOF) dan harga ini tergantung pada jarak terhadap tetangga lokalnya. Menurut Loureino (2004), pendekatan berbasis cluster pada dasarnya adalah memandang cluster dengan ukuran kecil sebagai data pencilan. Dalam pendekatan ini, cluster kecil (yaitu cluster dengan anggota yang secara signifikan lebih sedikit dibandingkan anggota cluster lain) dipertimbangkan sebagai data pencilan. Keunggulan pendekatan berbasis cluster adalah metode tersebut tidak perlu disupervisi. Di samping itu, pendekatan berbasis cluster dapat digunakan dalam *mode incremental* (sesudah mempelajari cluster, data baru dapat disisipkan dalam system dan diuji sebagai data pencilan).

Dalam tulisan ini untuk mendeteksi pencilan data digunakan metode K_means clustering. Adapun Algoritma yang digunakan dalam mengelompokkan data menjadi cluster (K_means clustering) adalah sebagai berikut:

Langkah 1 : Tentukan jumlah cluster.
Langkah 2 : Alokasikan data ke dalam cluster secara random.
Langkah 3: Hitung centroid atau rata-rata data yang ada di masing-masing cluster.
Langkah 4: Alokasikan masing-masing data ke cluster dengan centroid atau rata-rata terdekat.
Langkah 5: Kembali ke langkah 3, apabila masih ada data yang berpindah cluster atau apabila nilai centroid ada yang di atas nilai ambang yang ditentukan atau apabila perubahan nilai pada fungsi obyektif yang digunakan di atas nilai ambang yang ditentukan.

Dalam mendefinisikan cluster kecil, mengikuti Loureino (2004) yaitu cluster dengan jumlah titik lebih kecil dari setengah rata-rata titik dalam K cluster. Mode incremental yang digunakan dimulai dengan K=2, K=3 dan seterusnya sampai didapat data pencilan.

PEMBAHASAN

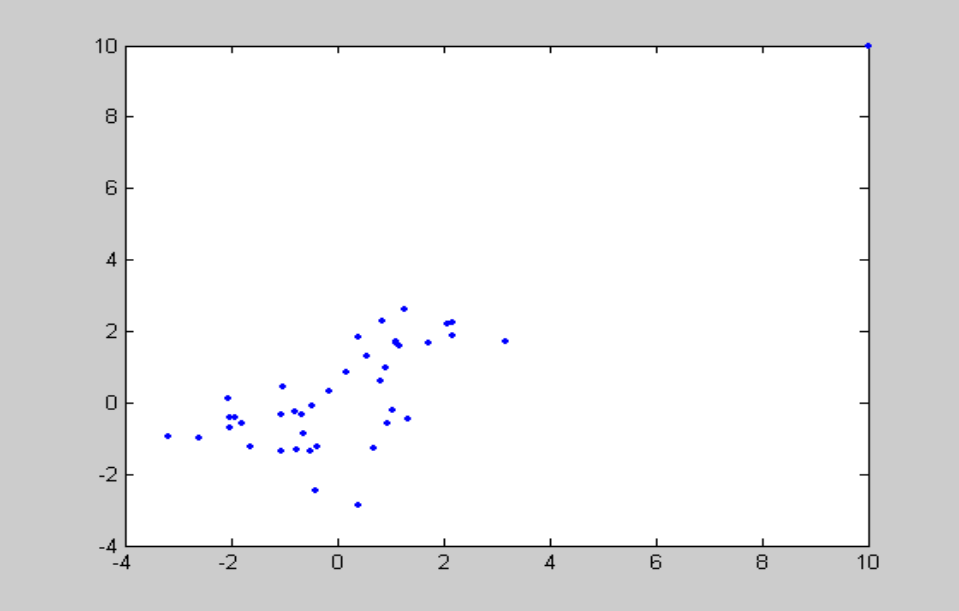
Dalam pembahasan ini akan diselidiki efektifitas algoritma K_means dalam mendeteksi adanya data pencilan. Dimulai dari kumpulan data awal, yaitu data dua dimensi yang ditunjukkan dalam tabel 1, dari gambar 1 terlihat jelas bahwa kumpulan data awal mempunyai sebuah pencilan yang terletak diujung kanan. Pada pembahasan disini diambil data secara acak sejumlah 41 pasang, kemudian akan dilihat apakah deteksi data pencilan menggunakan metode K_means Clustering bisa berjalan atau tidak.

Dimulai dengan kumpulan data acak sejumlah 41 pasang pada tabel 1 berikut

Tabel 1. Tabel Data Awal

No	x	y	No	x	y	No	x	y
1	0.5674	1.2944	15	0.8636	2.2902	29	-3.1707	-0.9597
2	-0.6656	-0.3362	16	1.1139	1.6686	30	-1.0592	-0.3229
3	1.1253	1.7143	17	2.0668	2.1908	31	-2.0106	-0.4311
4	1.2877	2.6236	18	1.0593	-0.2025	32	-0.3855	-1.2556
5	-0.1465	0.3082	19	0.9044	0.9802	33	-0.4923	-1.3775
6	2.1909	1.8580	20	0.1677	0.8433	34	0.6924	-1.2959
7	2.1892	2.2540	21	-2.6041	-1.0000	35	-0.4087	-2.4751
8	0.9624	-0.5937	22	-0.7427	-1.3179	36	-1.6436	-1.2340
9	1.3273	-0.4410	23	-2.0565	0.0950	37	-0.6197	-0.8816
10	1.1746	1.5711	24	0.4151	-2.8740	38	-2.0091	-0.6852
11	0.8133	0.6001	25	-1.8051	-0.5718	39	-1.0195	0.4435
12	1.7258	1.6900	26	-0.4713	-0.1044	40	-1.0482	-1.3510
13	0.4117	1.8156	27	-0.7807	-0.2690	41	10.0000	10.000
14	3.1832	1.7119	28	-1.9219	-0.4221			

Diagram pencar dari data awal terlihat dari gambar di bawah



Gambar 1. diagram pencar dari data awal

Menggunakan K_means clustering untuk k = 2 dengan program MATLAB didapat hasil seperti Tabel 2 berikut

Tabel 2. Hasil peng clusteran dengan k=2

Nomor	Nomor kluster	x	y
1	1	0.5674	1.2944
2	2	-0.6656	-0.3362
3	1	1.1253	1.7143
4	1	1.2877	2.6236
5	2	-0.1465	0.3082
6	1	2.1909	1.8580
7	1	2.1892	2.2540
8	2	0.9624	-0.5937
9	2	1.3273.	-0.4410
10	1	1.1746	1.5711
11	2	0.8133	0.6001

12	1	1.7258	1.6900
13	1	0.4117	1.8156
14	1	3.1832	1.7119
15	1	0.8636	2.2902
16	1	1.1139	1.6686
17	1	2.0668	2.1908
18	2	1.0593	-0.2025
19	1	0.9044	0.9802
20	2	0.1677	0.8433
21	2	-2.6041	-1.000
22	2	-0.7427	-1.3179
23	2	-2.0565	0.0950
24	2	0.4151	-2.8740
25	2	-1.8051	-0.5718
26	2	-0.4713	-0.1044
27	2	-0.7807	-0.2690
28	2	-1.9219	-0.4221
29	2	-3.1707	-0.9597
30	2	-1.0592	-0.3229
31	2	-2.0106	-0.4311
32	2	-0.3855	-1.2556
33	2	-0.4923	-1.3775
34	2	0.6924	-1.2959
35	2	-0.4087	-2.4751
36	2	-1.6436	-1.2340
37	2	-0.6197	-0.8816
38	2	-2.0091	-0.6852
39	2	-1.0195	0.4435
40	2	-1.0482	-1.3510
41	1	10	10

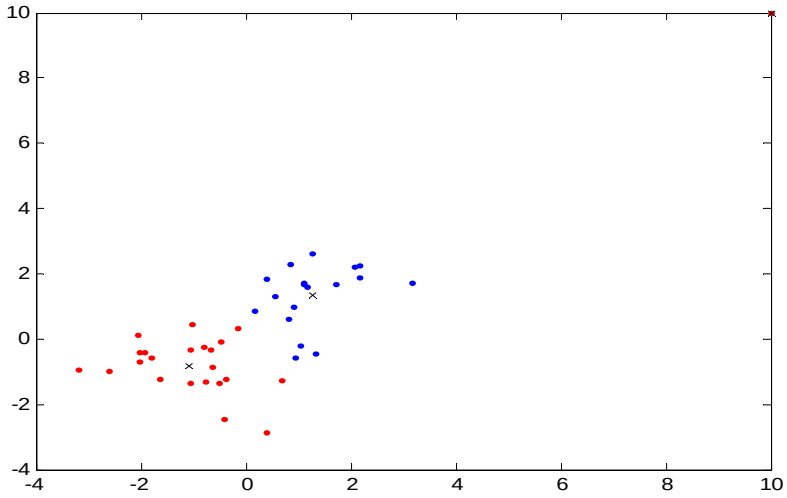
Jumlah anggota kluster 1 adalah 14 dan jumlah anggota kluster 2 adalah 27, sehingga terlihat bahwa untuk jumlah klaster k=2 tidak terdeteksi adanya data pencilan. Dengan program MATLAB didapat pusat kluster 1 adalah (-0.7268; -0.6708) sedang pusat kluster 2 adalah (2.0575; 2.4045)
Kemudian dicobakan menggunakan K_means clustering untuk k=3 dengan program MATLAB didapat hasil sebagai berikut

Tabel 3. Hasil peng clusteran dengan k=3

Nomor	Nomor kluster	x	y
1	2	0.5674	1.2944
2	1	-0.6656	-0.3362
3	2	1.1253	1.7143
4	2	1.2877	2.6236
5	1	-0.1465	0.3082
6	2	2.1909	1.8580
7	2	2.1892	2.2540
8	2	0.9624	-0.5937
9	2	1.3273	-0.4410
10	2	1.1746	1.5711
11	2	0.8133	0.6001
12	2	1.7258	1.6900
13	2	0.4117	1.8156
14	2	3.1832	1.7119
15	2	0.8636	2.2902

16	2	1.1139	1.6686
17	2	2.0668	2.1908
18	2	1.0593	-0.2025
19	2	0.9044	0.9802
20	2	0.1677	0.8433
21	1	-2.6041	-1.0000
22	1	-0.7427	-1.3179
23	1	-2.0565	0.0950
24	1	0.4151	-2.8740
25	1	-1.8051	-0.5718
26	1	-0.4713	-0.1044
27	1	-0.7807	-0.2690
28	1	-1.9219	-0.4221
29	1	-3.1707	-0.9597
30	1	-1.0592	-0.3229
31	1	-2.0106	-0.4311
32	1	-0.3855	-1.2556
33	1	-0.4923	-1.3775
34	1	0.6924.	-1.2959
35	1	-0.4087	-2.4751
36	1	-1.6436	-1.2340
37	1	-0.6197	-0.8816
38	1	-2.0091	-0.6852
39	1	-1.0195	0.4435
40	1	-1.0482	-1.3510
41	3	10	10

Jumlah anggota kluster 1 adalah 22, jumlah anggota kluster 2 adalah 18, sedang jumlah anggota kluster 3 adalah 1, sehingga terlihat bahwa terjadi pencilan data yaitu pada kluster ke 3. adapun gambar data dengan 3 buah kluster terlihat pada gambar 2 dibawah. Dengan program MATLAB didapat pusat kluster 1 adalah (2.2777; 2.6157), pusat kluster 2 adalah (-1.8499; -0.5854) dan pusat kluster 3 adalah (0.1220; -0.5221).



Gambar 2. diagram pencar dengan jumlah kluster k=3

KESIMPULAN

Dalam tulisan ini telah ditunjukkan bahwa deteksi data pencilan dapat dilakukan dengan menggunakan algoritma K_means clustering. Dengan menganggap jumlah cluster sebagai parameter, maka cluster dengan ukuran kecil dapat ditentukan dan selanjutnya cluster tersebut dapat dipandang sebagai data pencilan. Algoritma K_means memang dikenal

kebaikannya untuk jarak Eucledian. Banyak masalah lain yang pengclusterannya tidak berbasis jarak Eucledian seperti berbasis densitas, fungsi obyektif sehingga penelitian data pencilan dengan basis bukan jarak Eucledian masih terbuka untuk dilakukan.

DAFTAR PUSTAKA

Breunig, M., H. Kriegel, R. Ng and J. Sander, 2000, Lof: identifying density-based local outliers. In Proceedings of 2000 ACM SIGMOD International Conference on Management of Data, ACM Press, 93-104

Han, J. and M. Chamber , 2006. Data Mining: Concepts and Techniques, Morgan Kaufmann, 2nd ed.

Hawkins, D., 1980, Identifications of Outliers, Chapman and Hall, London

Hodge, V. and J. Austin, 2004. A Survey of Outlier Detection Methodologies, Artificial Intelligence Review, 22: 85–126.

Johnson,R.A,Wichren,D,2004.Applied Multivariate Analysis. Prentice Hall

Knorr, E., R. Ng, and V. Tucakov, 2000, Distance-based Outliers: Algorithms and Applications, VLDB Journal, 8(3-4): 237-253

Loureiro,A., L. Torgo and C. Soares, 2004. Outlier Detection using Clustering Methods: a DataCleaning Application, in Proceedings of KDNet Symposium on Knowledge-based Systems for the Public Sector. Bonn, Germany.

Niu, K., C. Huang, S. Zhang, and J. Chen, 2007. ODCC: Outlier Detection using Distance Distribution Clustering, T. Washio et al. (Eds.) : PAKDD 2007 Workshops, Lecture Notes in Artificial Intelligence (LNAI) 4819, pp. 332-343, Springer-Verlag.

Zhang, J. and H. Wang, 2007. Detecting outlying subspaces for high-dimensional data: the new Task, Algorithms, and Performance, Knowledge and Information Systems, 10(3): 333-355.